Frank Stajano
Hyoung Joong Kim
Jong-Suk Chae
Seong-Dong Kim (Eds.)

Ubiquitous Convergence Technology

LNCS 4412

First International Conference, ICUCT 2006
Jeju Island, Korea, December 2006
Revised Selected Papers

Springer

Lecture Notes in Computer Science 4412

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

Frank Stajano Hyoung Joong Kim
Jong-Suk Chae Seong-Dong Kim (Eds.)

Ubiquitous Convergence Technology

First International Conference, ICUCT 2006
Jeju Island, Korea, December 5-6, 2006
Revised Selected Papers

Springer

Volume Editors

Frank Stajano
University of Cambridge
Cambridge CB3 0FD, United Kingdom
E-mail: fms27@cam.ac.uk

Hyoung Joong Kim
Korea University
Seoul 136-701, Korea
E-mail: khj-@korea.ac.kr

Jong-Suk Chae
ETRI
Daejeon, 305-700, Korea
E-mail: jschae@etri.re.kr

Seong-Dong Kim
KETI/Ubiquitous Research
Gyeonggi-do 463-816, Korea
E-mail: sdkim@keti.re.kr

Library of Congress Control Number: 2007923849

CR Subject Classification (1998): C.2, C.3, D.4, D.2, K.6.5, H.5.3, K.4

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-71788-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71788-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12044061 06/3180 5 4 3 2 1 0

Preface

Ubiquitous computing is already with us and is changing our lifestyle, way of thinking and quality of life. Everyday objects with embedded computing capabilities are now commonplace and, between mobile phones and RFID tags, further deployment proceeds at an unstoppable pace. The next major step of the ubiquitous computing evolution is the move, already partly underway, from isolated smart objects to distributed systems of smart objects and appropriate back-end infrastructure: microelectronics and communication technology converging with healthcare technology, communication technology, sports and entertainment, housing, vehicular technology, middleware, sensor networks and so on.

You will have noticed that many people in the field now use the word "ubiquitous" not to mean "present everywhere" but as a shorthand for "ubiquitous computing and communications"—leading to otherwise inexplicable locutions such as "the ubiquitous society". Rather than continuing to fight this synecdochical use we have chosen to go with the flow, in so far as the change in language is an indication of the global spread of the meme. We have therefore chosen "ubiquitous convergence" as a concise description of the above view: a systems-oriented perspective encompassing both the technology and its applications.

The First International Conference on Ubiquitous Convergence Technology (ICUCT) was held on Jeju Island, Korea on December 5–6, 2006. This was the first conference organized by the Institute of Electronics Engineers in Korea (IEEK) to celebrate its 60^{th} anniversary. This conference was organized to pave the way for the ubiquitous society by contributing to the development of ubiquitous technologies and their integration in the appropriate application domains. This volume collects the post-proceedings of the conference.

At ICUCT 2006 we accepted only 30 papers from around 640 submissions. We believe the acceptance rate of less than 5% is a clear indication of our commitment to ensuring a very high quality conference. This would not have been possible without the support of our excellent Technical Program Committee members who accurately reviewed and ranked an extraordinarily high number of papers under pressing deadlines. We express our extreme gratitude to all the Program Committee members for their dedication and hard work.

Due to the overwhelming number of submissions, it was impossible to evaluate all papers in one pass in the usual way. Thus, the evaluation process was twofold. In the first round, each reviewer reviewed and classed around 30 papers. After the first round of evaluation, 90 papers were selected. In the second round, 30 papers were accepted. One no-show paper was excluded from the post-proceedings. In addition, we invited Hide Tokuda and Yo-Sung Ho to deliver keynote talks, and we thank them for their valuable contributions. Yo-Sung Ho also wrote up his

talk as an invited paper. This volume therefore contains one invited paper and 29 refereed papers.

All accepted authors were asked to revise and update their papers after the conference based on the written comments from the reviewers and on the formal and informal feedback they received at the conference from other attendees following their presentation. In choosing which papers to accept we tried to achieve a balance among important topics while keeping the paper quality high. Mobile and wireless communication techniques, multimedia technologies, security issues, RFID, sensor networks, applications and convergence aspects of relevant technologies are covered in this conference. These papers address both theoretical and practical issues which, we believe, are of broad interest to our community.

We hope the reader will find this volume to be a timely collection of quality papers that will help to advance the field of ubiquitous convergence technology.

December 2006

Frank Stajano
Hyoung Joong Kim
Jong-Suk Chae
Seong-Dong Kim

Conference Organizers

Conference Chair

Injung Park (DKU, Korea)

Technical Program Committee Chairs

Frank Stajano (University of Cambridge, UK)
Hyoung Joong Kim (Korea University, Korea)

Financial Committee

Kyle J. Kim (KETI, Korea)

Publications Committee

Soo-Hyun Park (Kookmin University, Korea)
Beongku An (Hongik University, Korea)

Local Arrangement Committee

Sang Yep Nam (Kyungmoon College, Korea)
Do Hyun Kim (Jeju National University, Korea)

Technical Program Committee

Beongku An, Hongik University, Korea
Alastair Beresford, University of Cambridge, UK
Ki-Seok Chung, Hanyang University, Korea
Jon Crowcroft, University of Cambridge, UK
Anind Dey, Carnegie Mellon University, USA
Nigel Davies, Lancaster University, UK
Alois Ferscha, Johannes Kepler Universität Linz, Austria
Christian Flörkemeier, ETH Zurich, Switzerland
Enrico Gregori, CNR, Italy
Marco Gruteser, Rutgers University, USA
Yo-Sung Ho, GIST, Korea
Yu Huihua, Fudan University, China
Young Huh, KERI, Korea
Sozo Inoue, Kyushu University, Japan
Jae JI Jung, Hanyang University, Korea

Daeyoung Kim, Information and Communications University, Korea
Hyoung Joong Kim, Korea University, Korea
Seong-Dong Kim, KETI, Korea
Ho Yeol Kwon, Dangwon National University, Korea
Yong Joon Lee, ETRI, Korea
Petri Mähönen, Aachen University, Germany
Cecilia Mascolo, University College London, UK
Young Shik Moon, Hanyang University, Korea
Jae In Oh, Dankook University, Korea
Shingo Ohmori, NICT, Japan
Neeli Prasad, Aalborg University, Denmark
Kasim Rehman, Cambridge Systems Associates, UK
Tapani Ryhänen, Nokia Research Center, Finland
Alberto Sanna, Scientific Institute H San Raffaele, Italy
Ichiro Satoh, National Institute of Informatics, Japan
Hideyuki Tokuda, Keio University, Japan
Pim Tuyls, Philips Research, Netherlands
Pablo Vidales, Deutsche Telekom Laboratories, Germany
Dirk Westhoff, NEC Laboratories, Germany

Table of Contents

Keynotes

Multimedia

Applications

Mobile, Wireless, and Ad Hoc Networking

Smart Sensors and Sensor Networks

Privacy and Security

Web-Based Simulation, Natural System

Video Coding Techniques for Ubiquitous Multimedia Services

Yo-Sung Ho and Seung-Hwan Kim

Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Korea
{hoyo, kshkim}@gist.ac.kr

Abstract. Emerging ubiquitous multimedia services are expected to be available anytime, anywhere, and using different computing devices. Video compression is necessary for transmission of digital video over today's band-limited networks, or for storage constrained applications. This paper gives a short overview over previous video coding standards and analyzes in more detail H.264, which is the latest international video coding standard. Since scalable video coding (SVC) provides the capability of reconstructing lower resolution or lower quality signals from partial bitstream, it is a good paradigm to the streaming video application for ubiquitous multimedia service. Hence, we also discuss several coding techniques and frameworks for SVC including fine granular scalability (FGS).

Keywords: Video coding standard, ubiquitous, H.264, scalable video coding, fine granular scalability.

1 Introduction

The deployment of multimedia services such as audio/video-on-demand, digital library, remote camera surveillance, and distributed visual tracking is becoming ubiquitous. Since limited transmission bandwidth or storage capacity stresses the demand for higher video compression ratios, video compression has been a critical component of many multimedia applications available today. Meanwhile, international study groups, VCEG (Video Coding Experts Group) of ITU-T (International Telecommunication Union - Telecommunication sector) and MPEG (Moving Picture Experts Group) of ISO/IEC, have researched the video coding techniques for various applications of moving pictures since the early 1990s.

ITU-T developed H.261 as the first video coding standard for videoconferencing application. H.261 [1] supports video telephony and videoconferencing over ISDN circuit-switched networks. These networks operate at multiples of 64 kbps and the standard was designed to offer computationally simple video coding for these bitrate. The coding algorithm is a hybrid of transform coding and inter-picture prediction with an integer-accuracy motion compensation.The block-based inter-picture prediction is used to removes temporal redundancy between consecutive frames. If the time domain data is smooth with little variation then frequency data will make low frequency data larger

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 1–10, 2007.

and high frequency data smaller. Hence, discrete cosine transform (DCT) is used to convert data in time domain to data in frequency domain. In order to remove any further statistical redundancy in the motion data and transformed coefficients, variable length coding (VLC) is used.

The first MPEG standard, MPEG-1 video [2] was developed for the specific application of video storage and playback on Compact Disks. MPEG-1 video was conceived to support the video CD, a format for consumer storage and playback that was intended to compete with VHS videocassettes. The standard uses block-based motion compensation, DCT and quantization and is optimized for a compressed video bitrate of around 1.2 Mbps. MPEG-1 video is still widely used for PC and web-based storage of compressed video files.

Following on from MPEG-1 video, MPEG-2 video [3] (ITU-T adopted it as H.262) standard aimed to support a large potential market, digital broadcasting of compressed television. MPEG-2 video was a great success, with world wide adoption for digital TV broad-casting via cable, satellite and terrestrial channels. For several years, MPEG-2 video has been improved, but it is reaching its theoretical limitations. Additional improvements were attempted, using other techniques, such as fractals and wavelets, with no significant improvement in video results. The original MPEG-4 Visual[4] standard attempted to bring the object-oriented perception into the compression world, with limited success, due to its complexity and overhead. In order to cover the very wide range of applications such as shaped regions of video objects as well as rectangular pictures, MPEG-4 Visual [4] standard was developed. This includes also natural and synthetic video/audio combinations with interactivity built in.

In an attempt to improve on the compression performance of H.261, the ITU-T working group developed H.263 [5]. This provides better compression than H.261, supporting basic video quality at bitrate of below 30 kbps, and is part of a suite of standards designed to operate over a wide range of circuit-switched and packet-switched networks. The coding algorithm used in H.263 is similar to that used by H.261, however with some improvements and changes to improve performance and error recovery. Half pixel based motion compensation technique is used and some parts of the hierarchical structure of the data stream is provided optionally. There are now four negotiable options included to improve performance: Unrestricted Motion Vectors, Syntax-based arithmetic coding, Advance prediction, and forward and back-ward frame prediction. After finalizing the original H.263 standard for video telephony in 1995, the ITU-T Video Coding Experts Group (VCEG) started working on a long-term effort to develop a new standard for low bitrate visual communications. This effort leads to the H.26L standard draft, offering significantly better video compression efficiency than previous standards [6].

The organization of the paper is as follows. We first explain several key features of H.264 in Section 2 and present the basic coding structure for MPEG-4 FGS and the scalable extension of H.264 (JSVM) which is the newest SVC standard in Section 3. In Section IV, we show two kinds of experimental results: One is related to the comparison of coding efficiency between H.264 and MPEG-4 visual and the other one is related to between MPEG-FGS and JSVM. Section V draws conclusions and summarizes future perspectives of video coding techniques for ubiquitous multimedia services.

2 Overview of H.264/AVC

H.264 video coding standard has been developed to satisfy the requirements of applications for various purposes, better picture quality, higher coding efficiency, and more error robustness. In this Section, we describe an overview of H.264.

2.1 Profiles and Levels

A Profile specifies a subset of entire bitstream of syntax and limits that shall be supported by all decoders conforming to corresponding Profile. There are three Profiles in the first version: Baseline, Main, and Extended. Baseline Profile is to be applicable to real-time conversational services such as video conferencing and videophone. Main Profile is designed for digital storage media and television broadcasting. Extended Profile is aimed at multimedia services over Internet. Also there are four High Profiles defined in the fidelity range extensions [7] for applications such as content-contribution, content-distribution, and studio editing and post-processing : High, High 10, High 4:2:2, and High 4:4:4. High Profile is to support the 8-bit video with 4:2:0 sampling for applications using high resolution. High 10 Profile is to support the 4:2:0 sampling with up to 10 bits of representation accuracy per sample. High 4:2:2 Profile is to support up to 4:2:2 chroma sampling and up to 10 bits per sample. High 4:4:4 Profile is to support up to 4:4:4 chroma sampling, up to 12 bits per sample, and integer residual color transform for coding RGB signal.

The Profiles have both the common coding parts and as well specific coding parts as shown in Fig. 1. For any given Profile, Levels generally correspond to processing power and memory capability of a codec. Each Level may support a different picture size - QCIF, CIF, ITU-R 601 (SDTV), HDTV, S-HDTV, D-Cinema [7]. Also each Level sets the limits for data bitrate, frame size, picture buffer size, etc [7].

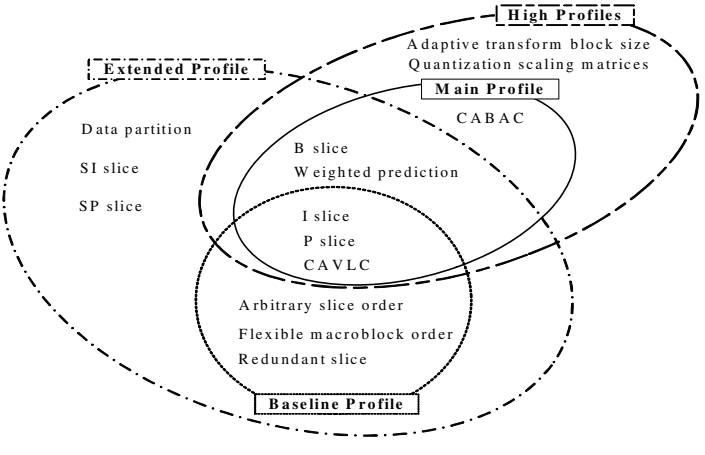


Fig. 1. Profiles and coding tools in H.264/AVC

2.2 Video Coding Algorithm

H.264 improves the rate distortion performance by exploiting advanced video coding technologies, such as variable block size motion estimation, multiple reference prediction, spatial prediction in intra coding, context based variable length coding (CAVLC) and context-based adaptive binary arithmetic coding (CABAC). The testing results of H.264/AVC show that it significantly outperforms existing video coding standards in both peak signal-to-noise ratio (PSNR) and visual quality [6].

For encoding a block or macroblock in intra-coded mode, H.264 predicts a block based on previously reconstructed blocks. The residual signal between the current block and the prediction is finally encoded. For the luminance samples, the prediction block may be formed for each 4x4 block, each 8x8 block, or for a 16x16 macroblock. One case is selected from a total of 9 prediction modes for each 4x4 and 8x8 luminance blocks; four modes for a 16x16 luminance block; and four modes for each chroma blocks.

Inter prediction is to reduce the temporal correlation with help of motion estimation and compensation. In H.264, the current picture can be partitioned into block sizes up to 4x4. For 16x16macroblock mode, there are four cases: 16x16, 16x8, 8x16 or 8x8, also four cases: 8x8, 8x4, 4x8 or 4x4 for 8x8 mode. Hence, the inter prediction process can form segmentations for motion representation as small as 4x4 block in size, using motion vector accuracy of one-quarter of the sample. Sub-pel motion compensation can provide significantly better compression performance than integer-pel compensation [7]. The process for inter prediction also involve the selection of the pictures to be used as the reference pictures from a number of stored previously-decoded pictures.

After inter prediction or intra prediction, the resulting prediction residual in a macroblock is split into small blocks according to the size of transform. H.264 uses also an adaptive transform block size, 4x4 and 8x8 (High Profiles only). In general transform and quantization require several multiplications resulting in high complexity for implementation. So, for simple implementation, the exact transform process is modified to avoid the multiplications. Then the transform and quantization are combined by the modified integer forward transform, quantization, scaling. For improved compression efficiency, H.264 also employs a hierarchical transform structure, in which the DC coefficients of neighboring 4x4 transforms for the luminance signals are grouped into 4x4 blocks and transformed again by the Hadamard transform. In order to utilize correlation among transform DC coefficients of neighboring blocks, the standard specifies the 4x4 Hadamard transform for luminance DC coefficients for 16x16 Intra-mode only, and 2x2 Hadamard transform for chroma DC coefficients.

Unlike fixed tables of variable length codes used in previous standards such as MPEG-1, 2, 4, H.261, H.262 and H.263, H.264 uses different VLCs in order to match a symbol to a code based on the context characteristics. In Baseline profile, all syntax elements except for the residual data are encoded by the Exp-Golomb codes and residual data is coded with more sophisticated entropy coding method called context-based adaptive variable length coding (CAVLC). In Main and High profiles, context-based adaptive binary arithmetic coding (CABAC) is can be used for all syntax elements including residual data. CABAC has more coding efficiency but higher complexity compared to CAVLC.

H.264 may suffer from blocking artifacts due to block-based transform in intra and inter prediction coding, and the quantization of the transform coefficients. The deblocking filter reduces the blocking artifacts in the block boundary and prevents the propagation of accumulated coded noise. H.261 has selectively suggested similar deblocking filter which was beneficial to reduce the temporal propagation of coded noise. However, MPEG-1, 2 did not use the deblocking filter because of high implementation complexity. However, H.264 uses the deblocking filter for higher coding performance in spite of implementation complexity. Filtering is applied to horizontal or vertical edges of 4 x 4 blocks in a macroblock. The luminance deblocking filter process is performed on four 16-sample edges and the deblocking filter process for each chroma components is performed on two 8-sample edges.

Table 1. Comparison of standards MPEG-2, MPEG-4 Visual and H.264

Feature	MPEG-2	MPEG-4 part 2	H.264/AVC
ME block size	8×8	16×16, 16×8, 8×8	16×16, 8×16, 16×8, 8×8, 4×8, 8×4, 4×4
Intra prediction	No	Transform Domain	Spatial Domain
Transform	8×8 DCT	8×8 DCT	8×8, 4×4 integer DCT 4×4, 2×2 Hadamard
Entropy coding	VLC	VLC	VLC, CAVLC, CABAC
Fractional ME	½-pel	¼-pel	¼ -pel
Reference picture	One	One	Multiple
In loop De-blocking filter	No	No	Yes
Picture types	I, P, B	I, P, B	I, P, B, SI, SP
Profiles	5 profiles	8 profiles	7 profiles
Transmission rate	2-15Mbps	64kbps - 2Mbps	64kbps - 150Mbps
Complexity	Medium	Medium	High

3 Scalable Video Coding

In ubiquitous environment, many challenges rise from the heterogeneity in multimedia client and server capabilities, and their end-to-end resource availabilities. For example, clients of a multimedia service may range from supercomputers to commodity PCs and smart handheld devices such as palm-tops. The network connections between the server and clients may range from high speed LANs to low speed dial-ups, from wire-line to wireless. Furthermore (and less addressed), even for clients with the same machine type and connection type, the amounts of resources available to each of them may still vary, depending on their location, workload, and the time they make service requests. In particular, the bottleneck resource in each client's resource requirement may be different. Therefore, to deal with the heterogeneity problem, any solution that only targets one specific type of bottleneck resource (for example, the network) may not be effective in all situations. Since scalable video coding (SVC) provides the capability of

reconstructing lower resolution or lower quality signals from partial bitstream, it is a good paradigm to the streaming video application for ubiquitous multimedia service.

Early video compression standards such as ITU-T H.261 [1] and ISO/IEC MPEG-1 [2] did not provide any scalability mechanisms. MPEG-2 was the first standard to include implementations of layered coding, where the standalone availability of enhancement information (without the base layer) is useless, because differential encoding is performed with reference to the base layer. All dimensions of scalability as mentioned above are supported (spatial, temporal, SNR); however, the number of scalable bitstream layers is generally restricted to a maximum of three in any of the existing MPEG-2 profiles. The video codec of the ISO/IEC MPEG-4 standard [5] provides even more flexible scalable profile called MPEG-4 FGS, including spatial and temporal scalability within a more generic framework, but also SNR scalability with fine granularity and scalability.

3.1 MPEG-4 FGS

As shown in Fig. 2, the basic information of the input signal is coded in the same way as the traditional block-based coding method in the base layer. In the enhancement layer, the residual signal that is not coded in the base layer is divided into 8x8 blocks and each block is DCT transformed. All the 64 DCT coefficients in each block are bitplane coded using four VLC tables [8].

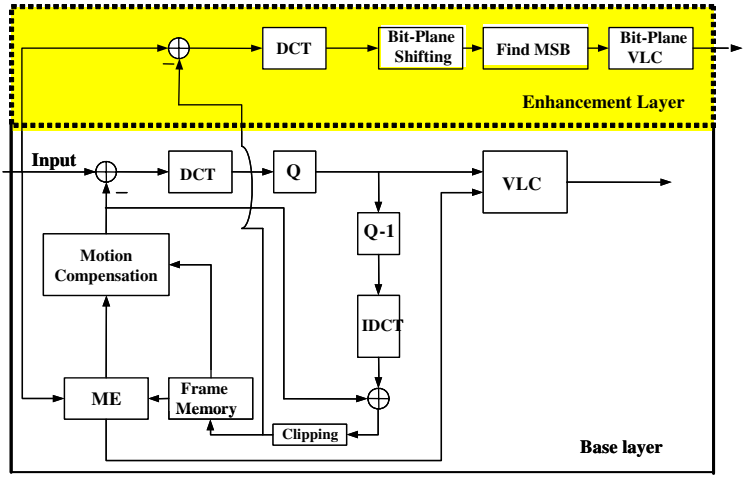


Fig. 2. MPEG-4 FGS encoder

There are many advantages of using FGS for Internet streaming video applications: it allows separation of encoding and transmission, the server can transmit enhancement layer at any bit rate without transcoding, it enables video broadcast on the Internet to reach a large audience, and it provides a solution to the video server overload problem.

However, compared with nonscalable coding, which is the upper bound for any scalable coding techniques, FGS is about 2-dB worse at the high end of the bitrate range.

3.2 Joint Scalable Video Model

In order to support fine granular SNR scalability, JSVM adopted progressive refinement (PR) slices [9]. Each PR slice is regarded as a FGS layer and coded with cyclical block coding as depicted in Fig. 3.

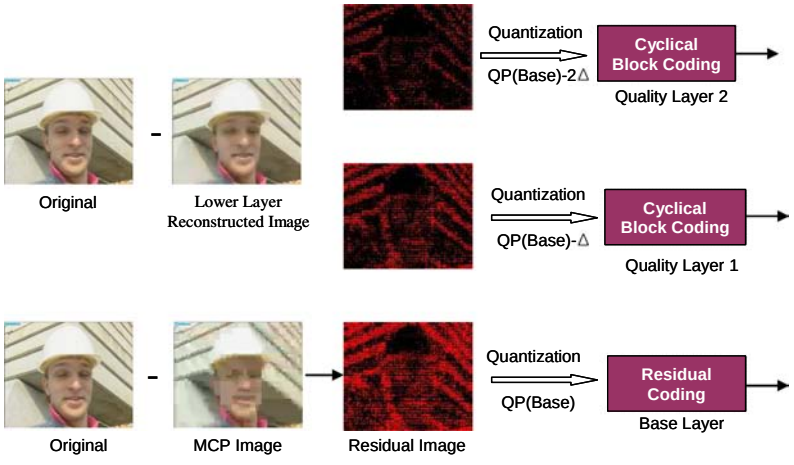


Fig. 3. FGS coding structure in JSVM

In the cyclical block coding scheme, the coding is basically partitioned into two passes, the significant and refinement passes. The significant pass first encodes the insignificant coefficients that have values of zero in the subordinate layers. Then, the refinement pass refines the remaining significant coefficients with range from -1 to +1. During the significance pass, the transform blocks are coded in a cyclical and block interleaved manner. On the other hand, the coding of the refinement pass is conducted in a suband-by-subband fashion [9] [10] [11].

In cyclical block coding, for each cycle, the coding of a block is continued until a non-zero coefficient in zigzag order is coded. Particularly, the coding of each cycle in a block includes an EOB symbol, a Run index and a non-zero quantization level. The EOB symbol is coded prior to the other symbols for signaling whether there are nonzero coefficients to be coded in a cycle. In addition, the Run index, represented by several significance bits, is used for recording the location of a non-zero coefficient. To further reduce the bit rate, each symbol is coded by a context-adaptive binary arithmetic coder [12] [13].

In Fig. 3, each FGS layer is represented as a group of multiple bit-planes. However, these bit-planes are coded by a cyclic block coding instead of traditional bit-plane coding used in MPEG-4 FGS. The coding order of transform coefficient levels has been

modified. Instead of scanning the transform coefficients macroblock by macroblock as it is done in the ”normal” slices, transform coefficient blocks are scanned in several paths, and in each path only a few coding symbols for a transform coefficient block are coded. Therefore, quality of the SNR base layer can be improved in a fine granular way. With the exception of the modified coding order, the CABAC entropy coding is reused, as specified in H.264/MPEG4-AVC [7].

4 Experimental Results

In the experiment, we first compare the coding efficiency between H.264 and MPEG-4 Visual. The basic test conditions for H.264 are set as follows:

1) MV search range is 16 pixels for CIF
2) RD optimization is enabled
3) Reference frame number equals to 1
4) GOP structure is IPPPP

Fig. 4 shows some comparisons of the coding efficiency of MPEG-4 Visual ASP and H.264 for the FOREMAN test sequence of CIF (352288) format. In these simulations, no rate control was used and Rate-Distortion (R-D) curves corresponding to encoding with different standards are presented. These are example plots and the results will vary from one encoder to another and from one test video sequence to another. From these plots we see that H.264 baseline profile provides about 2dB higher PSNR value over MPEG-4 Advanced Simple Profile.

In the second experiment, we have tested the coding efficiency of JSVM (version 5.2) according to the size of GOP. For GOP size of 2, we obtain the well-known prediction

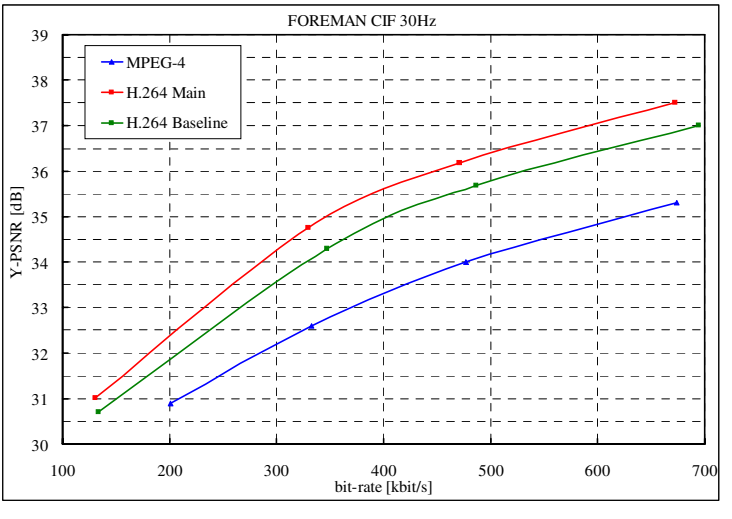


Fig. 4. Comparison of coding efficiency between MPEG-4 ASP and H.264

structure (IBPBPB) where one B-frame is encoded between two P or, alternatively, I-frames. For GOP size of 4, the coding structure of picture type is as follows: I B B B P B B B P. Fig. 5 shows some comparisons of the coding efficiency for GOP size of one, two, four, and eight. From Fig.5, we can know that the larger size of GOP provides better coding efficiency.

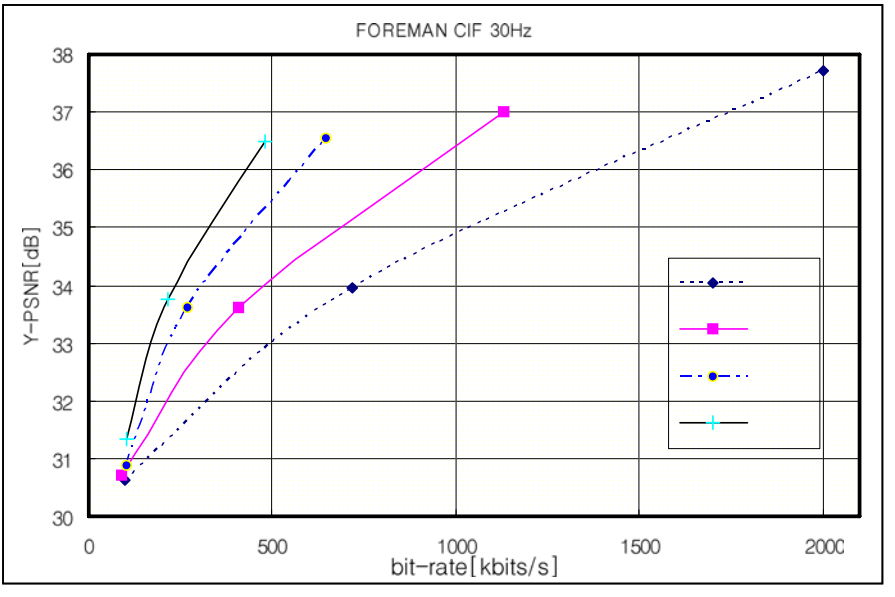


Fig. 5. Comparison of coding efficiency of JSVM according to GOP size

5 Conclusions

This paper gives a short overview over previous video coding standards. The new video standard known as H.264/AVC presents a rich collection of state-of-the-art video coding techniques and it can provide interoperable video broadcast or communication with degrees of capability that far surpass those of prior standards. Since the scalable extension of H.264 (JSVM) also provides fine granular scalable functionality and good coding efficiency, it is also a good paradigm to the streaming video application. Therefore, we believe these video coding technologies provide a powerful impact on ubiquitous multimedia services in the years to come.

Acknowledgements

This research was supported by MIC, Korea, under the ITRC support program supervised by IITA (IITA-2005-C1090-0502-0022) through the Realistic Broadcasting Research Center (RBRC) at the Gwangju Institute of Science and Technology (GIST).

References

1. H.261: International Telecommunication Union.: Video codec for audiovisual services at p x 64 kbit/s. ITU-T, (1993)
2. MPEG-1: ISO/IEC JTC 1.: Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 2: Video. ISO/IEC 11172 (2), (1993)
3. MPEG-2: ISO/IEC JTC1/SC29/WG11 and ITU-T.: ISO/IEC 13818-2: Informa-tion Technology-Generic Coding of Moving Pictures and Associated Audio In-formation: Video. ISO/IEC and ITU-T, (1994)
4. MPEG-4: ISO/IEC JTCI/SC29/WG11.: ISO/IEC 14 496:2000-2: Information on Technology-Coding of Audio-Visual Objects-Part 2: Visual. ISO/IEC, (2000)
5. H.263: International Telecommunication Union.: Recommendation ITU-T H.263: Video Coding for Low Bit Rate Communication. ITU-T, (1998)
6. Richardson I. E.G.: H.264 and MPEG-4 Video Compression. Wiley, (2003)
7. ITU-T Recommendation H.264 ISO/IEC 14496-10 AVC.: Advanced Video Coding for Generic Audiovisual Services. version 3, (2005)
8. Li, W.: Overview of fine granularity scalability in MPEG-4 video standard. IEEE Trans. on CSVT, vol. 11, no. 3, pp. 301-317, (2001)
9. Schwarz, H., Hinz, T., Kirchhoffer, H., Marpe, D., and Wiegand, T.: Technical Description of the HHI proposal for SVC CE1. ISO/IEC JTC1/SC29/WG11, Document M11244, Palma de Mallorca, Spain, Oct (2004)
10. Joint Video Team of ITU-T VCEG and ISO/IEC MPEG.: Scalable Video Coding-Working Draft 1. Joint Video Team, Document JVT-N020, Jan (2005)
11. Joint Video Team of ITU-T VCEG and ISO/IEC MPEG.: Joint Scalable Video Model JSVM. Joint Video Team, Document JVT-N021, Jan (2005)
12. Taubman, D.: Successive Refinement of Video: Fundamental Issues, Past Efforts and New Directions. Proc. SPIE, Visual Communication and Image Processing (2003) 649-663
13. Schwarz, H., Marpe, D., and Wiegand, T.: Scalable Extension of H.264/AVC. ISO/IEC JTC1/WG11 Doc. M10569/SO3, (2004)

Residual Signal Compression Based on the Blind Signal Decomposition for Video Coding

Sea-Nae Park[1], Dong-Gyu Sim[1], Seoung-Jun Oh[1], Chang-Beom Ahn[1], Yung-Lyul Lee[2], Hochong Park[1], Chae-Bong Sohn[1], and Jeongil Seo[3]

[1] VIA-Multimedia Center, Kwangwoon University
447-1, Wolgye-dong, Nowon-gu, Seoul 139-701, Korea
psea1118@kw.ac.kr
http://ipsl.kw.ac.kr
[2] DMS Lab., School of Computer Engineering, Sejong University, 98 Kunja-dong, Kwang-jin-gu, Seoul, Korea
[3] ETRI, Gajung-dong, Yusung-gu, Daejon, 306-700, Korea

Abstract. In this paper, a new residual signal compression method is proposed based on the blind signal decomposition for video coding. Blind signal decomposition is derived based on the fact that most of the natural signals in the real world could be decomposed into their basis signals and their weight values used in the composition process. In the proposed video coding system, composite data generated by adding two or more blocks are coded. Then the proposed decoder parses the coded bitstream and reconstructs the composite residual. The reconstructed composite residual is decomposed into the original residual blocks based on the blind signal decomposition. In the proposed system, the blind source separation is selectively used, depending on the performance of source separation. It is found that we can achieve approximately $2 \sim 3$dB gain by embedding our algorithm into an MPEG-4 baseline encoder.

Keywords: ICA, video, coding, BSD, residual coding, MPEG.

1 Introduction

Many video coding technologies have been standardized as MPEG-1/2/4 and H.264/AVC and widely used for many commercial multimedia applications [1]. Video compression is a key technology for better video quality with a constraint channel capacity. There have been many attempts in either enhancing video coding efficiency or adding new functionalities. However, all the coding standards are based on a hybrid motion-compensated transform coding, which is confronted with difficulties in improving the coding efficiency. These days, a model-based video coding technology has been proposed to resolve those problems. However, there is no new technology to significantly improve video coding efficiency [1],[2],[3].

In this paper, we propose a new residual signal compression method based on a blind source separation. The proposed video coding system encodes composite

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 11–19, 2007.

data generated by adding two blocks. Then the composite residual is reconstructed at a decoder side. The reconstructed composite residual is decomposed into the original residual blocks based on the blind signal decomposition. The latter is selectively used depending on the performance of source separation.

2 Independent Component Analysis (ICA) Algorithm

Independent component analysis is widely used as one approach for blind signal decomposition [4],[5]. It was suggested for solving the cocktail-party problem. At a cocktail party, there are mixed sounds consisting of peoples voices, music, and other types of noise. If we want to either talk with a friend or listen to music in the party, we need to separate either the friends voice or the music from the mixed sound. Human beings can do this without any problem. However, there is no perfect mathematical or computerized method to decompose all the source signals, called as basis signals. For the decomposition, we need to estimate not only the basis signals but also the mixture matrix that represents how the basis signals are mixed. Independent component analysis was hence proposed to find basis signals and a mixture matrix assuming that those basis signals are statistically independent [6],[7]. This approach was successfully ap-plied to many medical images and signal processing applications.

Let basis signals be denoted by $(s_1, s_2, \ldots, s_i)$ and $(x_1, x_2, \ldots, x_j)$ represent observed signals generated with a mixing matrix A. The observed signal can then be defined by

$$x_i = \sum a_{ij} s_j \tag{1}$$

where a_{ij} is each element of A. By replacing x_i, s_j, and a_{ij} with vector notations, $X = \{x_1, x_2, \ldots, x_j\}^T, S = \{s_1, s_2, \ldots, s_i\}^T$, and A. *Eq.* 1 can be denoted by

$$X = AS \tag{2}$$

We can obtain the mixture matrix W, by

$$W = A^{-1} \tag{3}$$

$$X = AS = W^{-1}S \tag{4}$$

and the basis signals can be computed by

$$S = WX \tag{5}$$

3 Proposed Video Compression Based on the BSD

Without any additional information, independent component analysis (ICA) can separate mixed signals by assuming that the basis signals are statistically independent. If the ICA can decompose the mixed signals into basis signals, the intentionally mixed signal would also be decomposed. Figure 1 is the block diagram for the proposed video coding system by applying the ICA algorithm to an

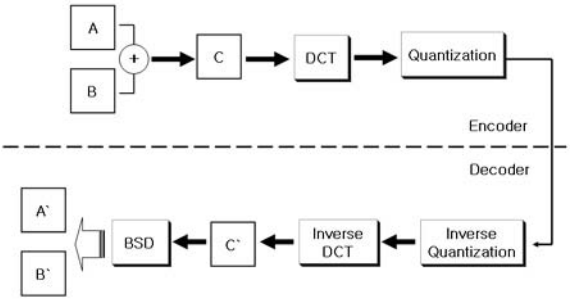


Fig. 1. The proposed ICA-based residual composition and decomposition

intentional composite residual signal. At the encoder side, the block $C(x, y)$ is composed by adding the blocks $A(x, y)$ and $B(x, y)$, where $A(x, y)$ and $B(x, y)$ are residual blocks by inter- or intra-predictions. Then C is coded by the conventional transform and quantization method, and the coded bits are sent to the decoder. The decoder reconstructs A'and B' blocks by decomposing the reconstructed C' block with the BSD. The blocks A and B are considered as basis signals, and the reconstructed A' and B' correspond to the original A and B blocks. The reconstructed blocks contain the error caused from the quantization and imperfection of the BSD. However, we can achieve coding efficiency by coding one composition block instead of two blocks. In this application, we know the mixture model of basis signals and how many signals are composed, so that the basis signals are likely to be accurately estimated.

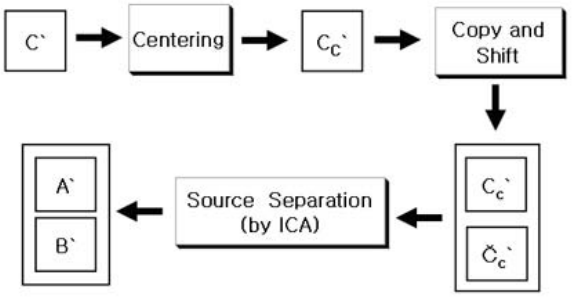


Fig. 2. Decomposition of the mixed block by the ICA

The decomposition procedure based on the ICA is performed with three steps, as shown in Figure 2. For the first step called centering, the reconstructed C_c' is compensated by the average of the block, and this is defined by

$$C_c{}'(x, y) = C'(x, y) - \frac{\sum_{x=0}^{X} \sum_{y=0}^{Y} C'(x, y)}{XY} \tag{6}$$

At the second step, $\check{C}_c$ should be generated by shifting $C'_c(x, y)$. For applying the ICA algorithm to the source separation, the dimension of the input mixed signal should be the same as that of the basis signals to be estimated. That is, the virtual mixed signal, $\check{C}_c$ is generated by copy and shift operations.

4 Proposed Video Compression Method by the Selective BSD

Figure 3 shows the block diagram of the proposed ICA-based video coding system. As shown in the figure, the residual blocks are mixed by adding the residual pixel values that are transformed and quantized. Then the entropy coded bitstream is decoded, and the mixed signal which is degraded by the quantization is reconstructed. The degraded mixed signal via quantization is decomposed into two residual blocks by the BSD. These two decomposed two residual blocks are similar to the original residual block. However, they are different not only because the composite block is quantized but also the BSD cannot be perfectly separated into its original sources.

Most of the conventional video coding systems including the MPEG-4 part 2 compress six 8×8 blocks per macroblock for the 4:2:0 format. One macro block consists of four 8×8 luminance blocks and two 8×8 chrominance blocks, as shown in Figure 4. The proposed algorithm should compress only three mixed blocks. Note that the transform coding of the H.264/AVC is conducted on 4×4 blocks.

As mentioned before, the separated blocks are not identical to the original residual blocks because two error sources introduced in the proposed video coding system. Figure 5 shows two residual basis blocks (S), and the composite and virtual composite blocks (X). In addition, S represents the separated signal block.

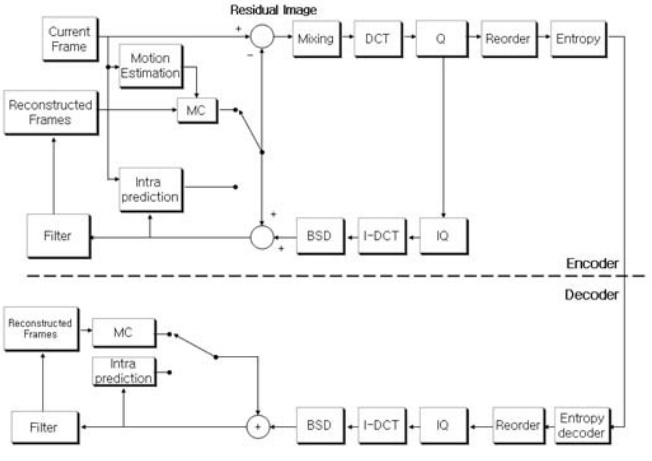


Fig. 3. The proposed ICA-based video coding system

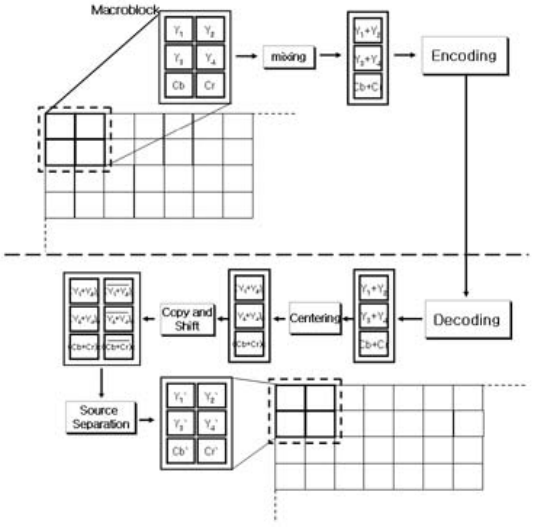


Fig. 4. The proposed block composition and decomposition

S looks similar to the original source signal but is not identical to it. The error is produced by the quantization, and the ICA can occasionally not perfectly decompose the source signals, depending on the characteristics of the source signals. Furthermore, ICA is based on the iterative statistical method so that it

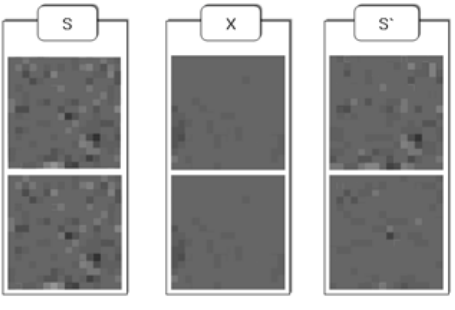


Fig. 5. The separated block S‘ and original block S

may not converge into the solution. Figure 6 shows the proposed selective ICA-based video coding system that selects the conventional transform coding or the ICA-based residual coding, depending on the reconstruction errors. Figure 7 shows the flowchart of the proposed decision flow whether or not the conventional DCT or ICA-based residual coding is used. At first, we need to calculate the pure ICA error by subtracting the error of the conventional DCT method from

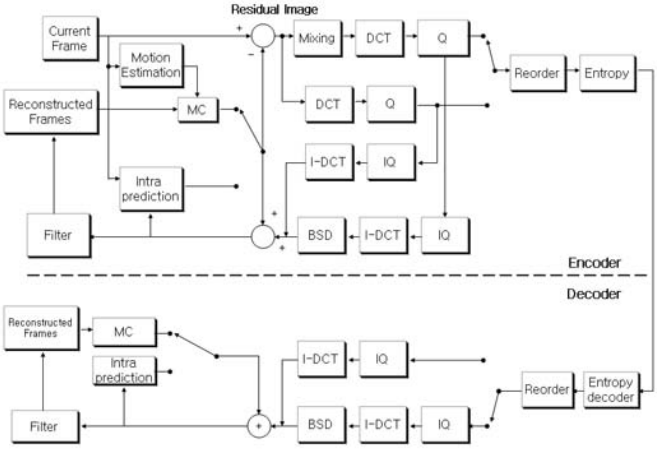


Fig. 6. The proposed selective ICA-based video coding system

the error of the ICA-based one. Then if this pure ICA error is larger than the threshold, the conventional DCT approach is activated. Otherwise, the encoder selects the proposed ICA-based transform coding. We need to send an indicating bit to represent whether the ICA-based coding is used or not.

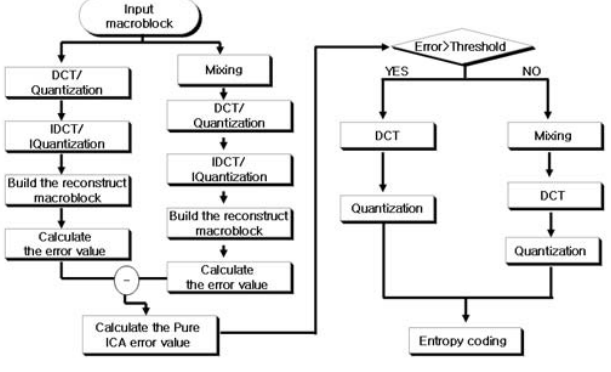


Fig. 7. Decision flow of the ICA-based mixed coding/conventional DCT-based coding

5 Experiment Results

The coding performance of the proposed system was evaluated with several standard video sequences and was compared with that of the MPEG-4 baseline video coding. We used "Stefan," "Mother_daughter," and "Mobile" sequences of

176 × 144 (QCIF). In our experiment, we set QP from 4 to 12, and the intra frame interval was set to 10.

Figure 8 shows several examples of reconstructed frames with the conventional and proposed coding methods. Figure 8(a) is the reconstructed image with MPEG-4 baseline, Figure 8(b) is the reconstructed image with the proposed algorithm applying the ICA mode to all the macro-blocks. Figure 8(c) shows the reconstructed images with the proposed selective ICA-based method. We found that the reconstructed images with the ICA-based method have annoying artifacts in subjective quality than those of the anchor. However, the proposed selective ICA-based method has comparable quality to the anchor algorithm with relatively smaller bitrate usage.

Fig. 8. Examples of the reconstruct images with MPEG-4, ICA-based, and selective ICA-based algorithms. (a) MPEG-4 baseline, (b) ICA-based method, (c) Selctive ICA-based method (Threshold = 2).

Table 1 shows the PSNR and the generated bits for each image of the sequences shown in Figure 8. In "Stefan" sequence, the bitrate of the ICA-based video coding method is one fourth of that of the MPEG-4 video baseline. However, we found that the PSNR drop is approximately 5dB. Because of the "Stefan" sequence has high motion activity so that the ICA is deteriorated and does not converge on a proper solution for several macro-blocks. As shown in the table, the proposed selective ICA-based algorithm yields almost the same quality with one third of the bitrate of the MPEG-4 video baseline by using the conventional video coding method for five macro-blocks. We also achieved significant gain for the rest of sequences. However, the selective ICA-based method has the best performance for all the sequences.

Figure 9 shows the PSNR in terms of bitrates for the three test sequences. We found that the proposed selective ICA-based algorithm can achieve approximately

Table 1. PSNR and generated bits for example image in Fig. 8

Sequence name	MPEG-4 baseline		ICA-based video coding		Selective ICA-based video coding (Threshold =2)		
	PSNR	Bits	PSNR	Bits	PSNR	Bits	Select ratio of ICA MB
Stafan	31.6	1879	25.1	588	30.8	605	95/99
Mother_daughter	35.8	550	35.8	110	35.8	350	69/99
Mobile	29.9	1957	26.4	572	29.5	708	98/99

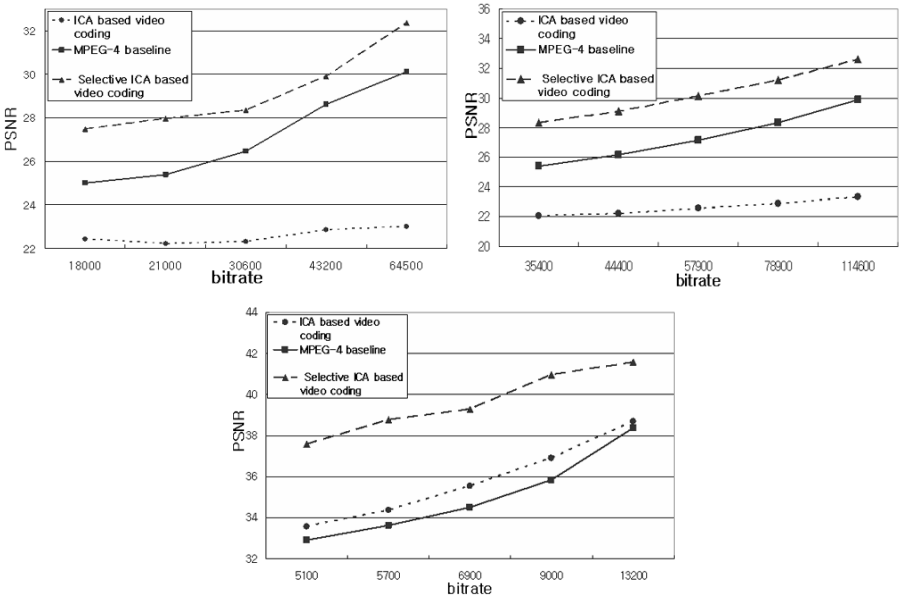


Fig. 9. Comparison of the RD performance for MPEG-4 and the proposed methods (a) Mother_daughter (b) Mobile (c) Stefan

$2 \sim 3$ dB gain by embedding the proposed algorithm into the MPEG-4 baseline encoder. Generally, the proposed method exhibits better performance for low-activity videos than high-activity videos. The selective ICA-based method yields the best perform-ance, but without the selective approach, it occasionally deteriorated for high-activity videos. For the "Mother_daughter" sequence, the ICA-based algorithm without the selective approach shows moderate RD performance because the sequence has low motion activity, resulting in a lower frequency of deterioration of the ICA.

6 Conclusion

In this paper, a new residual signal compression method was proposed based on the blind signal decomposition for video coding. The composite data of two or more residuals blocks were transformed and quantized. Then the composite block was reconstructed by inverse quantization and IDCT. The reconstructed composition block was decomposed into the residual blocks. The blind source separation was selectively used depending on the performance of source separation. We found that the proposed selective ICA-based video coding can obtain approximately $2 \sim 3$ dB gain, compared with the MPEG-4 video baseline. Further study will be focused on decreasing the amount of the pure ICA error by modifying the BSD algorithm with other a priori assumptions.

Acknowledgments. The present research has been in part conducted with the research grant of "Seoul R&BD Program" and "Future video/audio codec" project from ETRI.

References

1. Wiegand, T., Sullivan, G. J., Bjontegaard, G., Luthra, A.: Overview of the H.264 / AVC Video Coding Standard. IEEE Trans. on Circuits and Systems for Video Technology (2003) 560-576
2. Overview of the MPEG-4 Standard, MPEG-4 Overview. ISO/IEC JTC1/SC29/WG11 14496-10 N4668 (2002)
3. Coding of Moving Picture and Audio, Draft of Version 4. ISO/IEC JTC1/SC29/WG11 14496-10 (E) N7081 (2005)
4. Amari, S., Cichocki, A., Yang, H.H.: A New Learning Algorithm for Blind Signal Separation. Advances in Neural Information Processing Systems 8 (1996) 757-763
5. Bach, F. R., Jordan, M. I.: Kernel Independent Component Analysis. J. Machine Learning Res. 3 (2002) 1-48
6. Lee, T.W., Girolami, M., Sejnowski, T. J.: Independent Component Analysis using an Extended Infomax Algorithm for Mixed Sub-Gaussian and Super-Gaussian Sources. Neural Comput. 11 (1999) 417-441
7. Hyvarinen, A., Oja, E.: A Fast Fixed-Point Algorithm for Independent Component Analysis. Neural computation 9 (1997) 1483-1492

Personalized Life Log Media System in Ubiquitous Environment

Ig-Jae Kim, Sang Chul Ahn, and Hyoung-Gon Kim

Imaging Media Research Center, KIST,
39-1, Hawolgokdong, Seongbukgu, Seoul, Korea
{kij,asc,hgk}@imrc.kist.re.kr

Abstract. In this paper, we propose new system for storing and retrieval of personal life log media on ubiquitous environment. We can gather personal life log media from intelligent gadgets which are connected with wireless network. Our intelligent gadgets consist of wearable gadgets and environment gadgets. Wearable gadgets include audiovisual device, GPS, 3D-accelerometer and physiological reaction sensors. Environment gadgets include the smart sensors attached to the daily supplies, such as cup, chair, door and so on. User can get multimedia stream with wearable intelligent gadget and also get the environmental information around him from environment gadgets as personal life log media. These life log media(LLM) can be logged on the LLM server in realtime. In LLM server, learning-based activity analysis engine will process logged data and create meta data for retrieval automatically. By using proposed system, user can log with personalized life log media and can retrieve the media at any time. To give more intuitive retrieval, we provide intuitive spatiotemporal graphical user interface in client part. Finally we can provide user-centered service with individual activity registration and classification for each user with our proposed system.

Keywords: life log system, spatiotemporal interface, activity analysis.

1 Introduction

Recently, a large number of researches have been proposed for recording and retrieval for the information of personalized daily life due to the development of ubiquitous computing devices. We call these personalized media life log media(LLM). Life log media include the thing that one can see, the sound that one can hear, the information where one is, the state how one feels and so forth. Realtime retrieval of continuously captured personalized LLM will assist to enhance user's memory. To do this, we propose the new system with semi-automatic annotation technique based on our speech and activity analysis engine. We can give a cognitive assistants to people who want to organize their activity. To do this, we, first, record events using intelligent gadgets which is composed of wearable gadgets and environmental gadgets. Using captured LLM, we apply learning based activity classification technique based on multimodal analysis. After activity analysis, meta data for retrieval is created automatically. With

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 20–29, 2007.

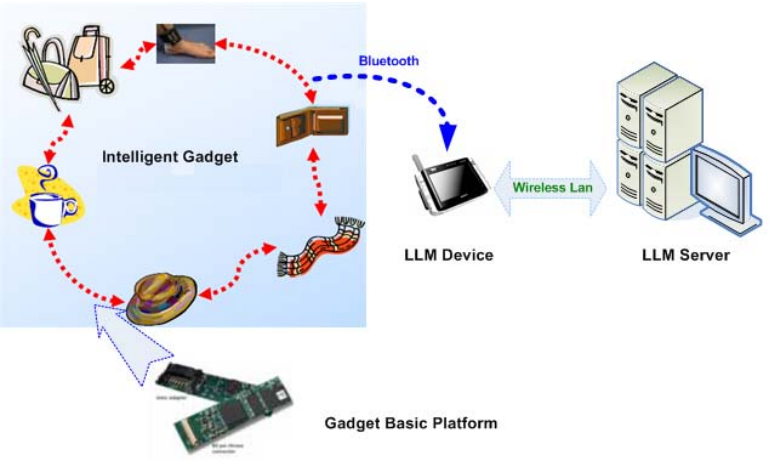


Fig. 1. Our system environment

described techniques, we can provide access to captured life log media at multiple levels of granularity and abstractions, using appropriate access mechanism in representations and terminology familiar to application users.

Our proposed system is composed of several parts, such as intelligent gadget, LLM device, LLM server and LLM browser. Figure 2 shows the relation of each component. In intelligent gadget, there are two components, wearable gadget and environmental gadget. Wearable gadget includes GPS, camera, microphone, body sensors, and HMD which give a functionality of I/O for P-LLM. Environmental gadget is built on the small and low power processing module and has a wireless interface like a zigbee or bluetooth. It is attached to our daily objects to give an information which is given to the LLM device of user. LLM device carried by user can get the information from intelligent gadget and send captured P-LLM to the LLM server at the idle time. In our LLM server, we can analyze the speech signal to identify speaker and classify speech from environment noise. We can, also, analyze the video signal to detect the registered objects and human faces. Besides, we can classify the pre-defined activity using multimodal sensor fusion technique. After analyzing LLM data in the server, Meta data are associated with A/V media data automatically and are used to retrieve. In our LLM client shown in Figure 3, we make a web-based browser with spatiotemporal query interface and tree-based activity search interface, such that user can query intuitively and see the retrieved results at their own LLM device. The rest of this paper is organized as follows. Section 2 reviews the related work in life log system. We present the whole scheme of our proposed system in section 3 and detail explanation about analysis for life log media data in section 4. Finally, we show the experimental results and conclusions in section 5, 6 respectively.

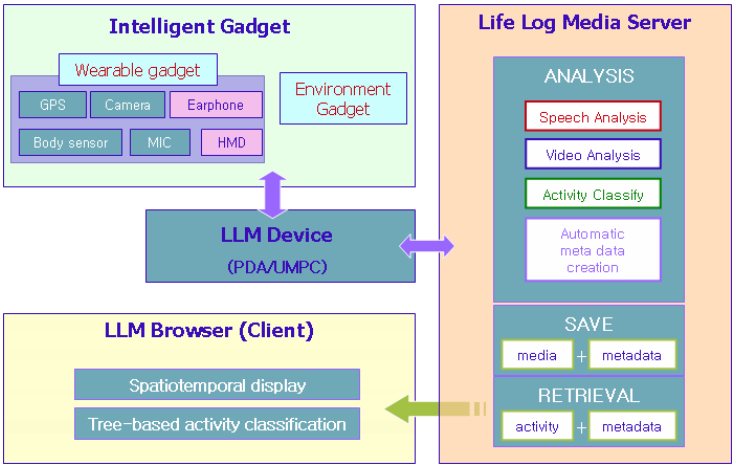


Fig. 2. Whole scheme for our proposed system

2 Related Work

There have been proposed several techniques for personal life log system . Gemmell et al. introduced SenseCam which is a device that combines a camera with a number of sensors[2]. Data from SenseCam is uploaded into a MyLifeBits repository, where a number of features, but especially correlation and relations, are used to manage the data. Mann described EyeTap which facilitate the continuous archival and retrieval of personal experiences, by way of lifelong video capture[3]. Vemuri et al. presented a method for audio-based memory retrieval[6]. They developed a pc based memory retrieval tool allowing browsing, searching, and listening to audio and associated speech-recognizer-generated transcripts. Aizawa et al. used audiovisual information as content to detect the conversation scenes and GPS data was applied as context to extract spatiotemporal key frames from time and distance sampling[1]. Tancharoen et al. extended their previous work including content based talking scene detection and context based key frame extraction using GPS data[9]. Recognizing general human activity or special motions is important key for automatic annotation. Recognizing general user activity has been tried by various authors. Randell et al. have done early investigations of the problem using only single 2-axis accelerometers[8]. Kern et al. presented a hardware platform to use multiple acceleration sensors that are distributed over the body[7]. They could capture 3-dimensional acceleration data from up to 48 positions on the human body. It is especially designed for robustness, allowing for recording even very dynamic activities, such as playing badminton or climbing. Chambers et al. focus on the recognition of complex gestures using Hidden Markov Models[4]. Kern et al. summarizes work on automatically annotating meeting recordings, extracting context from body-worn acceleration sensors alone, and combining context from three different sensors (acceleration, audio, location) for estimating the interruptability of the user[10].

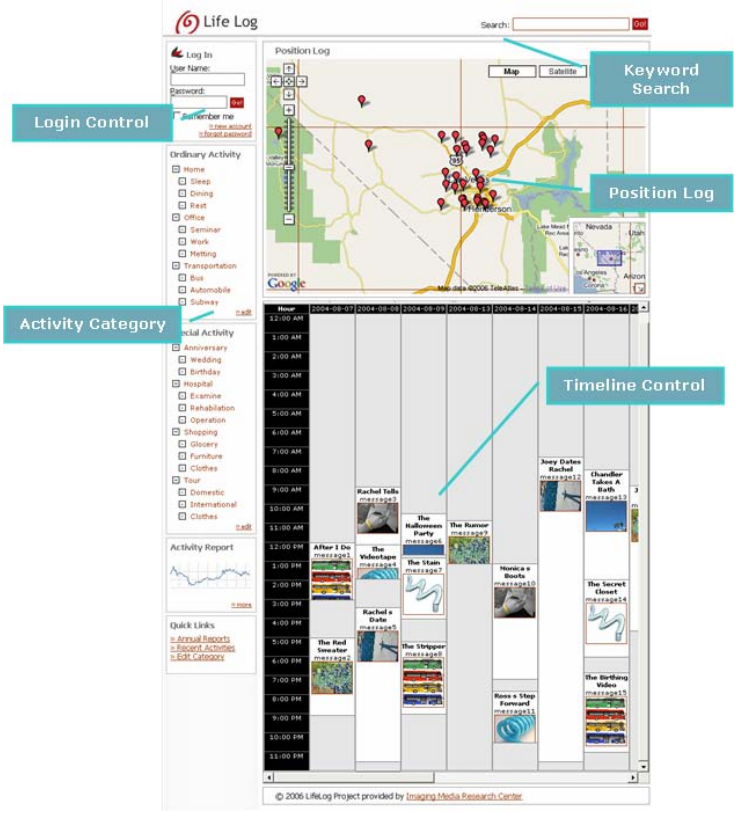


Fig. 3. Our web-based Life log media browser : User can query intuitively with spatiotemporal query interface and activity category

3 System Configuration

As shown in figure 2, our proposed system consists of four components. In intelligent gadgets, there are two modules, wearable gadgets and environment gadgets. User can capture the LLM data and show the retrieval results with wearable gadgets, such as camera, microphone, HMD and body sensors. These wearable gadgets are connected to the LLM device which user always carry. LLM device can also be connected to the environment gadgets which is embedded into the articles for daily use. In figure 1, we show the example of intelligent gadgets. We implemented the information gathering module and wireless communication module, such as zigbee or bluetooth onto the gadget basic platform and attached it to the objects for daily use. These intelligent gadgets are connected to the LLM device with wireless module and then LLM device can gathering the user's information in realtime. In fact, user's logging data must be sent to the LLM server in realtime, however, it is impossible to be connected to the server in everywhere. Practically, we use LLM device as a buffer for storing user's logging

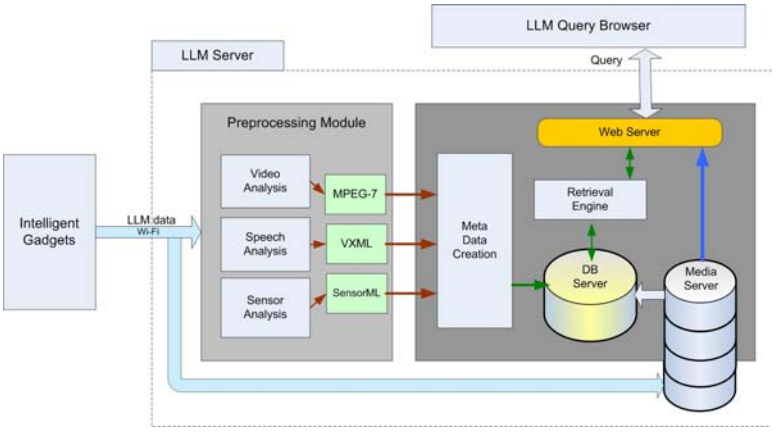


Fig. 4. LLM server system architecture

data temporarily in case the network connection is unavailable. And also, LLM device is used as a terminal for querying and browsing the retrieved results. After the LLM data being transmitted to the server, the LLM server analyze the user's activity.

We implemented the LLM server on the Windows XP. Figure 4 shows the architecture of LLM server. For web server and database server, we use IIS server, ASP.NET framework and PostgreSQL. In LLM server, user-dependent activity analysis can be done using multimodal data fusion technique, such as automatic video, speech and sensor data analysis. User can query and see the retrieved results through the LLM browser, which is served as web service and give user friendly spatiotemporal interface. In the following each section, more detailed explanation for each component will be given.

4 Activity Analysis

As mentioned in previous section, the captured LLM data from intelligent gadgets will be uploaded to the LLM server. After LLM being uploaded, the analysis will be started to classify the activity. Basically, our proposed system is based on the semi-automatic activity analysis. For the activity analysis, we use multimodal sensor fusion technique which is based on the speech analysis, video analysis and pattern classification from various sensors, such as accelerometer, gyro, physiological reaction sensors and environment sensors. In audio analysis, we extract the information of time, sex distinction, the number of speaker and captured environment as a meta data. Those information can be taken from pattern classification using several feature vectors such as MFCC, ZCR, Cepstrum energy and spectral differences. In video analysis, we used machine learning technique to detect the indoor location and registered objects. First, we select some pictures for specific area and familiar objects, such as corridor,

meeting room, monitors and so forth, and then put them into the training data set. We also processed on video stream to detect scene change using histogram matching method and identify the registered users using fisher classifier. From audiovisual analysis and action detection from smart sensors, we can transform them to LLM meta data according to the hierarchical meta data structure in figure 5.

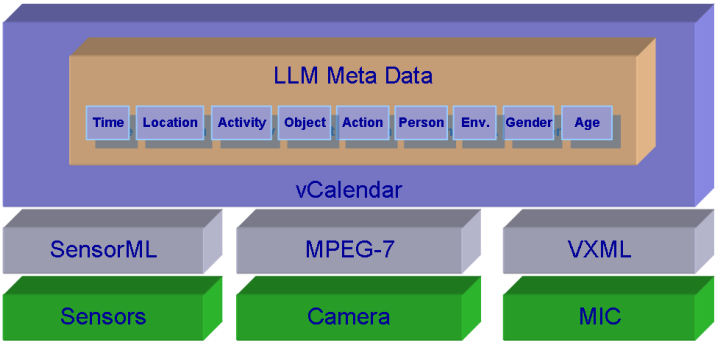


Fig. 5. LLM meta data structure

Multiple sensor data give a user's action information to classify the user's activity, such as lie down, running and fast moving. We use four 3D-accelerometers to detect the user's activity. These are attached above the each knee and on the each wrist of the user. Figure 6 shows the hierarchical sensor fusion process for LLM. We adopted bayesian analysis for multiple sensor fusion to classify the user's activity. An example of meta data creation using sensor fusion process is shown in Figure 7. However, it is impossible to define the user's whole activities automatically because the definition of activity is a sort of subjective evaluation, therefore even though the same sensor value will not be assigned the same category for the different users. For this reason, we define the general category for user's activity in advance, such as ordinary activity and extra-ordinary activity. In ordinary activity is related to the activity in home or office. Generally, the activities occurred outside of those area, they are classified as extraordinary activities. In addition to these pre-defined activities, users can add their own activity through our learning based structure.

To provide the user-dependent service, we need the definition for the individual activity for each user. To do this, as mentioned previous section, we let the user annotate personally at once for the repeated behavior with same objects and same time, as it is called *learning process*. After then, our activity inferring engine can automatically annotate for the same action. Although there are some intervention of user, when the user register the special activity on his/her own browser, in our activity analysis engine, it is more robust way for classify the user's activity in comparison with full-automatic activity classification method.

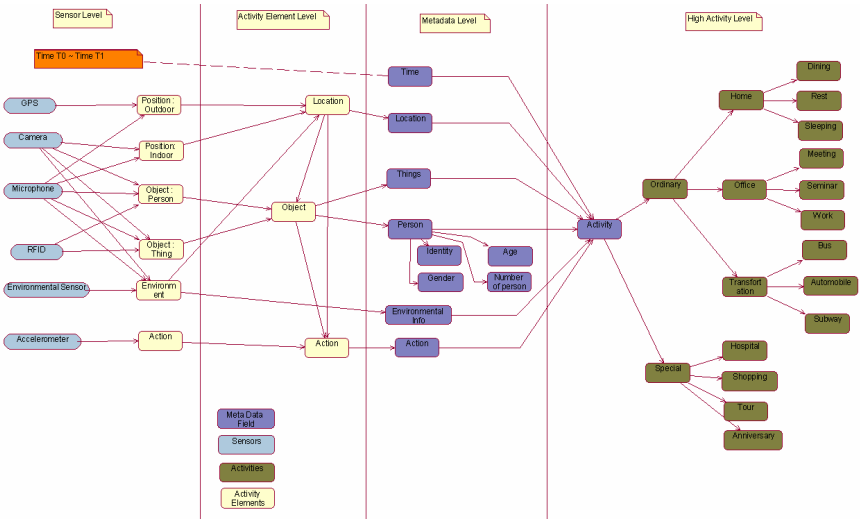


Fig. 6. Hierarchical sensor fusion process for LLM

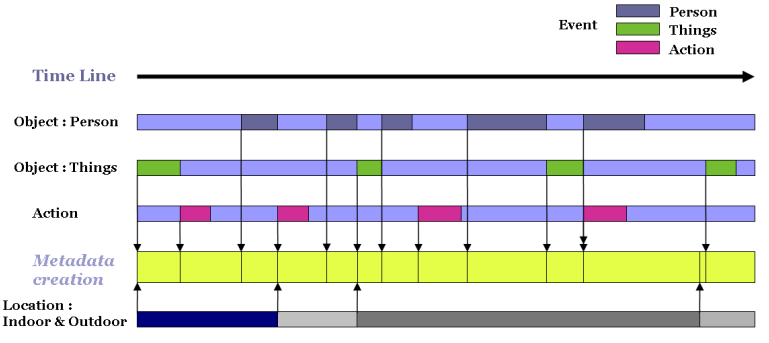


Fig. 7. Meta data creation from sensor fusion process

After learning process, we can calculate the probabilities of each sensors as a hypothesis for maximum likelihood estimation.

5 LLM Browser

After the automatic annotation is completed, users can search the AV data whenever they want through our LLM browser. In our LLM browser, we provide the user friendly graphical interface with spatiotemporal query interface and visualization, tree based menu selection method and categorized activity selection. Besides, users can access whenever/wherever they want as web service. We show

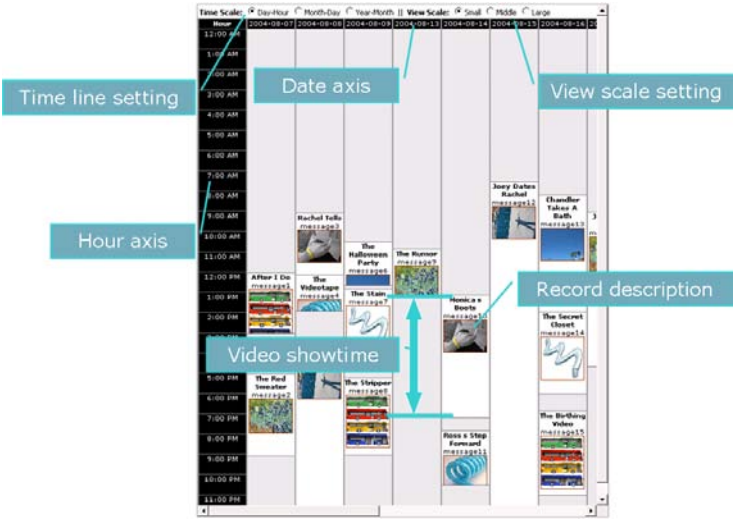


Fig. 8. Timeline control in LLM browser

the whole view of our LLM browser in Figure 3 and explain more specifically in following subsections.

5.1 Timeline Control

In LLM browser, we develop scalable timeline control interface which is shown in figure 8. In timeline setting, user can see the annotated description, corresponded video showtime in scalable view form. If user want to see larger thumnail image in time line windows, user can control the size of windows[Fig. 9]. As time goes by, user's data will increase, so that user cannot find his logged information at a glance. For more efficient search, we developed adaptive timeline control mechanism. User can select Day-hour, Month-day and Year-month pairs according to user's interest.

5.2 Map Control

We made our map control using Google Map API 2.0. There are several controls for zoom, position control, map overview and tooltip. Besides mentioned functionalities, we can display the user's logged position using custom overlay functionality. If the user's logged position can be transferred by LLM device automatically, server writes XML file which include the user's logged information and then our map control module reads the XML file to render points on the map. If user want to find logged information in some area, user can confine geographical area to select the boundary in the map by clicking the mouse button. If user click the title of map, LLM browser can show the record in timeline and

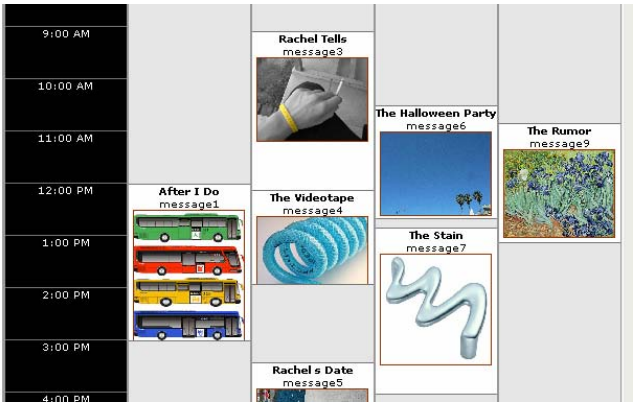


Fig. 9. View scale control in LLM browser

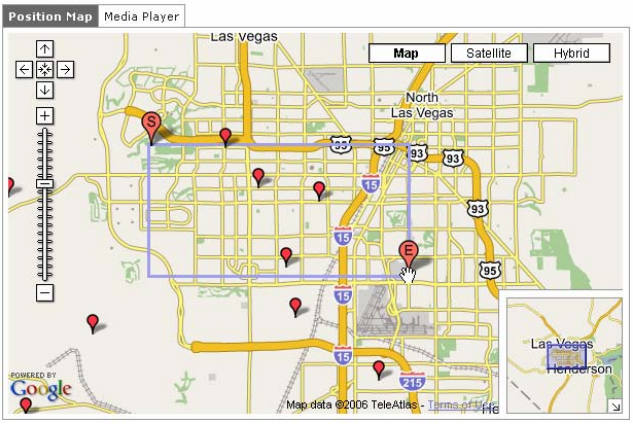


Fig. 10. Example of setting search boundary in the map

highlights it. In the same way, user click the title of timeline finds the record in the map, and also highlights itself as well.

6 Conclusion

In this paper, we present a new system for capturing and searching of life log media on networked environment. Our proposed system has four components, such as LLM device, Intelligent gadgets, LLM server and LLM client, which is connected on wireless network. LLM device is working as a gateway between intelligent gadgets and LLM server to connect them if network service is available. LLM device works as a terminal for querying and viewing the query results.

LLM server analyze the user's activity from stored LLM data. Intelligent gadgets which are attached to the daily supplies provide A/V data and multiple sensor data of user to be used for activity analysis and memory enhancement. LLM browser provides intuitive query interface to the user. For analysis of activity, we developed the learning based activity classification technique to annotate the captured LLM data. This learning based classification is a semi-automatic approach but we find it is more robust and adequate for user-dependent activity analysis. In this paper, we proposed a new platform for personal life log system. In this system, it is important that we can classify the user's activities accurately, therefore we have to investigate more robust activity classification technique in near future. We also, will make more compact and user-friendly LLM device with long battery life, which is basic problem to be solved.

References

1. Aizawa, K., Tancharoen, D., Kawasaki, S., and Yamasaki, T. 2004. Efficient Retrieval of Life Log Based on Context and Content. ACM Workshop CARPE 2004, 22–31.
2. Gemmell, J., Williams, L., Wood, K., Lueder, R., and Bell, G. 2004. Passive Capture and Ensuing Issues for a Personal Lifetime Store. ACM Workshop CARPE 2004, 48–55.
3. Mann, S. 2004. Continuous Lifelong Capture of Personal Experience with EyeTap. ACM Workshop CARPE 2004, 1–21.
4. Chambers, G., Venkatesh, S., West, G., and Bui, H. 2002. Hierarchical recognition of intentional human gestures for sports video annotation. In *In proceeding of IEEE Conference on Pattern Recognition. 2002*, Vol. 2, 2002, 1082–1085.
5. Torralba, A., Murphy, K., Freeman, T., and Rubin, M. 2003. Context-based vision system for place and object recognition. In *In Proceedings of International Conference on Computer Vision 2003*
6. Vemuri, S., Schmandt, C., Bender, W., Tellex, S. and Lassey, B. 2004. An audio-based personal memory aid. In *In Proceedings of Ubicomp 2004* Ubiquitous Computing, 2004, 400–417.
7. Kern, N., Schiele, B., and Schmidt, A. 2003. Multi-Sensor Activity Context Detection for Wearable Computing. In *In European Symposium on Ambient Intelligence 2003* 2003.
8. Randell, C., and Muller, H. 2000. Context awareness by analyzing accelerometer data. In *In proceeding of Fourth International Symoposium on Wearable Computers 2000*, 2000, 175–176
9. Tancharoen, D., Yamasaki, T., and Aizawa, K. 2005. Practical Experience Recording and Indexing of Life Long Video. ACM Workshop CARPE 2005, 61–66.
10. Kern, N., Schiele, B., Junker, H., Lukowicz, P., Troster, G., and Schmidt, A. 2004. Context Annotation for a Live Life Recording. In *In proceeding of Pervasive 2004*, 2004.

An Embedded Variable Bit-Rate Audio Coder for Ubiquitous Speech Communications

Do Young Kim[1] and Jong Won Park[2]

[1] Multimedia Communications Team, Electronics and Telecommunications Research Institute, 161 Gajeong-Dong, Yuseong-Gu, Daejeon, Rep. of Korea 305-700
dyk@etri.re.kr

[2] Department of Information Communications Engineering, Chungnam National University, 220 Gung-Dong, Yuseong-Gu, Daejeon, Rep. of Korea 305-764
jwpark@cnu.ac.kr

Abstract. In this paper, we propose an embedded variable bit-rate (VBR) audio coder to provide the fittest quality of service (QoS) and better connectivity of service for the ubiquitous speech communications. It has scalable bandwidth for narrowband to wideband speech signal, and embedded 8 32 kbit/s VBR corresponding to the network condition and terminal capacity. For the design of the embedded VBR coder, the narrowband signals are compressed by an existing standard speech coding method for the compatibility with G.729 coder, and then the other signals are compressed hierarchically on the basis of CELP enhancement and transform coding with temporal noise shaping (TNS) method. By the objective and subjective quality tests, it is shown that the proposed embedded VBR audio coder provides a reasonable quality compared with existing audio coders such as G.722 and G.722.2 in terms of mean opinion score (MOS) and perceptual evaluation of speech quality of wideband (PESQ-WB).

Keywords: Embedded Coder, G.729EV, MOS, PESQ-WB, Scalable Audio Coder, Ubiquitous Audio.

1 Introduction

The speech communications over the ubiquitous environment require better QoS than the existing telephony, better connectivity of service among various kinds of end points over the network, and the interoperability with the existing speech terminals. In this paper, we propose an embedded VBR audio coder so as to cope with the above requirements mainly for the ubiquitous speech communications. To meet the requirements, we consider the wideband speech coder which covers the full energy of human speech, and the embedded VBR audio coder in order to adapt its bit-rates dynamically from 8 to 32 kbit/s according to the variation of network especially between the different wireless networks and the capacity of the remote terminals [1]. Moreover, for backward compatibility with the popular narrowband speech coder used in the existing network, the proposed audio coder has a standard G.729 speech coder [2], [3], [4].

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 30–39, 2007.

Following this introduction, we will review the embedded VBR coder model that has been developed in the current state-of-the-art digital communication networks in Section 2. In Section 3, we will describe the structure of the proposed audio coder. In Section 4, we will evaluate the performance of the proposed audio coder by using the objective and subjective quality test method. Finally, we will present our conclusions in Section 5.

2 Embedded Audio Coder Model

Audible frequency range of human voice is from 20 to 20000 Hz. This audible frequency range can be divided into three parts, and we define these bands as narrowband (300∼3400Hz), wideband (50∼7000Hz), and audio band (20∼20000Hz). Because the energy of human speech is generally located in narrowband, narrowband speech coders have been developed and used. These speech coders started in PCM method [5], currently 8 kbps CS-ACELP which is standardized as ITU-T Recommendation G.729 is widely used [2], [3], [4]. With the advancement of network technology and internet service, many users have demanded the higher quality services and the research for wideband speech coder has been advanced. At present, G.722 [6] and G.722.2 [7] are widely used. These wideband coders have good performance for speech signals, but these cannot provide the embedded VBR functionality to give good connectivity over the IP network, as well as interoperability with the existing speech terminals. In this paper, we define that the embedded VBR coder is an audio coder that can generate variable bit-rates gracefully by the control of its application and provide scalable speech quality according to the changes of bit-rate by the structure of hierarchical bit-stream.

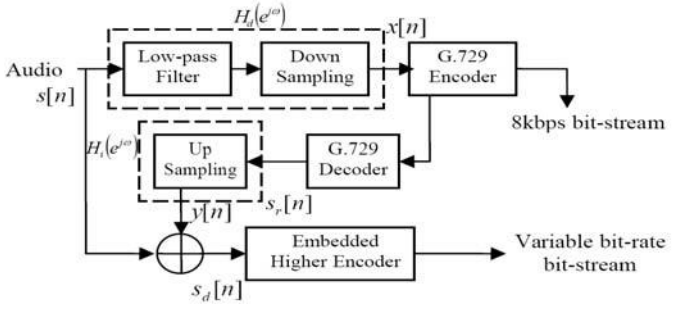


Fig. 1. A Block Diagram of the Embedded Audio Encoder Model

We propose a coder for the ubiquitous speech communications because its quality, media bandwidth, and interoperability can be controlled to cope with the requirements for the ubiquitous speech service. Fig. 1 and Fig. 2 show a block

diagram of the embedded VBR audio encoder and decoder model. Fig. 1 shows the block diagram of the embedded VBR audio encoder accomplished in the transmitter. First, a decimator $H_d(e^{jw})$ makes the input signal down-sampled to 8 kHz

$$X(e^{jw}) = H_d(e^{jw}) \cdot S(e^{jw}), \tag{1}$$

where $S(e^{jw})$ is the input signal and $X(e^{jw})$ is the down-sampled signal. $X(e^{jw})$ is the input signal to G.729 encoder. Then, G.729 encoder constructs 8 kbit/s bit-stream for transmitting the narrowband speech from 50 Hz to 3.4 kHz. In order to compute the remaining signal which is not coped with by G.729 encoder, the decoded speech signal is generated by the G.729 decoder, then it is up-sampled to adjust a sampling rate to the original signal by an interpolator $H_i(e^{jw})$.

$$Y(e^{jw}) = H_i(e^{jw}) \cdot S_r(e^{jw}), \tag{2}$$

where is the reconstructed signal by the G.729 decoder and is the up-sampled signal. The remaining signal,, is computed by the difference between the original signal and the up-sampled reconstructed signal as

$$s_d[n] = s[n] - y[n] \tag{3}$$

In this computation, we consider the delay from G.729 coder. That is, it is considered that 5 ms look-ahead, pre- and post-processing delay times in G.729 coder and processing times needed for decimation and interpolation [8]. After computing the remaining signal, the embedded high-layer encoder is performed. It generates variable bit-rate bit-stream. Fig. 2 shows the block diagram of em-

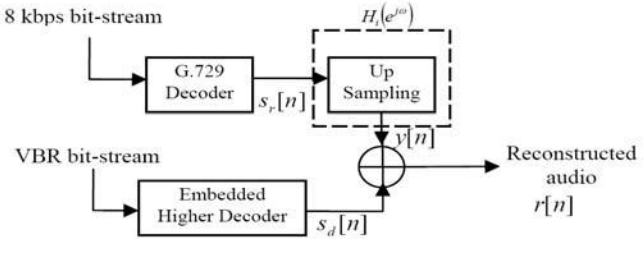


Fig. 2. A Block Diagram of the Embedded Audio Decoder Model

bedded VBR audio decoder accomplished in receiver. Transmitted bit-stream is divided into two parts, 8 kbps bit-stream and VBR bit-stream. 8 kbps bit-stream is decoded by G.729 decoder and VBR bit-stream is decoded by the embedded high-layer decoder, respectively. After up-sampling of the decoded signal by G.729 to adjust a sampling rate, two signals are merged and the complete signal $r[n]$ is reconstructed.

$$r[n] = x_r[n] + s_d[n], \tag{4}$$

where $x_r[n]$ is the reconstructed narrowband signal by G.729 decoder, and $s_d[n]$ is the reconstructed remaining signal by the embedded high-layer decoder. When we reconstruct the final signal, we consider the delay from G.729, the embedded high-layer coder, and an interpolator similar to computing Equation (4).

3 Proposed Embedded Variable Bit-Rate Audio Coder

To enhance the quality of voice for the ubiquitous applications, wideband speech codec technology was the first consideration for better quality of media source itself, since the performance of speech codec affects the quality of VoIP directly. And in order to provide robustness against the fluctuation of effective bandwidth over the ubiquitous network, we paid attention to the embedded VBR wideband speech codec described in Section 3. G.729 based embedded VBR coder(G.729EV) [9] was defined at SG16 of ITU-T. The main features of ToR of G.729EV coder are summarized in Table 1.

Table 1. Main Terms of Reference for G.729EV

Terms	Requirement
Core layer	G.729
Bandwidth	[300, 3400] ∼ [50, 7000] Hz
Sampling rate for input signal	16 kHz
Frame size	20 ms
Bit rates	8, 12∼32 kbit/sec
Granularity of bit-rates	2 kbit/s
Algorithm delay	< 60 ms
Complexity	< 40 WMOPS
Quality	Not worse than G.722@56kbit/s
Target application	Packet Voice(VoIP)

As shown in Table 1, the features of G.729EV provide high-quality internet telephony service for wireline and wireless networks providing two strong advantages. One is bit-level interoperability with legacy G.729 core codec, which is very popular for VoIP services in Asia, Europe, and North America. Also, since the frame size of G.729 is very short as 10ms, it gives the easier interoperability with the mobile phone. The second and main advantage of G.729EV for internet telephony over ubiquitous network is its scalability for the capacity of terminals and bandwidth. It adapts data rates according to the status of network from 8 and 12∼32kbit/s with the steps of 2 kbit/s.

3.1 Encoder

The block diagram of the proposed encoder is given in Fig. 3. 16 kHz wideband input is low-pass filtered and then down sampled to 8 kHz. This down-sampled

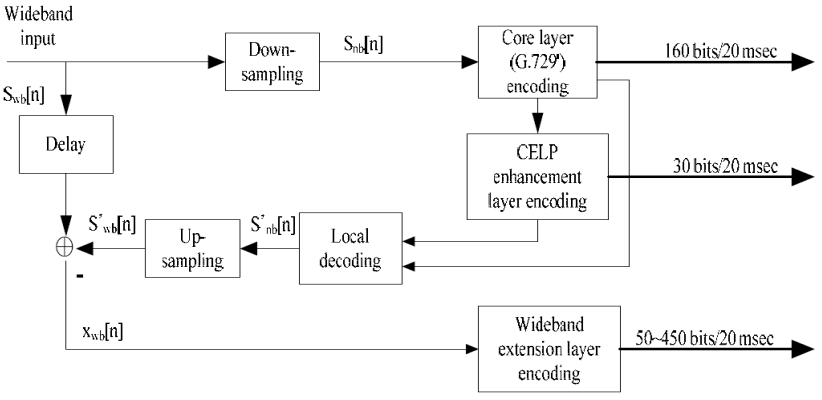


Fig. 3. Block Diagram of the Proposed Embedded VBR Audio Encoder

narrowband signal is encoded by the core layer and CELP enhancement layer. The core layer is based on ITU-T G.729 standard codec.

This layer is interoperable with ITU-T G.729 standard codec in the level of bitstream. The fixed codebook error signal of the core layer is processed in CELP enhancement layer in order to improve the quality of core layer. Thus the output of this layer is narrowband signal and the bit-rate is 1.5 kbit/s.

The difference signal between the delayed wideband input and up-sampled output of local decoder is processed in wideband extension layer. The difference signal is transformed using modified discrete cosine transform(MDCT). The coefficients are divided into several bands. The scale factor and normalized shape vector of each band are quantized respectively. The core layer is similar to G.729 standard codec except the LPC analysis window, pre-filtering, and post-filtering. The pre-filtering and post-filtering are suppressed. The length of the cosine part and the center location of the LPC analysis window are changed [10] and the look-ahead size is increased from 5 ms to 10 ms. The CELP enhancement layer is designed to improve the quality of core layer. In this layer, the fixed codebook error signal of core layer is represented by two algebraic pulses in every 10 ms. Signs and positions of the pulses are quantized with 15 bits. The pulses are scaled with the fixed codebook gain of core layer.

An input signal X_{wb}[n] of the wideband extension layer in the Fig. 3 is the difference between the delayed wideband input signal and the up-sampled version of locally decoded narrowband signal, and the signal is processed on every 20 ms frames. X_{wb}[n] is transformed first using MDCT. The MDCT is performed on 40 ms windowed signal with 20 ms overlap. The MDCT coefficients, X(k), is split into two typical bands, one for [0, 2.7 kHz] and the other for [2.7, 7.0 kHz]. The coefficients of the first band are quantized in MDCT domain and the coefficients of the second band are quantized on Linear Predictive Coding(LPC) residual domain. Finally, all of the quantized parameters are encoded and packed into a bitstream at the bit-packing block according to the predefined order.

3.2 Decoder

The decoder also comprises three layers: core layer, CELP enhancement layer and wideband extension layer as shown in Fig. 4. The operation of each layer

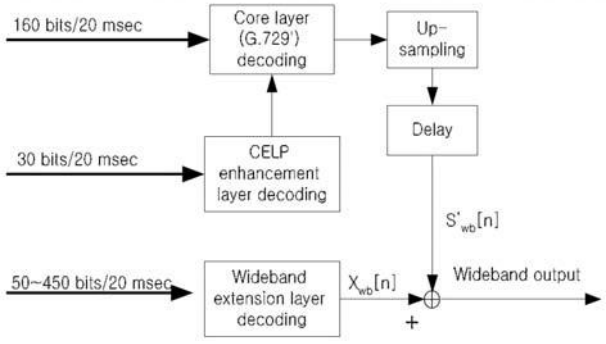


Fig. 4. Block Diagram of the Proposed Embedded VBR Audio Decoder

depends on the size of the received bit stream. A frame erasure concealment algorithm is also applied in order to improve the synthesized quality in frame erasure condition. The frame erasure concealment algorithm of core layer is partly modified based on a state machine [10]. The pitch gain and fixed codebook gain is reconstructed by an attenuated version of the previous pitch gain and fixed codebook gain respectively. The attenuation coefficient depends on the state. In the case of voiced frame, the fixed codebook gain is set to zero. In wideband extension layer, an erased frame is recovered by multiplying a randomly generated shape vector by the attenuated scale factor of the previous frame.

4 Performance Evaluation

We evaluate the performance of the proposed embedded VBR audio coder in terms of the quality, algorithmic delay, complexity, and the size of memory. The quality of proposed coder was evaluated formally by ITU-T subjective [11] and objective test methods [12]. The complexity was calculated by weighted million operations per second(WMOPS) defined in ITU-T P.191 [13].

4.1 Quality Evaluation

In order to evaluate the speech quality, we perform the subjective and objective tests for obtaining MOS and PESQ-WB scores which are formally defined each in the ITU-T P.800 and P.862.2 recommendations. Table 2 shows MOS score for narrowband speech, and compares its quality with the existing G.729A coder in order to evaluate the enhancement of quality.

Table 2. MOS Scores for narrowband speeches

Coder	Male	Female	Mean
Direct	4.375	4.229	4.302
G.729A	3.667	3.688	3.678
Proposed@8k	4.042	3.938	3.990

Table 3 shows the MOS score for the wideband speech, and compared its quality with the existing G.722 at 48 and 56 kbit/s, and G.722.2 at 8.85 kbit/s coder in order to evaluate the quality.

Table 3. MOS Scores for the wideband speeches

Coder	Male	Female	Mean
Direct	4.521	4.563	4.542
G.722.2@8.85k	4.063	3.917	3.990
G.722@48k	3.896	3.875	3.885
G.722@56k	4.146	4.125	4.135
Proposed@14k	4.313	4.354	4.333
Proposed@24k	4.313	4.188	4.250
Proposed@32k	4.458	4.271	4.365

Table 4 shows the mean value of PESQ-WB scores of the proposed audio coder in order to evaluate the linear quality enhancement of the embedded VBR audio coder for wideband. For the objective test, we use 5 languages; Korean, French, Japanese, German, and English.

Table 4. PESQ-WB Scores for the Proposed Coder

Bit-rates(kbit/s)	PESQ-WB Score	Bit-rates(kbit/s)	PESQ-WB Score
14	3.28	16	3.29
18	3.37	20	3.55
22	3.58	24	3.60
26	3.61	28	3.62
30	3.62	32	3.69

Table 5 also shows better quality of the proposed audio coder compared with reference coders at -16, -36 dB signal levels and noise conditions. All of the experiments involved four talkers (two males/two females), three samples per talker, and three panels of 8 listeners each (24 listeners total).

The result concludes that the proposed audio coder has 10.85% better quality than the existing G.729A coder at the same 8 kbit/s in terms of MOS score.

Table 5. MOS Scores at different signal levels and noise conditions

Proposed		Reference	
8k (-16dB)	3.760	G.729A@8k (-16dB)	3.688
8k (-36dB)	3.896	G.729A@8k (-36dB)	3.500
32k (-16dB)	4.188	G.722A@56k (-16dB)	4.344
32k(-36dB)	4.292	G.722A@56k (-16dB)	3.458
8k (Music)	4.240	G.729A@8k (Music)	4.219
8k (Office)	4.198	G.729A@8k (Office)	4.083
8k (Babble)	4.625	G.729A@8k (Babble)	4.354
32k (Music)	4.396	G.722A@56k (Music)	3.969
32k (Office)	3.396	G.722A@56k (Office)	4.219
32k (Babble)	4.458	G.722A@56k (Babble)	4.188

At 14kbit/s mode of the proposed coder which is the minimal bit-rate of wideband, it shows 10.86% quality enhancement than G.722.2 at 8.85kbit/s. And at 32kbit/s mode of the proposed coder, it shows 10.86∼11.24% quality enhancement than G.722 at 48 and 56 kbit/s.

4.2 Algorithm Delay

The algorithm delay of the proposed audio coder is 40.75ms, which comprises 20 ms framing delay, which is the same value of frame size of the coder, 10ms look-ahead, 10ms MDCT overlapping window, and 0.75ms up/down sampling delay.

4.3 Complexity and Memory

To calculate the complexity for the implementation of the proposed coder, we use the WMOPS which are defined in ITU-T P.191 recommendations. The complexity and the size of memory are summarized in the Table 6 and Table 7 respectively. In table 6, the complexity is evaluated in the worst case. The total values are given by the sum of the three layers and other functions such as resampling. The overall complexity of the proposed codec is about 37.85 WMOPS.

Table 6. The Worst Case Computational Complexity(WMOPS) of the Proposed Coder

Components	Encoder	Decoder
Core layer	11.683	2.612
CELP enhancement layer	5.707	0.042
Wideband enhancement layer	9.692	4.265
Other functions	2.519	1.372
Total	29.601	8.291

Table 7. Memory Requirement of the Proposed Coder(Word)

Memory Types	Encoder	Decoder	Total
PROM	3,704	2,943	6,647
DROM	22,865		22,865
DRAM	4,295	3,897	8,192
Total	37,704		

The DROM takes into account all the constant tables. Same tables are used in both encoder and decoder. Thus the DROM in table 6 is the summation of the encoder and decoder. The DRAM corresponds to the memory of all the static variables and worst case of the dynamic RAM usage.

5 Conclusion

In this paper, we have proposed an embedded VBR audio coder in order to provide the fittest quality of service and better connectivity of service for the speech communications over the ubiquitous network environment. After dividing input signal into narrowband and wideband, it performs coding procedure in each part hierarchically and the bit-rate for providing the fittest quality between the ubiquitous end points is determined dynamically according to the channel conditions and terminal capacities. This embedded VBR architecture of the proposed audio coder provides the fittest speech quality of service, better connectivity of service, and excellent service completion ratio among the various ubiquitous end points. Therefore, the proposed audio coder is useful for high-quality speech communications such as voice over IP, conversational e-learning, audio conferencing, remote monitoring, conversational internet games, and other multimedia services over the ubiquitous network. For the interoperability with legacy voice terminal, the proposed audio coder has the compatibility with existing G.729 coder. Moreover, it uses G.729 enhancement coder to improve quality of the narrowband signal and to provide the basis of better quality of the higher band. This higher band, which covers all bandwidth of human speech, provides better quality compared with the voice quality of analog telephones. The merit of the interoperability of the proposed coder enables wider reuse of the existing voice over IP systems such as G.729/G.729a terminals and gateways. As a result, the proposed coder has a reasonable performance compared with the existing wideband audio coder by the subjective and objective evaluation measures of speech quality.

Acknowledgments. Authors acknowledge the contribution of Jongmo Sung, Mi Suk Lee, and Hyun-Woo Kim in the development of the embedded VBR coder. The work was supported by the Ministry of Information and Communication, Republic of Korea.

References

1. Do Young Kim, Mi Suk Lee, H.W.J.H.K.K.: Scalable speech and audio coding technologies for wireless network. In: Proc. of KICS. Volume 22., Seoul, KICS, KICS (2005) 1397–1407
2. G.729: Coding of speech at 8kbps using conjugate-structure algebraic-code-excited linear-prediction (cs-celp). In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
3. G.729A: G.729 annex a: Reduced complexity 8 kbit/s cs-acelp speech codec. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
4. G.729B: G.729 annex b: A silence compression scheme for g.729 optimized for terminals conforming to recommendation v.70. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1996)
5. G.711: Pulse coded modulation(pcm) of voice frequencies. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1988)
6. G.722: 7 khz audio coding within 64 kbit/s. In: ITU-T Recommendation, Geneva, ITU, ITU-T (1988)
7. G.722.2: Wideband coding of speech at around 16kbit/s using adaptive multi-rate wideband (amr-wb). In: ITU-T Recommendation, Geneva, ITU, ITU-T (2002)
8. G. H. Lee, Y. H. Lee, H.K.K.D.Y.K., Lee, M.S.: A scalable audio coder for high-quality speech and audio services. In: Proc. of the 9th Western Pacific Acoustics Conference, Seoul (2006) 178–185
9. ITU-T: Q10/16 meeting report, Geneva, ITU, ITU-T (2004)
10. ITU-T: High-level description of etri candidate codec for g.729ev, Geneva, ITU, ITU-T (2005)
11. P.800: Methods for subjective determination of transmission quality, Geneva, ITU, ITU-T (1996)
12. P.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs, Geneva, ITU, ITU-T (2005)
13. P.191: Software tools for speech and audio coding, Geneva, ITU, ITU-T (1993)

Performance Enhancement of Error Resilient Entropy Coding Using Bitstream of Block Based SPIHT

Jeong-Sig Kim and Keun-Young Lee

Image Communication Lab., School of Information and Communication Engineering, SungKyunKwan University, 300 Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, Republic of Korea
condor@mickey.skku.ac.kr,kylee@ece.skku.ac.kr

Abstract. To provide the bit stream with some level of protection against channel errors, standard image and video techniques add a controlled amount of redundancy. This redundancy may take the form of resynchronization makers which enable the decoder to restart the decoding process from a known state in the event of transmission errors. The Error Resilient Entropy Code(EREC) is a well known algorithm designed to reduce the added redundant information. We propose a more error robust algorithm, ERREREC, for the bit stream of DCT and SPIHT based coding techniques, which greatly improves its ability to maintain the compressed image quality in the event of random errors. The simulation results of proposed algorithm shows that the quality of transmitted image is improved for the principal coding techniques.

Keywords: Error Resilient Entropy Coding, Set Partitioning in Hierarchical Tree, VLC(Variable-Length Code), Wavelet Transform, DCT.

1 Introduction

One inherent problem with any communications system is that information may be altered or lost during transmission due to channel noise. This effect of such information loss can be devastating for the transport of compressed image and video.

Transmission errors can be roughly classified into two categories: random bit errors and erasure errors. Random bit errors are caused by the imperfections of physical channels, which result in bit inversion, bit insertion, and bit deletion. When fixed-length coding is used, a random bit error will only affect one code word. But if VLC is used, random bit errors can desynchronize the coded information such that many following bits are undecodable until the next synchronization code word appears. Erasure errors can be caused by packet loss in packet networks due to physical defects, or system failures for a short time. Random bit errors in VLC can also cause effective erasure errors since a single bit error can lead to many following bits being not decodable and hence useless. Therefore, there is no need to treat random bit errors and erasure errors separately.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 40–49, 2007.

Many error handling techniques have been suggested, such as Layered Coding[1], Forward Error Correction(FEC)[2], Multiple Description Coding(MDC)[3], Error-Resilient Coding[4],[5],[6] and so on. The layer coding ought to guarantee the delivery of the most important base layer over channel, since a loss in base layer can lead to a disastrous defect in the decoded visual quality. FEC do not work well in highly compressed digital image and video, since FEC uses lots of redundancy bits. The source coder with MDC assumes that all coded bits will be treated equally and that all are subject to loss in parallel channels. And the relatively overhead associated with MDC is appropriate only for channels that have relatively high loss or failure rates. Error-resilient coding which reduces redundancy due to channel coding, nonetheless protects against error propagation. An error-resilient coding technique for image and video transmission has been proposed which uses a bit rearrangement technique in Error Resilient Entropy Coding (EREC)[4]. The EREC algorithm has attracted considerable attention, because it allows the spatial propagation of transmission errors to be significantly improved without any sizeable overhead. The ERREC(Efficient and Robust EREC)[6] algorithm described in our previous work significantly improves on the EREC[4] algorithm, by considering the statistical distribution of the long and short blocks.

In this paper, we propose a new method which combines the ERREC[6] algorithm with the SPIHT[7] algorithm, in order to accomplish better error resilience. We use the structures of the spatial orientation tree[7], in which each tree is coded independently, and group the wavelet coefficients according to the corresponding spatial blocks. By doing so, a highly efficient source coding can be obtained in order to search for self-similarity fully in the coefficients across the different scale-levels of the wavelet transform. Then, EREC is applied to reorganize these variable-length data bits into fixed length slots for transmission over erroneous channels. At the receiving end, the start bit of each slot can be automatically determined in order to synchronize the data bits. Error propagation has less noticeable effects on bits located at low frequency band.

We simulated the proposed algorithm, ERREC, for the bit stream of DCT and SPIHT based coding techniques, which greatly improves its ability to maintain the compressed image quality in the event of random errors. The simulation results of the proposed algorithm will show that the quality of transmitted image is improved for the two principle coding techniques.

2 Error Resilient Entropy Coding

Error resilient entropy coding(EREC) was originally proposed in [4] to handle the sequential transmission of variable-length coded DCT data blocks over noisy channels. In the case of DCT based block image coding schemes such as JPEG, H.263 etc., the length of the coded binary bits in one block is generally different from those of the other blocks. The key idea behind the EREC is the re-organization of the variable-length data blocks into fixed slots.

The EREC scheme allows for resynchronization at the start of each block without the additional overhead of inserting redundant resynchronization markers. It does this by cleverly rearranging the existing blocks so that they fit into a predetermined number of fixed length slots. The basic operation of the EREC is to rearrange the M variable-length blocks (of length b_i) of data into fixed-length slotted structure (M slots with length S) in such a way that the decoder can independently find the start of each block and start decoding it. The encoder first chooses a total data bits length size Ltotal which is sufficient to code all the data. Therefore, total bits length L_{total} of M slots is given by eq.(1). The value L_{total} needs to be coded as a small amount of protected header information.

$$L_{total} = \sum_{i=1}^{M} S_i \geq \sum_{i=1}^{M} b_i \tag{1}$$

where b_i is variable-length bits of each block. Overall variable-length bits of image M blocks do not exceed the total slot length L_{total} bits. Slot length S is determined as the average variable-length code bits of image M blocks given eq.(2).

$$S = \left\lceil \frac{1}{M} \left(\sum_{i=1}^{M} b_i \right) \right\rceil \tag{2}$$

where $\lceil \bullet \rceil$ is the ceiling function as integer value.

Next the encoder splits L_{total} into M slots of length S chosen to be even approximately. An $N-$stage algorithm is used to place the variable-length blocks of data into each of the fixed-length slots. At each stage k, a block i with left uncoded data searches in slot j. Therefore this relates to offset (Φ_k) at $k-$stage which is searching $j = i + k(mod\ M)$ for space to code some or all of the remaining data from block i. Φ is an offset sequence.

3 Set Partitioning in Hierarchical Tree

Set partitioning in hierarchical tree(SPIHT) image coding algorithm was developed in 1996 by Said and Pearlman[7] and is another more efficient implementation of the embedded zero tree wavelet (EZW) algorithm. The wavelet transform using 9/7 tap wavelets is applied to an image.

The main algorithm works by partitioning the wavelet decomposed image into significant and insignificant partitions based on the following function.

$$S_n(T) = \begin{cases} 1 \ , & max_{(i,j)\in T} \{|C_{i,j}|\} \geq 2^n \\ 0 \ , & otherwise \end{cases} \tag{3}$$

where $S_n(T)$ is the significance of a set of co-ordinates, T, and $C_{i,j}$ is the value of the coefficient value at co-ordinate i, j. There are two passes in the algorithm - the sorting pass and refinement pass. The sorting pass is performed on the list of insignificant sets(LIS), the list of insignificant pixels(LIP) and the list

of significant pixels(LSP). The LIP and LSP consist of nodes containing single pixels, while the LIS contains nodes having descendants.

During the sorting pass, those co-ordinates of the pixels which remain in the LIP are tested for significance by using eq.(3). The result, $S_n(T)$, is sent to the output. Those co-ordinates which are significant will be transferred to the LSP, as well as their sign bit to the output. Those sets in the LIS which consists of nodes with descendants will also have their significance tested. Those that are found to be significant will be removed and partitioned into subsets. Those subsets with a single coefficient which are found to be significant will be added the LSP, or if they are insignificant, they will be added to the LIP.

During the refinement pass, the n^{th} most significant bit of the coefficients in the LSP is sent to the output. Then, the value of n is decreased by 1, and the sorting and refinement passes are repeated. This continues until either the desired rate is reached or $n = 0$, and all the nodes in the LSP send all their bits to the output. The latter case will result in almost perfect reconstruction, as all of the coefficients are processed completely. Moreover, the bit rate can be controlled precisely in the SPIHT, because the output is produced in the form of single bits and the algorithm can be terminated at any time.

3.1 Block Composition of a Content Similarity Tree of SPIHT

After the hierarchical wavelet decomposition, coefficients corresponding to the same orientation componets of the image at different decomposition levels are grouped into a wavelet tree structure. This wavelet tree is rooted in the lowest frequency sub-band and, out of each 2×2 root node, one node has no descendent and every other coefficient has four offspring in the high frequency sub-band of the same orientation, thus forming trees. In each tree, If one node is insignificant with respect to a given threshold, all of its offspring will most probably be insignificant with respect to the same threshold. This is a well known self-similarity property in insignificant across different wavelet frequency scale levels.

The set of four coefficients at the lowest frequency band and their descendents (if any) can reconstruct the square block-based wavelet transform. The left-hand side of Fig.1 shows the spatial orientation wavelet tree structure formed by the self-similarity property and the right-hand side of Fig.1 is its equivalent square block image content when all coefficients in the wavelet tree are grouped[8]. These data are the frequency component for a specific image area with the same block size at the corresponding position.

We use the structure of the spatial orientation tree[7] using self-similarity, in which each tree is coded independently, and group the wavelet coefficients according to the corresponding spatial blocks. After the coefficient grouping, one tree corresponds solely to one image block. For each wavelet tree, the SPIHT is employed to encode it independently. By doing so, multiple wavelet tree coding can be considered as the re-organization of the single SPIHT coded bit-stream into multiple variable-length segments, and each variable-length segment corresponds to a single wavelet tree and, thus, a single spatial block in the original image. In consequence of multiple wavelet tree coding, the coefficients of low

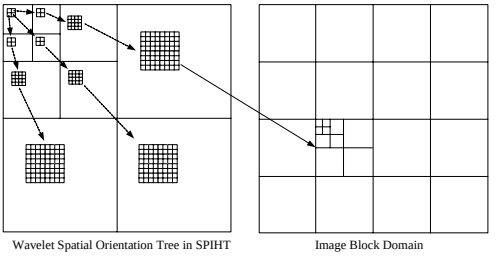


Fig. 1. Wavelet Tree and block composition from the tree

frequency band that is located in front of the each slot can be used to synchronization of data bits. This is especially important in the case of the block based SPIHT, because the low frequency band include more important information than the high frequency band. So, the coefficients of low frequency band in multiple wavelet tree less affect than that of the single SPIHT bit-stream in random bit errors and the error propagation has less noticeable effects on bits located at low frequency band.

Therefore, we propose a new algorithm which adds the ERERC function. The proposed algorithm has good performance.

4 Robust Bit-Streaming Algorithm

In image compression algorithms, DCT and Wavelet are the most popular transform therefore we suggested the algorithm which is the simultaneously effective performance enhancement of Error Resilient Entropy Coding between DCT and Wavelet. A Simple block diagram of DCT and Wavelet is shown Fig.2.

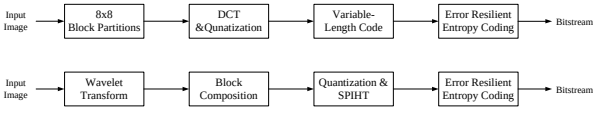


Fig. 2. DCT and SPIHT based image coding

The data in the blocks is allocated to the corresponding slots, starting from the beginning of each block. Blocks that are larger than the slot size are truncated and the remaining data are put into other slots having available space according to a predefined offset sequence. Therefore, at the receiving end, the start of each block can be automatically determined as the start of each fixed length slot. In the absence of channel errors, the decoder can follow the same algorithm to recover all variable-length blocks using the same offset sequence. When channel errors occur, error propagation in EREC decoding will more likely affect the data close to the end of each block than that close to the beginning of each block.

This characteristic fits well with the DCT and wavelet tree embedded coding we used.

In our scheme, coefficients are encoded from the lowest frequency band to the highest frequency band. Therefore, the importance of the coded data generally decreases along the bit-stream from the beginning to the end. In the case of noisy channels, error propagation will more likely affect the higher frequency band, which results in less distortion energy.

4.1 Efficient and Robust Error Resilient Entropy Coding

Efficient and Robust Error Resilient Entropy Coding(ERERC) uses the correlation of consecutive long and short blocks using high probability between the coding lengths of the transformed blocks. The initial searching position and offset value are selected by the VLC block properties.

We first describe some definitions that will be used to explain the ERERC method. Let $d_i, i = 1, 2, \ldots, M$ denote the $i-th$ block of data to be placed in M slots of equal length S. Here, M corresponds to the total number of output blocks from the Huffman encoder, and S is the average code length. Let $l(d_i)$ denote the number of bits in block d_i to be placed into the slots, and $l(m_i^n)$ be the number of bits in slot m_i at stage n of the algorithm. The indicator function is denoted by I in the definition given below, and we drop n for convenience.

Definition 1. The set $F = \{m_i, m_{i+1}, \ldots, m_k\}$ is called a full cluster if $I_{\{l(m_j) \geq S\}} = 1, j = i, i+1, \ldots, k, I_{\{l(m_{i-1}) < S\}} = 0$ and $I_{\{l(m_{k+1}) < S\}} = 0$.

Definition 2. The set $E = \{m_i, m_{i+1}, \ldots, m_k\}$ is called a partially full cluster if $I_{\{l(m_j) < S\}} = 0, j = i, i+1, \ldots, k, I_{\{l(m_{i-1}) \geq S\}} = 1$ and $I_{\{l(m_{k+1}) \geq S\}} = 1$.

Therefore, the probability of block d_i finding a partially full slot m_j is high for $j > i+1$ when $l(d_i) \geq S$, since those output blocks whose length exceeds S are more likely to be followed by similar blocks. That is, the clustering of blocks is highly correlated with the VLC block length. So, a block has to cross the full cluster and reach the partially full cluster in the consecutive stages to be placed in a slot. If $F_1, F_2, \ldots, F_r$ and $E_1, E_2, \ldots, E_s$ denote the r full clusters and the s partially full clusters, respectively, then the average length of each cluster is given by

$$L_f = \left\lceil \frac{1}{r} \left[C(F_1) + C(F_2) + \cdots + C(F_r) \right] \right\rceil \tag{4}$$

$$L_e = \left\lceil \frac{1}{s} \left[C(E_1) + C(E_2) + \cdots + C(E_s) \right] \right\rceil \tag{5}$$

where C denotes the cardinality of a set and $\lceil \cdot \rceil$ is the ceiling function. The searching step size(SP) is given below

$$SP = \left\lceil \frac{L_f + L_e}{2} \right\rceil \tag{6}$$

Since spatially neighboring blocks have similar properties, there is a chance to reduce the slot searching time. This property can reduce the searching step, which involves searching for empty slots and filling the unfilled bits in these slots with the extra bits from the remaining large slots, so that, in the proposed method, the extra bits from the large slots are placed in nearby slots. We calculate the maximum number of consecutive long blocks(L_l) and short blocks(L_s), rather than the average slot length(S). Then, we find the start position($P_{\max}$) of L_l consecutive long blocks and the start position($P_{\min}$) of L_s consecutive short blocks, using eqs.(7).

$$P_{\max} = \arg\ max_p \left(\sum_{i=p}^{p+L_l} l(d_i) \right) \qquad (7)$$
$$P_{\min} = \arg\ min_p \left(\sum_{i=p}^{p+L_s} l(d_i) \right)$$

By using these values, the initial position($P_{initial}$) is given by eq.(8)

$$P_{initial} = (P_{\min} - P_{\max} + M) \bmod M \qquad (8)$$

where M is the total number of blocks. $P_{initial}$ means the spatial distance between the consecutive long blocks with the summation of maximum block bits and the consecutive short blocks with the summation of minimum block bits, so the chance that the large slots can find small slots can be maximized with this offset value. Therefore, we suggest that the initial position be used as the initial offset value(Φ_1), and we use it to calculate the offset sequence values of the consecutive stages.

Once the SP is decided by eq.(5), the offset sequence values are calculated by eq.(9)

$$\Phi_{k+1} = \begin{cases} (\Phi_1 + k \cdot SP) \bmod M & , \quad k = odd \\ (\Phi_1 - k \cdot SP) \bmod M & , \quad k = even \end{cases} \qquad (9)$$

where $k = 1, 2, 3, \ldots$, and the '+' and '-' symbols refer to the forward and backward searching directions, respectively. Φ_{k+1} is the number of searching iterations. We must examine eq.(9) carefully. Since the offset value increases as SP, if we chose some SP, all offset values are not searched in any case (i.e, the offset value can have the same value as the previous offset after some iterations). Therefore, we find the necessary condition to avoid offset repetition. If a prime factor of the SP is $\{1, x_1, x_2, \ldots, x_m\}$ and a prime factor set of the total slot is $\{1, y_1, y_2, \ldots, y_n\}$, then offset repetition can be decided as follows.

If any value of $x_i (i = 1, 2, \ldots, m)$ does not match $y_j (j = 1, 2, \ldots, n)$ except 1, then all of the offset sequence values can be selected once by eq.(9). On the other hand, the values of the offset sequence cannot be selected once only if the above condition is violated. In that case, we adjust the SP value to assure that the necessary condition can be satisfied. The detailed operations are shown below,

Detail-1: When the same factors exist between the prime factors of the SP value and the prime factors of the total slot, with the exception of a prime factor of 1, eq.(9) cannot search all of the offset values and, hence, repetition occurs.

Detail-2: The value of SP increases one by one until the necessary condition is found. Then, new offset sequence values are decided by using the value of the new SP in eq.(9).

In the above explanation, the SP value can be incremented or decremented. Since the probability of finding an empty slot is higher with a large offset value, due to the spatial correlation of the neighboring blocks, we increment the SP value. ERERC changes the initial offset more adaptively than the existing EREC methods and depends on the characteristics of the image blocks.

4.2 Bit-Stream Structure of ERERC

By doing so in previous section 3.1, a single SPIHT coded bit stream can be considered as the multiple wavelet tree coding with variable-length code corresponding to a single spatial block in the original image. In using the multiple wavelet tree, the coefficients low frequency band including more important information are located in front of the each slot in order to synchronize the data bits. In the other side, the coefficients of low frequency band in single SPIHT can not located in front of bit stream and not synchronize the data bits. When bit errors occur, the coefficients of low frequency band in single SPIHT more affect than that of the multiple wavelet tree due to absence of synchronization.

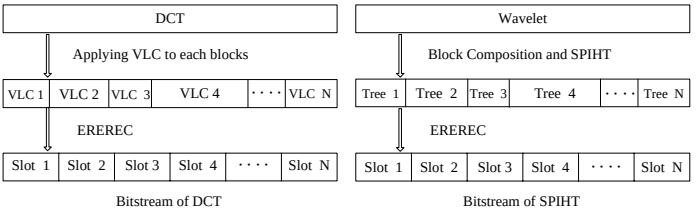


Fig. 3. Bit-Streaming of SPIHT and DCT based coding for ERERC

EREC re-organizes these multiple variable-length parts in order to obtain an error resilient coding. According to using the EREC, error propagation is conversed into the coefficient bits of high frequency band and the coefficients of low frequency band including more information can be more safety against the bit errors. Fig.3 shows the proposed basic idea compositing the blocks of DCT and SPIHT based coding for the ERERC algorithm.

5 Simulation Results

We simulated the 512×512 8 bits/pixel Boat image on the algorithms described in previous sections. EREC and ERERC algorithms are applied to the bit

(a) SPIHT-EREC (a) SPIHT-ERERREC (a) JPEG-EREC (a) JPEG-ERERREC

Fig. 4. Reconstructed image with 104 bit errors at 0.397bpp

streams obtained from 1024 trees of SPIHT and from 4096 blocks of DCT based coding in order to compare the performance of two methods.

In our experiment, we obtained self-termination by utilizing the value of the stop-layer for the encoding and decoding of each wavelet tree. Fig.4 shows the reconstructed image with 104 bit errors at 0.397bpp. Fig.5 shows the BER versus PSNR at 0.397bpp and the table 1 describes the numerical values of Fig.5.

We can summary the simulation results as follows; The quality of image is strongly affected by bit error. So the most of algorithms have low PSNR at the high error condition. In the given range of bit errors, ERERREC has high image quality that means more error resilient, when compared to previous EREC. SPIHT-ERERREC shows the best quality that is 2dB (in PSNR) higher than that of DCT-ERERREC from the image quality point of view.

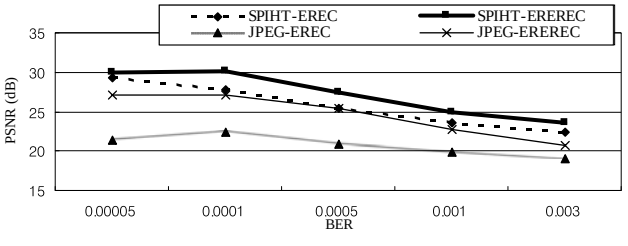


Fig. 5. BER vs. PSNR at 0.397bpp

Table 1. BER vs. PSNR at 0.397bpp

BER	*Number of Errors*	*SPIHT −EREC*	*SPIHT −ERERREC*	*JPEG −EREC*	*JPEG −ERERREC*
5×10^{-5}	5	29.3183	29.9360	21.4529	27.1074
1×10^{-4}	10	27.7035	30.1774	22.3958	27.1420
5×10^{-4}	52	25.3664	27.3722	20.9569	25.4653
1×10^{-3}	104	23.6026	24.8483	19.8813	22.6863
3×10^{-3}	313	22.4548	23.5462	19.0420	20.7721

6 Conclusion

The Error Resilient Entropy Code is a well known algorithm designed to reduce the added redundant information. We proposed a more error robust algorithm, ERREREC, for the bit stream of DCT and SPIHT based coding techniques, which greatly improves its ability to maintain the compressed image quality in the event of random errors. The simulation results of the proposed algorithm showed that the quality of transmitted image is improved for the two principal coding techniques.

Therefore, the proposed ERREREC algorithm showed better image quality than the previous EREC algorithm for the range of BER. So we can say that our proposed ERREREC algorithm has good performance on image quality and robustness for error prone wireless communications environment.

References

1. Mohammad Ghanbari and Vassilis Seferidis : Cell-loss concealment in ATM video codecs, IEEE Trans. Circuits Syst. Video Technol., vol. 3, no. 3, pp. 238-247, June 1993.
2. Hiroshi Ohta and Tokuhiro Kitami : A cell loss recovery method using FEC in ATM network, IEEE Journal on Selected Areas in Communi., vol. 9, no. 9, pp. 1471-1483, Dec. 1991.
3. V. A. Vaishampayan : Application of multiple description codes to image and video trans-mission over lossy networks, in Proc. 7th Int. Workshop Packet Video, pp. 55-60, Mar. 1996.
4. D. W. Redmill and N. G. Kingsbury : The EREC: An error resilient technique for coding variable-length blocks of data, IEEE Trans. Image Processing, vol. 5, pp. 565-574, Apr. 1996.
5. R. Chandramouli, N. Ranganathan and Shivaraman J. Ramadoss : Adaptive Quantization and Fast Error-Resilient Entropy Coding for Image Transmission, IEEE Trans. on Circuits and Systems for Video Technology, vol. 8, No. 4, pp. 411-421, August 1998.
6. Jeong-Sig Kim, Ju-Do Kim and Keun-Young Lee : The Efficient and Robust Error Resilient Entropy Coding of Compressed Image for Wireless Communications, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E88-A, no. 6, June 2005.
7. Amir Said and William A. Pearlman : A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees, IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 3, pp.243-250, June 1996.
8. Lei Cao and Chang Wen Chen : Robust Image Transmission Based on Wavelet Tree Coding and EREC, International Conference on Image Processing 2001, Proceedings 2001, vol. 3, pp. 222-225, Oct. 2001.

A Study on the Personal Program Guide Technique Within Ubiquitous Media Community Environment Using Multi-band Sensor Gateway

Sang Won Lee[1], Byoung Ha Park[1], Sung Hee Hong[1], Chan Gyu Kim[1], In Hwa Hong[1], Seok Pil Lee[1], and Sang Yep Nam[2]

[1] Digital Media Research Center, KETI,
68 Yatap-dong, Bundang-gu, Seongnam-si, Gyeonggi-do, 463-816, Korea
{swlee,bhpark,shhong,kimcg,hongih,lspbio}@keti.re.kr
[2] Dept. of Information Communication, Kookje College,
San-45, Changan-Dong, Pyongtaek-Shi, Kyeongki-do, 459-070, Korea
r13337@paran.com

Abstract. The media community service and user interface that can be supported in an ubiquitous environment is a basis for sharing through various media informations and personalized service. Experiments on the conceptual and technical elements and its application were carried out in this study. Some mobile devices, which were equipped with sensor network with a 900 MHz, 2.4GHz frequency substitution and a media server, which was equipped with sensor gateway for arbitration of each media device, were used to establish a multi-substitution ubiquitous media community environment. Under such an environment, the context information of a device was exchanged with another through a gateway in a media server and each device was able to interwork by context information generated on a different mobile device, in a different substitution, effectively.

1 Introduction

The sensor network is different from passive RFID and is semipermanent. Using a low power technology, it can transmit data to a long distance through ad hoc technology. It is developed in a 900 MHz and 2.4 GHz frequency band and studied in a various fields, such as bridge monitoring, necklace tag, temperature detection and position tracing. But we tried to apply sensor network to the media device for sharing a public media information and viewing a personal information and then eventually, to design a media community for ubiquitous society. So, we assumed that a current home looks like a ubiquitous and we established a local ubiquitous network under multi-band sensor networks. For instance, the media server is used as a sensor network gateway, each mobile device connect to it by sensor network and each device is able to exchange another with the media information or interwork each other through it. Receiving and transmitting sharing information between devices by sensor network, showing the necessary information based on user interface or interworking other networks through the

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 50–59, 2007.

media server, a wide media community environment is able to established. On this supposition, we adopted a media environment for digital broadcasting as a subject of study. The details of this paper is described as follows. The existing UI related study is introduced in a Chapter 2. The sensor working together technical contents between devices are shown in a Chapter 3. An individualization UI technology is described in a Chapter 4. The simple experiment results are shown in a Chapter 5.

2 The Trend of Researching Electronic Program Guides

Nowadays, a various EPG services are studied and built for applying digital broadcasting with a different method by many researchers. One of the examples is to use a only character information. This type of information includes the general characteristics of a TV program guide, that is, time, station, title and date, etc. of news, an interviews, sports, documentaries and movies. These are categorized into some more detailed types and is currently used by EPG of TiVo[1] and Diego [2]. It is a very generous method to navigate a electronic program guide and support services for Digital TV. Second is to use an audio Pattern Information. After analyzing an audio characteristic of a TV program, it separates a jenre of TV program using a probable variance of analyzed data. The analyzed audio patterns are saved in the storage and can be used for video indexing and segmenting of TV programs. The electronic program guides uses simple text based information about genre for whole programs, but this determines genre information at the level of program segments.[3] And third is to use medatata. The Metadata means 'data about data', and can be divided into 4 categories: Content Description Metadata, Instance Description Metadata, Segmentation Meta data, and User Preference. It can be provided more detailed program information than a current electronic program guide. In this way it is called, Advanced EPG.[4] Finally, it is to use a habit information. The broadcasting channel becomes larger than an analogue broadcast, and a lot of trouble occurs in a channel search or channel setting. In order to solve these troubles, it uses a recommendation method that an audience make a favorite TV program select based on the watching time, favorite program, preference, age, sex, job, etc. It is called, Personalized Program Guide(PPG).[5],[6]

3 Sensor Network Synchronization Technology Between Different Bands

3.1 SGMI Fragment Data Format

The SGMI data is transmitted to another device with a Fragment Data Section part in the Sync Agent Protocol Message structure and it has a 17 bytes length at this time. The front 1 byte is set up as the ID value, which can partition off SGMI data for 17 bytes, and the last byte is set up as the value that can confirm whether a normal data transmission exists. An explanation about of the

meaning of an each field is as follows for Fragment Data. The identity ID is a random value that can confirm an agreement of SGMI data transmitted in 4 bits in size, and a random value is used, and 16 kinds of SGMI data acknowledgement codes are made. The class ID is the value that SGMI data transmitted in 4 bit size can classify any kind of data into. 0x00 is meaningless with NULL, 0x01 is UI, 0x02 is the User Preference, 0x03 is the EPG data, and 0x04 0x0F is the reserved value. SGMI Data means the SGMI data that are transmitted from source to 15 bytes size, and Checksum is the value that can confirm whether the data transmitted to Fragment Data Section are normally data that have been received in 1 byte size.

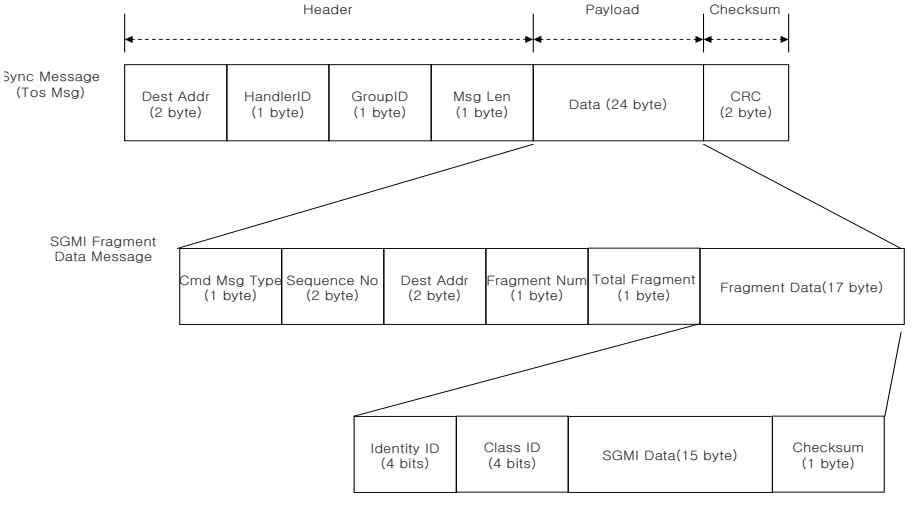


Fig. 1. A Structure of SGMI Data in TOS Message

3.2 Sync. Agent

Sync.Agent is an agent technology that it is able to control the sensor network between devices and transmit data on it. It is necessary for a user to share information with a local community even if a user does not intend to send contextual information through the sensor network module attached to the media device. An explanation about working together between devices using Sync. Agent is as follows.

Figure 2 shows the operation stream of a synchronization protocol. It broadcasts using the MMP Sync Agent which generates a MMP Discovery Req Message for User SGMI DB configuration in a Data Aggregator module of an embedded media station for a different kind device user SGMI data collection if media server is run, and accesses the sensor network for each band. It transmits an ID Beacon Message which the environment recognition Sensing value

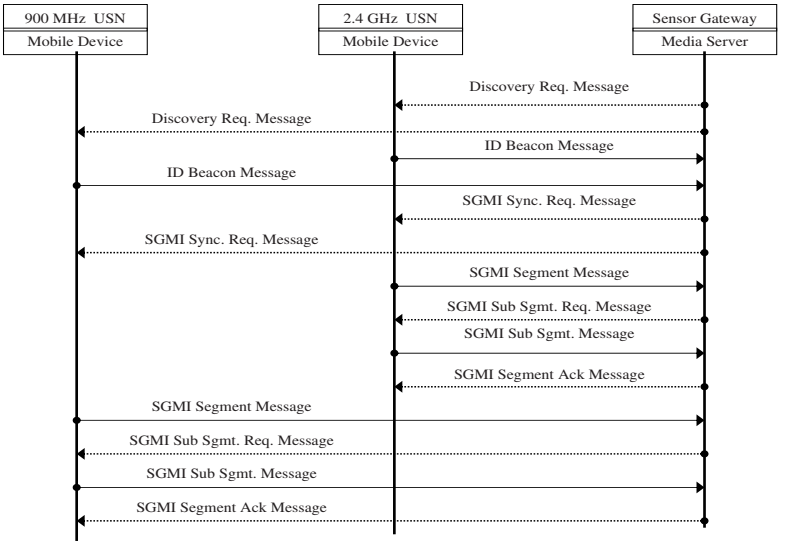


Fig. 2. Sync. Agent Synchronization Protocol Operation Flow

that Sync Agent Device Id, SGMI Version, to have gained of the MMP device that received the MMP Discovery Req Message which a broadcasting becomes with each band is simple was included in to media server for one second. It is composed Sync Agent Tree existing through Id Beacon Message parsing transmitted through multi-hop routing to a band star. It transmits SGMI Sync Req Message to the specific device that corresponds to each band through thed accessed Sync Agent if it carries out the User Profile and Version Matching that exists in the User SGMI DB as for the media server while making its round a node of Sync Agent Tree, and transmission of new SGMI data are requested. It is not a compulsive transmission, and it transmits multi-hop routing through the re-broadcasting method of a Sync Agent node in the case of this message. That is, it distinguishes whether its own ID value can be compared with the Destination field value, and the node that received the SGMI Sync Req Message was transmitted from a message to itself. It broadcasts again, and it transmits the message that it is not a message transmitted to itself, and is received by all nodes in the RF Range. The Sync Agent of MMP carries out segmentation in order to transmit the User SGMI Info accumulated in a device. Because the Payload size (25bytes) of the SGMI Segment Message transmitting User SGMI Info is limited, it is transmitted through a message fragment. It is assigned to the media server that received the first SGMI Segment Message according to the Total Fragment value of the Message Slot, and the SGMI Segment Message is transmitted in the order it is saved in the corresponding Slot and Assembling make another User SGMI Messag. The SGMI Sub Sgmt Req Message ordering

the Segment that was not received again is transmitted from the transmitted Segment Message to an MMP device by packet loss if receipt becomes inferior. Reset does the Slot which massage receipt judges an impossibility if the SGMI Sub Segment Message request that is not received in media server is requested by three times, and Req Count reaches 3, and massage receipt is canceled, and was received. Also, a receipt failure node is deleted from the Sync Agent Tree generated by the media server. If the divided User SGMI Info is received all in order, and the SGMI Slot is filled, the media server lets the SGMI Segment Ack Message to be transmitted to an MMP device, and it is received successfully.

4 The Context and Method for UI Presentation

The user interface is different from the technical interface (the interface used between devices). The technical interface is performed through direct interaction, but the user interface is performed indirectly through the interface with a computer. Also, it is not hard to make this interface. It is played in the order that the data in a system are seen, or may be changed. A command had more easily and clearly communicated through the technical interface. These can be seen with the basic structure of the basic user inter-face, and UI-related researchers have thought about the basics. D-UI (Dynamic UI) supports enough six Factor of Usability in this study,[7] and is done with the aim of providing a user interface that can adapt itself to variety in an ubiquitous environment. The details are as follows.

4.1 The Transmitted Context Information

Program Context. Seeing the ATSC(Advanced Television Systems Committee) document, a various standard table information transmitted from the digital terrestrial /cable broadcasting station is described, and a various information such as time information, program information, version information is transmitted along the MPEG-2 stream using these tables. An EIT(Event Information Table) that transmits information related to a program is mainly used in this study and all kinds of information about some program, that is, Video PID, Audio PID, the start/stop time and the like is tranmitted. [8]

User Context. A user context is composed of 4 types of informations: user ID, Device ID, user symbol information and device information. The User ID is used to confirm that transmission of the users context information to a network was authorized. The User ID value, which confirms that the user is registered to use the device, is extracted from the related media device. An random hexadecimal number value for the user was used in this study. Adding a mobility to a dynamic EPG in a wide meaning, the device ID is the value that describes the type of related media device. The broadcast-related information that a user prefers in a media device (jenre, actor, etc) is recorded in the user preference information.

UI Context. Because the UI context used in this study is applied to electronic program guide technology in digital broadcasting, it will be described with restriction. The background screen, date, time interval, broadcasting station and a program output screen were drawn on one screen on an electronic program guidebook. The screen explained a program with 6 sub-screens. The structure of this part and 1:1 mapping were got completed in this study, and to connect to a component one individual was made. EPGBack, EPGData, EPGTimeIntercal, EPGStation, EPGProgram, EPGDescription are the components that correspond individually, and a structure that can express the information that a user set up information about this with fluid by a foundation was designed.

4.2 Software Structure

MIPM(Meta Information Processing Module). The SGMI(Self Generated Meta Information) information that is the inputted Meta information is transmitted from the outside through a network (a sensor, TCP/IP) to each media device. This information is composed of user-related preference information or UI information and other a lot of information, and is transmitted in a regular structure to a remote media device. The Meta Information Processing Module was necessary, and the structure is shown in Figure 3 in order to extract UI data of the form that a user wanted through these data.

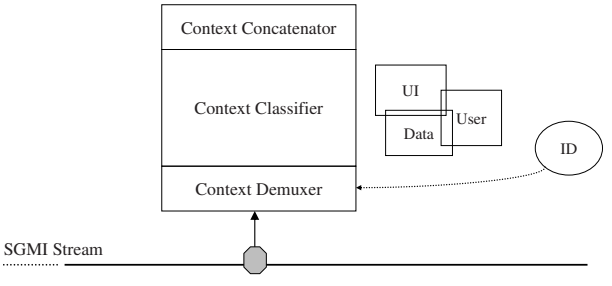


Fig. 3. A Meta Information Processing Structure

Three MIPM were executed (Context Demuxer, Context Classifier, Context Concatenator.). The Context Demuxer extracts data to make ID agree with the ID value set up at the receivubg device when a lot of data are transmitted to the SGMI Stream. The Context Classifier classifies the data extracted from the Context Demuxer. This information includes what kind of data have been classified in the SGMI Stream, and role to gather the information at the place that did data to have information with a base at the same day is carried out. The Context Concatenator is made from the total data size from the source before the transmitted data have been classified in the Context Classifier from source.

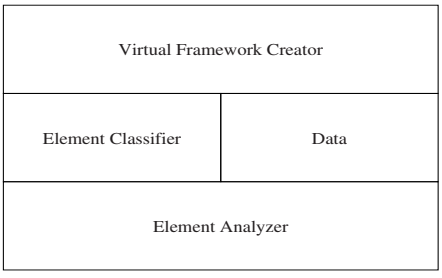


Fig. 4. A Rendering Structure

That is, the size of the date should be small, and the data that have been divided are put together, and the original data are made.

RM(Rendering Module). The data skipped through MIPM are different from the actual GUI with data of a text form.Therefore, the middle intermediate which can let convert it in the middle in order to express data of a text form to a screen with a picture must be, but it is Rendering Module(RM) to play this. The meaning of each Element is interpreted, and, as for the data that RM deals with, this can make a structure to see from XML-like data skipped in pure MIPM.Four RM are executed: the Element Analyzer, Element Classifier, Virtutal Framework Creator and Data Storage. The Element Analyzer separates the XML-like data. They are analyzed with the structure Element for Element, the User Preference for UI and Element for data are separated, and can be dealt with. The Element Classifier separates these analyzed data into a group. That is, data for UI do it in order to connect the each object to an event in GUI, or User Preference does it in order it is classified particularly, and to be able to deal with each ID. The Virtual Framework Creator functions before being shown in an actual screen in order total configuration tries to be expressed with a virtuality, or to monitor it. It is played in order it is marked by an actual screen, and to be able to all inspect it internally. Data Storage saves the data transmitted through a network, but the data of this time are saved according to the rules that were predefined so that it is used suitably when UI is expressed.

GPM(Graphic Processing Module). The GPM deals with the D-UI finally so that the GUI that a user requests on an actual screen is expressed, and the interaction is presented to a user.

The GPM performs three executions. Virtual Framework Analyzer, Graphic API Mapper, and Presentation Engine. The Virtual Framework Analyzer analyzes the Virtual Framework information generated in RM. The information that was analyzed here is communicated with the Graphic API Mapper, but graphic API and 1:1 mapping are done so that it can be marked by an actual screen. This information is displayed on the screen by the actual graphic engine.

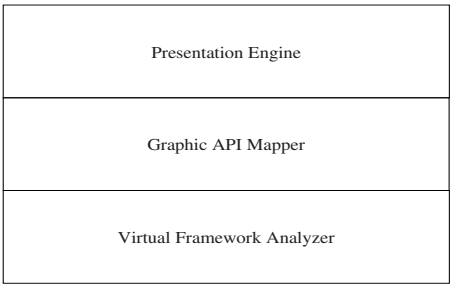


Fig. 5. A Graphical Process Structure

5 Experiment

If a mobile media device get into the local media community environment, the user information and UI information saved in a device is transmitted to the media server through a sensor network, and the information transmitted to a media server is saved per user. At this time, it is confirmed whether there is relevant user information on the media server if a user requests an EPG screen display in a mobile device. If there is a registered user information, it is displayed EPG on a screen based on the information by using a electronic program guide program. If not, it is displayed EPG supported basically. Figure 6 is a mobile device equipped with sensor network for experiment and Figure 7 shows an experiment environment.

The user context information that is transmitted from a mobile device to a media server is used for EPG context extraction from the MPEG-2 stream. An UI context information transmitted from a mobile device is used for configuring UI components and properties so that an complete electronic program guide screen is made. Because there are a lot of media community environments in real life and we cannot configure it, so we was only used the restriction of a cable

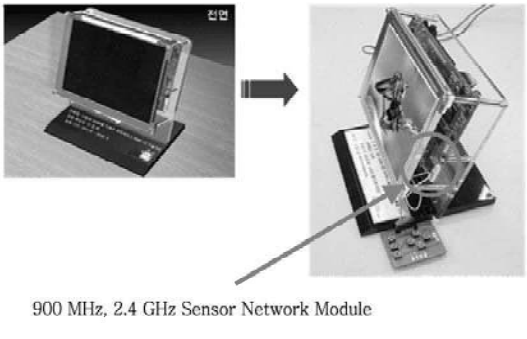


Fig. 6. A mobile device equipped with sensor network for experiment

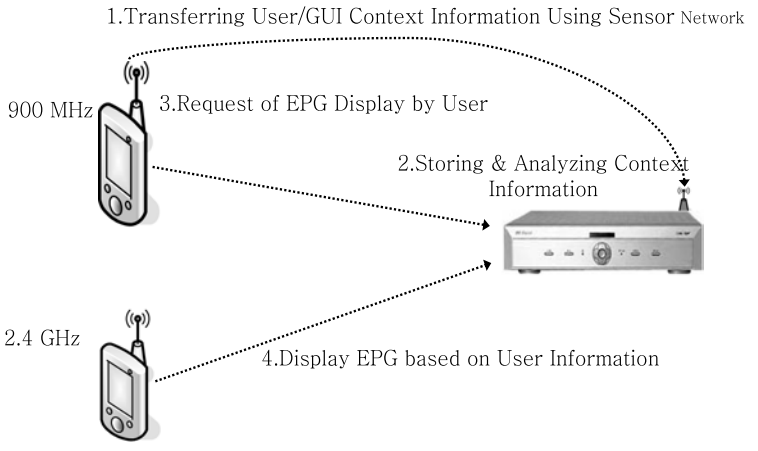


Fig. 7. An experiment environment

network based on an ATSC standard in this study. The local media community environment was composed through a sensor gateway of the media server with a sensor network of different bandwidths and lots of mobile device are arbitrated to communicate with each other by a media server. In this study, we showed the possibility of the EPG service with a User Context/UI Context/EPG Context using a sensor gateway. But it took a considerable time to transmit the EPG Context and the same large data restricted the transmission speed of a sensor network, and the bottle-neck phenomenon to have obeyed transmission occurred. But the application was able to show that considerable possibilities for a media environment that uses a sensor network.

6 Conclusion

Although a lot of burdens follow, it was very effective in a measure to use a sensor network within a virtual ubiquitous environment. While one is unconscious, some data were transmitted to another device, reproduced in a media device and displayed on a screen. Even though a media device is under a multi-band, the media community was established through media server with sensor gateway and the information was shared each other. As a result of a study, we proved that a sensor network can be used to transmit some information in a media environment and applied to a ubiquitous media environment sufficiently.

Acknowledgments. This research is supported by the ubiquitous Computing and Network (UCN) Project, the Ministry of Information and Communication (MIC) 21st Century Frontier RD Program in Korea.

References

1. TIVO, www.tivo.com
2. Data Sheet, www.moxi.com
3. Jasinschi, R.S. Louie,J. Automatic TV program classification based on audio patterns, Euromicro Conference, Proceedings. 27th(2001)370-375
4. Marija Leban, Internet Search for TV Content Based on TV Anytime, EUROCON 2003, Vol.2 (2003) 70-73
5. L.Ardssono, C.Gena, P.Torasso, et al., Personalized recommendation of TV programs, LECT NOTES ARTIF INT2829, (2003) 475-486
6. Isobe, Fujiwara, Kaneta, Morita Uratani, Development of a TV reception navigation system personalized with viewing habits, IEEE Transactions on Consumer Electronics, Vol.51(2005) 665-674
7. Soren Lauesen, User Interface Design, ADDISON WESLEY, (2005)
8. ATSC Standard: Program and System Information Protocol for Terrestrial Broadcast and Cable (Revision B), (2003)

Remote Diagnostic Protocol and System for U-Car

Doo-Hee Jung[1], Gu-Min Jeong[2], Hyun-Sik Ahn[2,⋆], Minsoo Ryu[3], and Masayoshi Tomizuka[4]

[1] Department of Electronics Engineering, Korea Polytechnic University, Korea
doohee@kpu.ac.kr
[2] School of Electrical Engineering, Kookmin University, Seoul, Korea
{gm1004,ahs}@kookmin.ac.kr
[3] College of Information and Communication, Hanyang University, Korea
msryu@rtcc.hanyang.ac.kr
[4] Department of Mechanical Engineering, University of California, Berkeley, USA
tomizuka@me.berkeley.edu

Abstract. This paper proposes a remote diagnostic protocol and system for U-Car, which has same functionality of conventional scanners, automobile diagnostic instruments. It consists of a remote server, mobile handsets with an application and a connecting device. The remote server has vehicle diagnostic database. The application program on mobile handsets relays diagnostic signals and controls the flow of data packets based on the dedicated protocol. The connecting device is used for converting the signal level and protocols between automobiles and handsets. Since the remote server diagnoses vehicle directly, diagnostic history can be accumulated automatically. Therefore, value added services such as connecting service-shop directly based on diagnostic results are possible. New automobile systems can be easily dealt only by changing diagnostic database of the remote server. Although the proposed system is constructed based on mobile handsets, it can be easily extended to car-PCs and other systems for U-car.

Keywords: U-car, Telemetry, Vehicle diagnostic system, Mobile handsets.

1 Introduction

With the advance of communication technologies, automobile systems become more and more intelligent. U-car means such intelligent car systems based on network capability. Many value-added services such as navigation services based on real-time traffic information, car office based on car-PCs, and automated highway system can be possible from network capability using various media. Among these, mobile handsets are important and powerful media.

Telemetry services are remote control and measurement services. Users can control local devices remotely using mobile handsets that connect to remote

⋆ Corresponding author.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 60–68, 2007.

control servers, which are physically connected to the controlled devices. Meanwhile, different from this common concept of telemetry services, the concept of Smart Telemetry Service(STS)[1] has been proposed. In STS, the local device is controlled from the remote server by physically connecting with mobile handsets, which is easy-to-carry and hence very important media in the ubiquitous environment.

Vehicle Diagnostic system is one of the important part in U-car systems. In automobile diagnosis market, the functionality of conventional scanner are imported to telematics devices that are customized for each type of car systems. Although new car systems accord to the international standard ODB-II, there are many previous car systems using other protocols. Therefore, it is difficult to make telematics devices that accommodate these various car systems.

In this paper, we propose a diagnostic protocol and system for U-car. The proposed system is based on Smart Telemetry Service technology and uses remote server, simple connecting devices(or converting board), and wireless handsets. It is possible to cover various car systems by downloading an application on mobile handsets and upgrading the server.

Moreover, to resolve time delay problem in mobile networks, a dedicated protocol is designed for vehicle diagnostic systems. In the proposed method, vehicle diagnostic operation is executed directly from remote server. Additional car system can be covered by only upgrading diagnostic database on the server. Various services can easily be deployed based on the diagnostic history accumulated automatically in the server.

This paper is organized as follows. At first, we explain the concept of Smart Telemetry Services and the configuration of remote diagnostic systems. Then the problem of time delay is discussed and the dedicated protocol to resolve this problem is explained briefly. To show the effectiveness of the proposed method, experimental results are shown. Finally, future works are discussed.

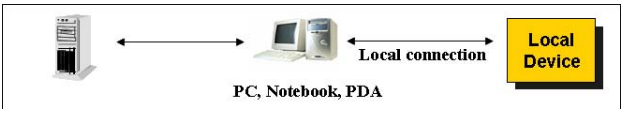
2 Smart Telemetry Services

Different from general telemetry services, Smart Telemetry Service(STS) is based on the mobile handsets carried by users. STS makes remote server control local devices directly which do not have network capability. That is, it requires users to carry mobile handsets or other devices with network capability. It is useful in the cases when most operations can be executed in the remote server without changing local devices.

It is also different from the previous data communication scheme that uses mobile handsets as simple CDMA modem. The difference of previous scheme and the proposed method is depicted in Fig.1.

Fig.1 (a) shows the previous data communication scheme. In this scheme, notebook or PDA systems connect to remote server by using mobile handsets as CDMA modem. In this case, it can be applicable only for network-capable devices such as notebook or PDA. That is, it cannot be applicable for the previous network-incapable local devices.

(a) a scheme for previous data communication services



(b) a previous scheme for network-incapable local devices

(c) the proposed scheme

Fig. 1. Comparison between previous schemes and the proposed scheme

Fig.1 (b) shows a typical scheme controlling these network-incapable local devices. In this scheme, using notebook or PDA, local devices can connect to remote server through wired or wireless network. Recently, market trend goes to converge mobile handsets and PDA. However, due to time and cost, it is not appropriate for common users.

Fig. 1 (c) shows the proposed scheme that remote server controls local devices through mobile handsets. In this case, mobile handset operates as a virtual cable interconnecting local devices and remote server over CDMA networks.

The proposed scheme is effective for controlling previous network-incapable local devices from remote server. It is done by downloading application for mobile handsets.

In this scheme, one of the important problem is the time-delay in wireless networks and internet. For most vehicle diagnostic protocols including ODB(On-Board Diagnostics)-II, it is required to exchange signals within pre-defined time for establishing connection and for keeping connection after connection establishment. Since such conditions cannot be guaranteed in wireless networks, it is necessary to design a dedicated protocol that resolves this time-delay problem.

3 Configuration of the Proposed Vehicle Diagnostic System

In Fig.2, the configuration of the proposed vehicle diagnostic system is depicted. The remote server and the application on mobile handsets exchange information

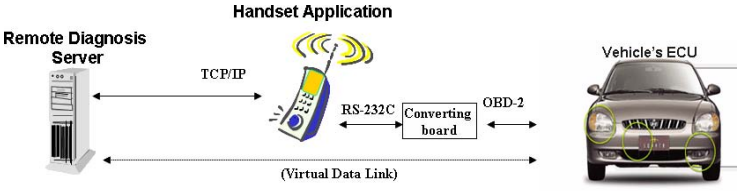


Fig. 2. Configuration of the proposed vehicle diagnostic system

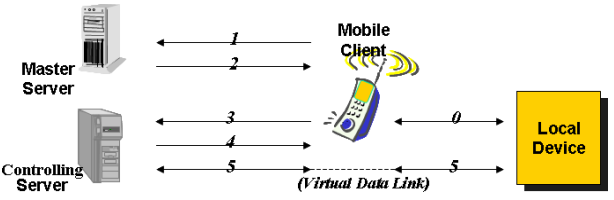


Fig. 3. Procedure of the proposed vehicle diagnostic systems

on TCP/IP. Mobile handset application does the functionality of user interface and relaying data for remote diagnostic operations to ECUs(Engine Control Units) through the converting board. Mobile handsets and the converting board are connected through UART with voltage level of 0 to 3.3[V] and the converting board changes voltage levels and communication schemes for the ODB-II protocol or others.

Users can select automatic diagnostic operation for pre-registered car system or do manual diagnostic operation for other car systems using user interface menus provided by the remote server. After selecting the type of car system and connecting the cable between the converting board and ECUs, users can start diagnostic operation. According to the protocol between mobile handsets application and the remote server, the control data from the remote server bypasses to ECUs through the converting board that converts signal levels and communication schemes.

Through the pre-defined establishment sequence and the diagnostic protocol, ECUs and the remote server exchange data for diagnostic operation. Then, the remote server analyzes the diagnostic data. Finally, users can check the results through mobile handsets. Such results are saved automatically on remote server and various additional services such as automatically connecting to nearby service shops can be provided easily based on the results.

In this paper, we concentrate on the parts of Diagnostic Trouble Codes(DTC) and MIL (Malfunction Indicator Light) codes which are frequently used for commercial stand-alone scanner. Although commercial stand-alone scanner has additional functionality like analog signal measurements, this functionality is rarely used. For the proposed cost-effective diagnostic system, we have no choice but to consider only the parts with digital signals.

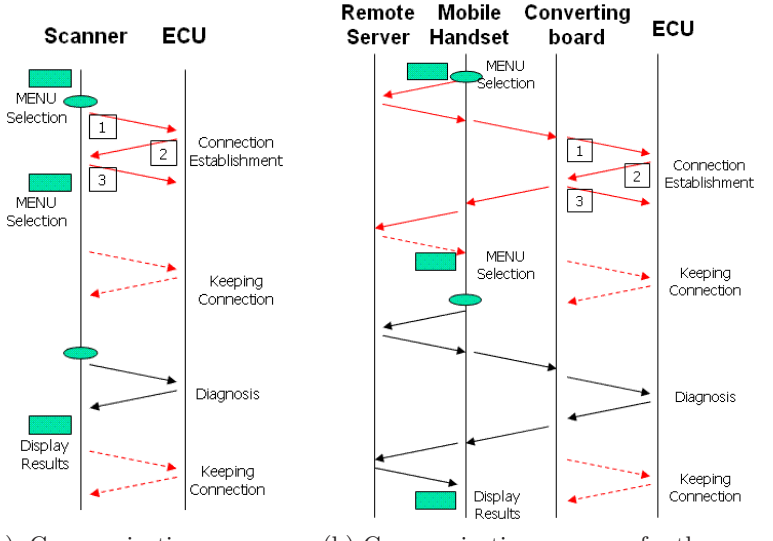
Fig.3 shows the procedure of the proposed system briefly. The procedure can be summarized as follows:

- Step 0: The handset and ECUs of the automobile are connected using a converting board.
- Step 1 and Step 2: The handset sends the service request signal to master server.
- Step 3 and Step 4: ECUs of the automobile (local device) is connected to the controlling server through the wireless handset.
- Step 5: Using virtual data link, data is transmitted and analyzed. After that, the results are displayed in the handset.

Note that the functionality of Master Server and Controlling Server in Fig. 3 can be located together in one remote server or can be distributed on more than two servers with load-balancing capability.

4 A Protocol for Resolving Time-Delay Problem

To illustrate the basic concept of the protocol for resolving time-delay problem in wireless networks, exchange of diagnostic data is depicted in Fig.4. Here, square parts represent messages for users from the stand-alone scanner(or the remote server) and elliptic parts represent the key input of users to the scanner(or mobile handsets). Arrow lines represent data flow between the scanner(or mobile handsets through the converting board) and ECUs.



(a) Communication sequence between the scanner and ECUs (b) Communication sequence for the proposed system

Fig. 4. Communication sequence of data for diagnostic operation

Command	Length	Baudrate Setting	Config. 1	Config. 2	Config. 3	Config. 4	Checksum
33H	12	8 bytes	1 byte	1 byte	1 byte	1 byte	1 byte

(a) Data format of Configuration Command

Command	Length	Initialization configuration	Data	Checksum
35H	n	1 byte	(n-1) bytes	1 byte

(b) Data format of Communication Setup Command

5 bps Code	1st ECU Response	Min. Delay for Response	1st Board Response
1 byte	n bytes (n=0,1,2,3)	1 byte	1 byte

(c) Data format of Data field in (b) (5 bps)

No. of Random Pattern Data	Random Pattern Data	1st ECU Response	Min. Delay for Response	1st Board Response
Nr	Nr bytes	n bytes (n=0,1,2,3)	1 byte	1 byte

(d) Data format of Data field in (b) (random pattern)

Fig. 5. Data format of the dedicated protocol

There are several timing requirements in the protocol between the scanner and ECUs. Typical examples are related to the connection establishment and keeping connection. When the scanner sends signal 1 to the ECU for connection establishment, the ECU responses signal 2 within pre-defined time to notify ready condition. In this case, the scanner should send signal 3 within a certain amount of time. However, if the remote server checks signal 2 and sends signal 3, such timing requirements cannot be satisfied. It is due to the unpredictable time-delay in wireless networks and internet.

To resolve this problem, we design a dedicated protocol that manipulates local response. Using this protocol, the application program on mobile handsets generates local response directly while sending response to the remote server. Therefore, timing requirement of the vehicle diagnostic protocol can be satisfied while notifying server that connection is successfully established.

After connection establishment, it is also necessary to send connection keeping signal within pre-defined time to make connection alive. If remote server sends directly these signals for keeping connection, due to the unpredictable time-delay in wireless networks and internet, it is impossible to guarantee timing requirement for keeping signal. Moreover, unnecessary data exchange on wireless network results in the increase of the cost for data communication. With the dedicated protocol for keeping connection, the application program on mobile handsets can generate these signals periodically after first receiving command related to this operation. Therefore, we can satisfy timing requirement for keeping connection and also reduce unnecessary data exchange through wireless networks.

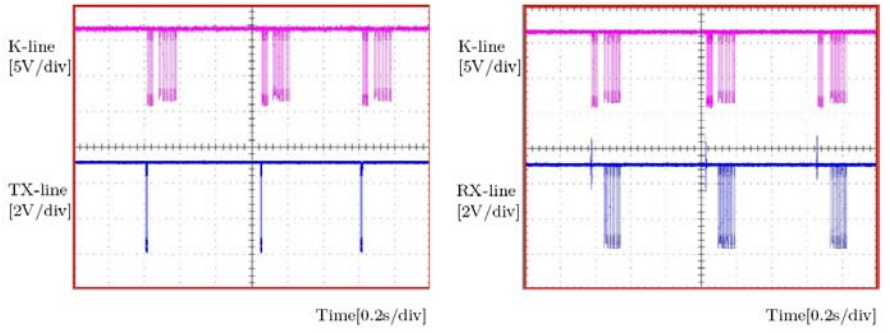
To show the details of the dedicated protocol, we depicted data format of the commands for the converting board in Fig 5. Using configuration command of Fig. 5(a), we can change baudrate and polarity, synchronization method, channel selection, pull-up selection, and handshaking method to support various car systems. Fig. 5(b) shows communication setup command that is used for establishing command. It selects initialization method of 5 bps mode and wakeup pattern generation mode which generates 50ms wakeup patterns. initialization configuration contains the length of 1st ECU response data and the length of 1st acknowledgement data. If initialization method is 5 bps mode, data fields in Fig. 5(b) is in the form of Fig. 5(c). For random pattern mode case(wakeup pattern generation mode), it is in the form of Fig.5(d). In both cases, data fields in Fig. 5(b) contains the value of initialization code, 1st ECU response data, minimum delay for response time, and 1st acknowledgement data. After checking with this 1st ECU response data, converting board sends 1st acknowledgement data to ECU after minimum delay for response time.

5 Experimental Results

To show the applicability of the proposed system, we construct a diagnostic program for remote server, an application program for mobile handsets and a converting board respectively. In constructing database on the remote server for vehicle diagnostics and the hardware/software systems for the converting board, we cooperated with Nex-tek Cooperation, one of the major makers of commercial stand-alone scanners[5].

A diagnostic program for the remote server works together with multi-process/ multi-thread TCP/IP server module and database for vehicle diagnosis. An application program on mobile handsets is implemented on BREW environment[6]. LCD display and keypad of mobile handsets act like dummy terminal through which users can select type of car systems or desired diagnostic operation. They can also be used for checking the results of diagnostic operation. 8-bit microcontroller (PIC16C73 of Microchips Co.) is used for the converting board. The program of the converting board is coded in assembly language to enhance real-time performance. For serial communication between the converting board and mobile handsets, internal serial ports are used. On the other hand, a software-generated serial communication scheme through digital input/output pins is used for communication between the converting board and ECUs. As mentioned above, schemes in the commercial system is referenced in designing hardware of the converting board for signal level conversion and output pattern selection for various car systems.

Fig. 6 shows the waveforms of diagnostic operation. the ECU of ATOS (Hyundai Motor Company)[7] is diagnosed. The upper waveforms of (a) and (b) show signal on K-line through which the converting board and the ECU communicate diagnostic data, while the lower waveform of (a) shows signal on



(a) Waveform of K-line and TX-line (b) Waveform of K-line and RX-line

Fig. 6. Waveforms of diagnostic operations

TX(transmit) line from mobile handsets to the converting board. After connection establishment, the application program on mobile handsets generates signal for keeping connection periodically as shown in the lower waveform according to the protocol and the command from the remote server.

Although the response of the ECU is sent back from the converting board to mobile handsets though RX(receive) line as shown in the waveform of (b), to remove unnecessary data exchange on wireless networks, it is discarded according to the protocol and the command from the remote server.

6 Conclusion

In this paper, we propose a diagnostic protocol and system for U-car with mobile handsets, which has same functionality of conventional scanners, automobile diagnostic instruments. Since diagnostic algorithms are located on the remote server, we can construct cost-effective vehicle diagnostic systems.

Experimental work shows the validity of the proposed protocol and system. Although the proposed system is constructed based on mobile handsets, it can be easily extended to car-PCs and other systems for U-car.

Further works can be done in two ways. One is for applying the proposed scheme to other systems for U-car which has cost-effective network capability than mobile handsets. The other is for elimination of wired connection between mobile handsets and the converting board by using wireless PAN(Personal Area Network) technologies.

Acknowledgment

This work was supported by the 2005 research fund of Kookmin University in Korea.

References

1. Dong-Heon Lee, "Method for Relaying Remote Control Signal, Mobile Communication Terminal Using the Same and Mobile Communication System Using the Same", Korea Patent 10-0441969-0000
2. Peter David, Ruben Fernandez, "OBD II Diagnostic: Secrets Revealed", Kotzig Publishing, Inc.
3. Peter David, "OBD II Fault Codes Reference Guide", Kotzig Publishing, Inc.
4. Gi-Seok Kim, Jung-Hun Oh, Dong-Wook Kang, and Ki-Doo Kim, "Implementation of Automotive Remote Diagnosis Function for Telematics Systems", The Korean Society of Broadcast Engineers Conference, pp. 305-308, 2004.
5. www.nex-tek.com
6. http://brew.qualcomm.com/brew/ko/
7. http://worldwide.hyundai-motor.com/

Map-Building and Localization by Three-Dimensional Local Features for Ubiquitous Service Robot

Youngbin Park[1], Seungdo Jeong[2], Il Hong Suh[1], and Byung-Uk Choi[1]

[1] College of Information and Communications, Hanyang University
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791, Korea
ybpark@mlab.hanyang.ac.kr, {ihsuh, buchoi}@hanyang.ac.kr
[2] Department of Electrical and Computer Engineering, Hanyang University
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791, Korea
sdjeong@hanyang.ac.kr

Abstract. In this work, we propose a semantic-map building method and localization method for ubiquitous service robot. Our semantic-map is organized by using SIFT feature-based object representation. In addition to semantic map, a vision-based relative localization is employed as a process model of extended Kalman filters, where optical flows and Levenberg-Marquardt least square minimization are incorporated to predict relative robot locations. Thus, robust map-building performances can be obtained even under poor conditions in which localization cannot be achieved by classical odometry-based map-building. To localize robot position and solve kidnap problem, we also propose simple, but fast localization method with a relatively high accuracy by incorporating our semantic-map.

1 Introduction

To navigate and plan a path in an environment, the intelligent ubiquitous service robot has to be able to build a map of the environment and to recognize its location autonomously. SLAM(Simultaneous Localization And Mapping) is a popular technique to accurately localize the robot and simultaneously build a map of the environment. Numerous studies of SLAM have been performed, as this has been one of the important issues in the intelligent mobile robot community for a long time. Since the 1990s, methods using extended Kalman filters have been focused on by a number of researchers interested in SLAM [3,8]. Most algorithms using extended Kalman filters combine two methods. One is relative localization, which is the method used to compute the current position with respect to an initial location. The other is the method using a map composed of landmarks.

Odometry is often used for relative localization. However, it has many errors caused not only by systematic factors such as a difference in wheel diameter, inaccurate gauging of wheel size, and others, but also non-systematic factors such as slip, poor condition of the floor, etc. [4] proposed an algorithm to reduce such systematic errors. Their algorithm consists of three steps: odometry error modeling, error parameter estimation using the PC-method, and estimation of

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 69–79, 2007.

the covariance matrix. Even so, it is hard to overcome the error of estimation caused by non-systematic factors.

Building the map of the environment is an issue with SLAM. The map consists of landmarks which are used for ubiquitous service robot localization. Davision uses three-dimensional corner points as landmarks, which are obtained by a Harris corner detector and stereo matching [5]. Se uses keypoints obtained from SIFT(Scale Invariant Feature Transform) [6,10]. However, simple coordinates or features within the image are not enough to facilitate interaction with humans.

Robot localization has been another issue in robot navigation system. A variety of methods have been proposed to locate robot position. One of those methods are to update robot position by utilizing odometer history. In indoor environments, those methods probably yield good results. But, when the robot moves without the odometer history update as in the kidnapped robot problems, robot cannot locate itself. And, robot position should be determined by other sensor information in the map. In solving the kidnapped robot problems, due to uncertainties and incorrectness of map and sensor information, estimation of robot position yields errors. Determining accurate robot position by estimated values from these incorrect measurements is also challenging task [9].

In this work, we propose the semantic-map building method for ubiquitous service robot. For building the semantic-map, we use vision-based relative localization process as the process model of our extended Kalman filter. This approach is robust enough for environments which cannot be supported by encoders that measure values such as the number of rotations of the robot's wheels. And we also propose a technique to locate robot position in a realtime while ensuring accuracy, where our semantic-map and stereo vision are employed.

2 OFM-Based Robot Localization and Semantic-Map Building

In this work, we propose a three-dimensional object feature model (OFM) with essential properties and propose a method to use the OFM for improving the performance of vision-based SLAM. Figure 1 shows the block diagram of our SLAM method using a three-dimensional OFM.

We use images taken by the stereo camera as the only sensor information in this work. Our system is composed of three parts: the real-time relative localization part which estimates robot location in real-time, the landmark recognition part which builds up and/or observes landmarks by using SIFT-based object recognition, and the data fusion part which combines two observation data using an extended Kalman filter.

2.1 Real-Time Relative Localization

Image-based relative localization is a method that applies the motion between corresponding points within consecutive image sequences to localize the robot. Thus, matching between feature points is a crucial factor for localization performance.

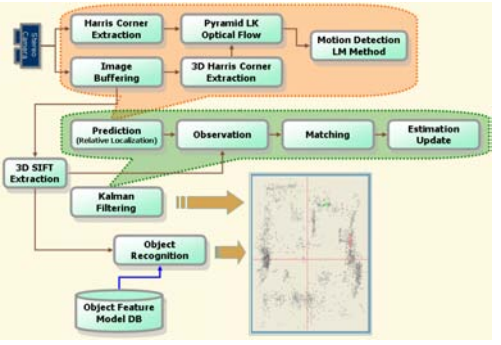


Fig. 1. Block diagram of the proposed map-building

Tracking of corner features. To track the camera motion, it is necessary to obtain the feature points in the current frame corresponding to the feature points in the previous frame. Note that the corner points satisfy the local smoothness constraint. Lucas-Kanade optical flow performs well in tracking corner points with that property [7]. However, Lucas-Kanade optical flow is best suited to small motion tracking. It is not sufficient to follow the movement of the ubiquitous service robot. Thus, pyramid Lucas-Kanade (PLK) optical flow using a Gaussian image pyramid is used for relative localization, where it is known to track a relatively wide area.

Relative localization. The feature points extracted from the initial frame are used as 3D point landmarks. Note that we use the stereo camera as the input sensor, and we obtain a stereo image pair. The 3D coordinates of the extracted feature point are calculated using the disparity of the stereo camera images. The stereo camera is previously calibrated. When we know the exact relationship between the 3D coordinates of the landmark and the corresponding 2D coordinates which are projected to the image, the projection matrix indicates the mapping relationship between points of 3D space and points of the image.

The projection matrix is composed of two parts. One is the intrinsic parameter matrix which includes internal information of the camera. This matrix represents the relationship between the camera coordinate system and the image coordinate system. The other part of the projection matrix is the extrinsic parameter matrix. This matrix represents the translational and rotational relationship between the 3D space coordinate system and the camera coordinate system.

Therefore, if the projection matrix is estimated with coordinates of corresponding points in 3D space and in the 2D image, we can obtain the motion of the camera and the relative location of the robot. It is a non-linear problem to estimate the projection matrix with numerous corresponding points. In this work, we estimate the rotation and translation components using the Levenberg-Marquardt least square minimization (LM LSM) method.

2.2 Extended Kalman Filter-Based SLAM Using Object Landmarks

In most previous research works on SLAM, landmarks were composed of lines of the environment. They were obtained using a range finder or feature points of camera images. These maps only involve coordinates of feature points. Thus, the use of these maps for high level purposes, such as delivery tasks or interactions with humans, requires an anchoring technique to link point-based landmarks with their associated semantics. Alternatively, in this work, objects are recognized using SIFT and then features of recognized objects are registered as landmarks. Therefore, we can create the semantic-map supporting the location of main objects in the environment without any additional anchoring process.

In most SLAM methods using Kalman filters, odometry-based robot kinematics or dynamics are used as the process model. However, those performances are very poor because of systematic factors such as the difference of wheel diameter and non-systematic factors such as slip. Moreover, in the case that a wheel encoder does not exist, such as for a humanoid robot, odometry-based methods are difficult to apply.

Therefore, to resolve such drawbacks, we propose the extended Kalman filter-based SLAM that integrates PLK optical flow and LM algorithm-based relative localization with object landmarks.

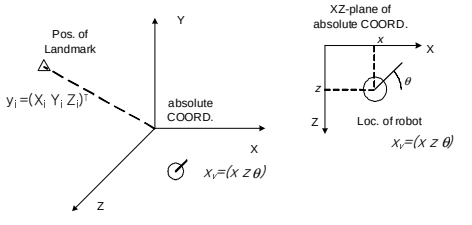


Fig. 2. System coordinate for the robot location and the landmark position

The state vector and covariance. We assume that the robot is located on a plane as in Fig. 2, so position and orientation of the robot is represented by $x_v = (\, x \; z \; \theta \,)^T$. The position of each landmark is denoted as $y_i = (\, X_i \; Y_i \; Z_i \,)^T$. Here, SIFT keypoints of objects recognized by the three dimensional OFM are represented as absolute coordinates. To regulate uncertainty of landmarks and relationships among them we use the system state vector and covariance model proposed by [1]. Thus, we can represent the state vector p and covariance matrix Σ as

$$p = \begin{pmatrix} x_v \\ y_1 \\ y_2 \\ \vdots \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2_{x_v x_v} & \sigma^2_{x_v y_1} & \sigma^2_{x_v y_2} & \cdots \\ \sigma^2_{y_1 x_v} & \sigma^2_{y_1 y_1} & \sigma^2_{y_1 y_2} & \cdots \\ \sigma^2_{y_2 x_v} & \sigma^2_{y_2 y_1} & \sigma^2_{y_2 y_2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1}$$

Three dimensional object feature model. The three dimensional object feature model(3D OFM) is defined as the model which is composed of the three dimensional SIFT keypoints extracted from object images. To make the three dimensional object feature model, we rotate and object by 20 degrees with respect to the center of gravity and then take the images using the calibrated stereo camera. This process is repeated to get 18 views images. SIFT keypoints are extracted from each image and are given three dimensional coordinates calculated with the stereo vision technique. SIFT keypoints having three dimensional coordinate are called as the three dimensional SIFT keypoints. The three dimensional SIFT keypoints in objects recognized by using 3D OFM are used as landmarks.

Process model. In EKF, we define the process model as the estimation of current location and its covariance for the robot and landmarks referring to the state vector and its covariance matrix for the previous system. We apply the displacement $\Delta x, \Delta z, \Delta\theta$ obtained by PLK optical flow and LM algorithm-based relative localization to the process model. Let the process model of our work be given as

$$x_v(k+1|k) = f(x_v(k|k), u(k)) = \begin{bmatrix} x \\ z \\ \theta \end{bmatrix} + \begin{bmatrix} \Delta x_r \\ \Delta z_r \\ \Delta\theta \end{bmatrix}. \tag{2}$$

In (2), Δx_r and Δz_r are given as

$$\begin{bmatrix} \Delta x_r \\ \Delta z_r \end{bmatrix} = \begin{bmatrix} cos(\theta+\Delta\theta) & sin(\theta+\Delta\theta) \\ -sin(\theta+\Delta\theta) & cos(\theta+\Delta\theta) \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta z \end{bmatrix}. \tag{3}$$

The current location for the robot $x_v(k+1|k)$ is estimated by function f as shown in (2). $x_v(k|k)$ and $u(k)$ represent displacement of the previous location and motion of the robot respectively. Δx_r and Δz_r are the displacement which are transformed to the reference coordinate system of the map from the estimated displacement Δx and Δz referring to the camera coordinate system of the previous location.

$$y_i(k+1|k) = y_i(k|k), \forall i. \tag{4}$$

In (4), $y_i(k+1|k)$ represents the estimated current position of the i-th landmark, and we assume that landmarks are fixed.

The covariance for the system $\sigma^2_{x_v x_v}, \sigma^2_{x_v y_i}$ and $\sigma^2_{y_i y_j}$ are obtained as

$$\begin{aligned} \sigma^2_{x_v x_v}(k+1|k) &= \nabla_{x_v} f \sigma^2_{x_v x_v}(k|k) \nabla_{x_v} f^T + \nabla_u f \sigma^2_u(k|k) \nabla_u f^T, \\ \sigma^2_{x_v y_i}(k+1|k) &= \nabla_{x_v} f \sigma^2_{x_v y_i}(k|k), \\ \sigma^2_{y_i y_j}(k+1|k) &= \sigma^2_{y_i y_j}(k|k), \end{aligned} \tag{5}$$

where $\nabla_{x_v} f$ and $\nabla_u f$ represent the Jacobian of the estimation function f for the state vector of the robot and the displacement respectively. These are given as

$$
\begin{aligned}
&\nabla_{x_v} f = \left[\, \tfrac{\partial f}{\partial x} \; \tfrac{\partial f}{\partial z} \; \tfrac{\partial f}{\partial \theta} \,\right], \quad \nabla_u f = \left[\, \tfrac{\partial f}{\partial \Delta x} \; \tfrac{\partial f}{\partial \Delta z} \; \tfrac{\partial f}{\partial \Delta \theta} \,\right], \\
&where \quad \partial f/\partial x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T, \quad \partial f/\partial z = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T, \\
&\frac{\partial f}{\partial \theta} = \begin{bmatrix} -sin(\theta+\Delta\theta)\Delta x + cos(\theta+\Delta\theta)\Delta z \\ -cos(\theta+\Delta\theta)\Delta x - sin(\theta+\Delta\theta)\Delta z \\ 1 \end{bmatrix}, \\
&\partial f/\partial \Delta x = \begin{bmatrix} cos(\theta+\Delta\theta) & -sin(\theta+\Delta\theta) & 0 \end{bmatrix}^T, \\
&\partial f/\partial \Delta z = \begin{bmatrix} sin(\theta+\Delta\theta) & cos(\theta+\Delta\theta) & 0 \end{bmatrix}^T, \\
&\frac{\partial f}{\partial \Delta \theta} = \begin{bmatrix} -sin(\theta+\Delta\theta)\Delta x + cos(\theta+\Delta\theta)\Delta z \\ -cos(\theta+\Delta\theta)\Delta x - sin(\theta+\Delta\theta)\Delta z \\ 1 \end{bmatrix}.
\end{aligned} \tag{6}
$$

Here, $(\,\Delta x \;\Delta z \;\Delta\theta\,)$ is the movement of the robot estimated by the LM algorithm.

In (5), σ_u^2 is the covariance matrix due to the noise in the process of camera motion estimation.

Measurement model. Let the measurement model be given as $h_i(k+1) = h(y_i, x(k+1|k))$. Observation for landmark is the three dimensional coordinate of landmarks based on the calibrated stereo camera coordinate system. Measurement prediction model of h_i is given as

$$
h_i(k+1) = [R][y_i - x'_v], \quad where \quad x'_v = \begin{bmatrix} x \\ 0 \\ z \end{bmatrix}, R = \begin{bmatrix} cos\theta & 0 & -sin\theta \\ 0 & 1 & 0 \\ sin\theta & 0 & cos\theta \end{bmatrix}. \tag{7}
$$

Three dimensional absolute coordinate of landmark is denoted as y_i. And h_i is the function to transform absolute coordinate to camera coordinate system. x'_v represents planar coordinate of the robot based on the absolute coordinate system. Rotation between the absolute coordinate system to the camera coordinate system is denoted by R. θ in matrix R denotes the orientation of the robot.

Observation and matching. Matching between landmarks and observed three dimensional coordinate of SIFT $z_j = [X_j \; Y_j \; Z_j\,]$ is accomplished through SIFT matching algorithm. To determine searching area for matching, we can use the innovation matrix and its covariance.

The innovation matrix $v_{ij}(k+1)$ between the predicted measurement $h_i(k+1|k)$ for landmark y_i and observation z_j is given as

$$
\begin{aligned}
v_{ij}(k+1) &= [z_j(k+1) - h(y_i, p(k+1|k))]\,, \\
\sigma^2_{IN,ij} &= \nabla_{x_v} h_i \cdot \sigma^2_{x_v x_v} \cdot \nabla_{x_v} h_i^T + \nabla_{x_v} h_i \cdot \sigma^2_{x_v y_i} \cdot \nabla_{y_i} h_i^T \\
&\quad + \nabla_{y_i} h_i \cdot \sigma^2_{y_i x_v} \cdot \nabla_{x_v} h_i^T + \nabla_{y_i} h_i \cdot \sigma^2_{y_i y_i} \cdot \nabla_{y_i} h_i^T + \sigma^2_{R,i},
\end{aligned} \tag{8}
$$

where $\sigma^2_{R,i}$ represents the covariance of the measurement.

Searching area is determined with the Mahalanobis distance and threshold constant g^2 such as

$$
v_{ij}^T(k+1) \cdot \sigma^{-2}_{IN,ij} \cdot v_{ij}(k+1) \le g^2 \tag{9}
$$

Using SIFT matching algorithm, matching is accomplished between the landmark y_i and the observation satisfying the criterion shown in (9).

Estimation of the system state vector and covariance. Kalman gain K can be calculated as

$$K = \sigma^2 \nabla_{x_v} h_i^T \cdot \sigma_{IN_i}^{-2} = \begin{pmatrix} \sigma_{x_v x_v}^2 \\ \sigma_{y_i x_v}^2 \\ \vdots \end{pmatrix} \frac{\partial h_i^T}{\partial x_v} \sigma_{IN_i}^{-2} + \begin{pmatrix} \sigma_{x_v y_i}^2 \\ \sigma_{y_i y_i}^2 \\ \vdots \end{pmatrix} \frac{\partial h_i^T}{\partial y_i} \sigma_{IN_i}^{-2}, \tag{10}$$

where $\sigma_{x_v x_v}^2$, $\sigma_{x_v y_i}^2$ and $\sigma_{y_i y_i}^2$ are 3×3 blocks of the current state covariance matrix Σ. $\sigma_{IN_i}^2$ is the scalar innovation variance of y_i.

The updated system state and its covariance can be computed as

$$\begin{aligned} p(k+1|k+1) &= p(k+1|k) + K \cdot v_i(k+1), \\ \sigma^2(k+1|k+1) &= \sigma^2(k+1|k) - K \cdot \sigma_{IN_i}^2(k+1) \cdot K^T. \end{aligned} \tag{11}$$

This update is carried out sequentially for each innovation of the measurement.

Registration and deletion of landmarks. When new three dimensional SIFT feature points are found on a recognized object using 3D OFM, absolute coordinates of those are computed using $y_n(x_v, h_n)$ of the measurement model. The computed absolute coordinates are added into the system state vector. We can also delete landmarks by deleting all rows and columns related to target landmarks from the covariance matrix.

3 Semantic-Map-Based Localization

The kidnapped robot problem refers the case where robot is lifted and manually repositioned in a different location in the environment, and has to relocate itself based on new sensor evidence. Our goal is to solve the kidnapped robot problems to locate robot position by only stereo vision information together with our semantic-map.

The relationship between absolute coordinate and relative robot coordinate system as shown in Fig. 2 is given as

$$\begin{bmatrix} x_{abs} \\ z_{abs} \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_{rel} \\ z_{rel} \end{bmatrix} + \begin{bmatrix} x_t \\ z_t \end{bmatrix}, \tag{12}$$

where θ is the angle between relative and absolute coordinate and $[x_t \ z_t]^T$ is the robot position in absolute coordinate. Subtraction with two matched 3D SIFT points, we can obtain the relation as

$$\begin{bmatrix} \Delta x_{abs} \\ \Delta z_{abs} \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} \Delta x_{rel} \\ \Delta z_{rel} \end{bmatrix}, \tag{13}$$

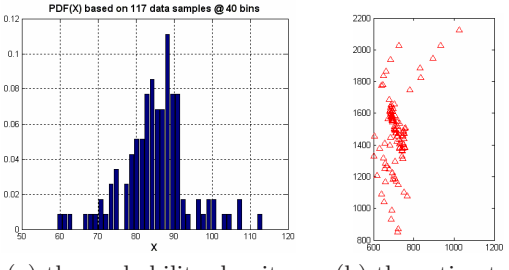
where $[\Delta x_{abs}\ \Delta z_{abs}]^T$ and $[\Delta x_{rel}\ \Delta z_{rel}]^T$ are the vectors of difference between matched points in absolute coordinate system, and in relative coordinate system, respectively. So, we can obtain the angle θ between the absolute coordinate system and the relative coordinate system by using two matched 3D SIFT points. The θ can be obtained by

$$cos\theta = \frac{[\Delta x_{abs}\ \Delta z_{abs}][\Delta x_{rel}\ \Delta z_{rel}]^T}{\|[\Delta x_{abs}\ \Delta z_{abs}]^T\| \cdot \|[\Delta x_{rel}\ \Delta z_{rel}]^T\|}. \tag{14}$$

And the robot location is computed by

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} x_{abs} \\ z_{abs} \end{bmatrix} - \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x_{rel} \\ z_{rel} \end{bmatrix}. \tag{15}$$

Ideally, if absolute and relative vectors of two pairs are correct, the robot can locate itself accurately. However, the estimated position has uncertainties because of errors in measurements and map building.



(a) the probability density function of estimated θ (b) the estimated location

Fig. 3. The statistics of the robot localization

As shown in Fig. 3, the angle distribution indicates Gaussian-like distribution. In such a distribution, this function has a peak at the mean value. Thus, the mean value of θ is determined as the direction of robot in the absolute coordinate system. The process of robot localization is summarized as follows;

1. Eliminate outliers of θ : eliminates the outside of standard deviation
2. Compare the magnitudes of absolute and relative vectors :
   If $\|[\Delta x_{abs}\ \Delta z_{abs}]^T\|/\|[\Delta x_{rel}\ \Delta z_{rel}]^T\| > 1.5$
   or $\|[\Delta x_{abs}\ \Delta z_{abs}]^T\|/\|[\Delta x_{abs}\ \Delta z_{abs}]^T\| < 0.66$, the robot position estimated by Eq. (14) is discarded.
3. Determine the angle of the robot : the averaged value of θ is adopted as the estimated angle of the robot in the absolute coordinate system.
4. Determine robot position with the θ : averaged vector of the robot location is chosen as the estimated location of the robot.

Fig. 4. The real scene of the indoor environment in which the map-building is performed

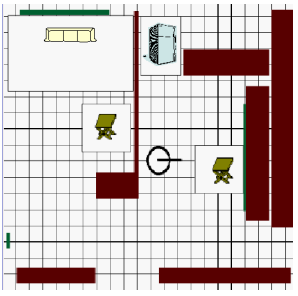


Fig. 5. The topological map of the experimental space

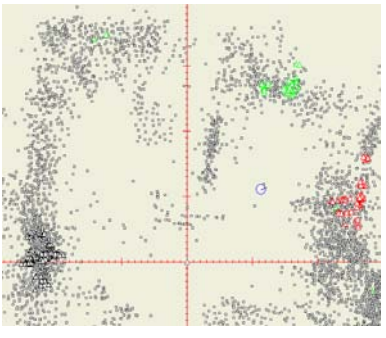


Fig. 6. The map building by only odometer

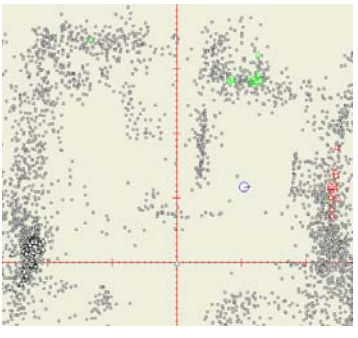


Fig. 7. The map building by odometer and 3D SIFT

4 Experimental Results

4.1 Performance Evaluation for Map-Building and Localization

To determine how much the map-building is accurate, we solve the kidnapped robot problem at a given position of the robot. By comparing the solution of kidnap problem with the given position, we can measure the accuracy of the map.

Figure 4 shows the scene of the real indoor environment in which the map-building is performed. The circle in Fig. 5 represents the actual position of robot in the absolute coordinate system. And, robot is made to find its location by solving the kidnapped robot problems with one of maps in Fig. 6 to Fig. 9 as well as 3D SIFT features, not using odometer information. The resulting estimated position in each map is marked with circle as shown in Fig. 6 to Fig. 9.

Figure 6 shows the map which is built by relative localization by odometer. we abbreviate this method as ODO-LOC. Due to noisy odometer information, we obtain the distorted map as in Fig. 6.

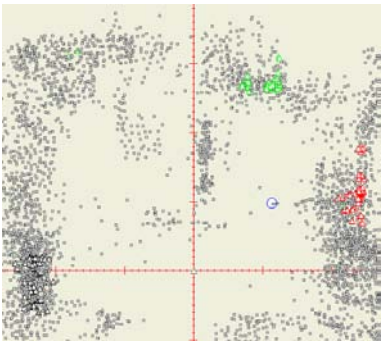


Fig. 8. The map building by odometer and 3D Harris corners

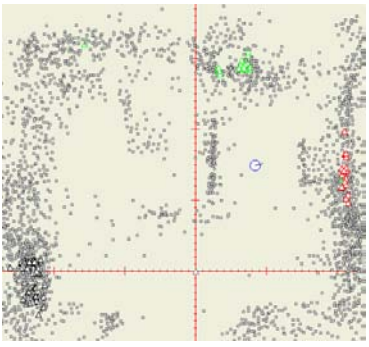


Fig. 9. The map building by 3D Harris corners and 3D SIFT

Table 1. Averaged error between the robot actual position and the position estimated

METHOD	X (mm)	Z (mm)	θ (degree)
ODO-LOC	277.17	880.19	19.47
ODO-SIFT-KAL	107.37	558.70	10.34
HARI-LOC	174.30	758.83	8.85
HARI-SIFT-KAL (proposed method)	127.01	218.51	8.87

Figure 7 shows the map which is built by extended Kalman filter using odometer and 3D SIFT features. We will call this method as ODO-SIFT-KAL. As shown in Fig. 7, the map appears to be more accurate when compared with the map built by ODO-LOC. This can be also observed from Table 1 by comparing solutions of kidnap problems for the ODO-LOC map and ODO-SIFT-KAL map.

Figure 8 shows the map which is built by relative localization by optical flows of 3D Harris corners. We abbreviate this method as HARI-LOC. Figure 9 shows the map-building by our proposed method in Sec. II, which will be called as HARI-SIFT-KAL. From Table 1, the map by HARI-SIFT-KAL is observed to be the most accurate map in this experiment.

5 Concluding Remarks

In this work, we have proposed autonomous semantic-map building combined with vision-based robot localization. We have used nonsymbolic SIFT-based object features with symbolic object names as landmarks and we have used vision-based relative localization as the process model of our EKF. Thus our method is able to be applied successfully to environments in which encoders are not available, such as humanoid robots.

In most previous methods, landmarks have been composed of points without semantics. Thus, an additional anchoring technique was often required for

interaction. However, in our work, symbolic object names with their 3D feature location have been used as landmarks. Using such symbolic object names as landmarks is very useful when humans interact with the robot.

We also solve the kidnapped robot problem to determine how much the map-building is accurate. Note that the more is a map accurate, the more is correct the solution of kidnap problem. Consequently, in our proposed HARI-SIFT-KAL map-building, the solution of kidnap problem shows the most accurate result. Thus, it is concluded that the map by HARI-SIFT-KAL appears to be most accurate in the map-building.

Acknowledgement

This work is supported by the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Korea Ministry of Commerce, Industry and Energy. And, all correspondences of this work should be addressed to I. H. Suh.

References

1. A. J. Davison and W. Murray: Simulatoneous Localization and Map-Building Using Active Vision. IEEE Transaction on Pattern Analysis and Machine Intelligence. Vol. 24 (2002) 865–880
2. A. J. Davision: Real-Time Simultaneous Lacalisation and Mapping with a Single Camera. Proceedings of Ninth IEEE International Conference on Computer Vision. Vol. 2 (2003) 1403–1410
3. M.W.M. G. Dissanayake, P. Newman, S. Clark, and H. F. Durrant-Whyte: A Solution to the Simultaneous Localization and Map Building (SLAM) Problem. IEEE Transaction on Robotics and Automation. Vol. 17 (2001) 229–241
4. N. Doh, H. Choset, and W. K. Chung: Accurate Relative Localization Using Odometry. Proceedings of IEEE International Conference on Robotics and Automation. Vol.2 (2003) 1606–1612
5. C. Harris and M. J. Stephens: A Combined Corner and Edge Detector. Alvey Vision Conference. (1988) 147–152
6. D. G. Lowe: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. (2004) 91–110
7. B. Lucas and T. Kanade: An Interative Image Registration Technique with an Application To Stereo Vision. Proceedings of DARPA IU Workshop. (1981) 121–130
8. P. S. Maybeck: The Kalman Filter:An Introduction to Concepts. Spinger-Verlag. (1990)
9. I. Shimshoni: On Mobile Localization From Landmark Bearings. IEEE Transactions on Robotics and Automation. Vol. 18, No. 6 (2002) 971–976
10. S. Se, D. G. Lowe, and J. Little: Mobile Robot Localization And Mapping with Uncertainty using Scale-Invariant Visual Landmarks. International Journal of Robotics Research. Vol. 21 (2002) 735–758

LRMAP: Lightweight and Resynchronous Mutual Authentication Protocol for RFID System*

JeaCheol Ha[1], JungHoon Ha[2], SangJae Moon[2], and Colin Boyd[3]

[1] Dept. of Information Security, Hoseo Univ., 336-795, Korea
jcha@hoseo.edu
[2] School of Electrical Eng. and Computer Science, Kyungpook National Univ., 702-701, Korea
{short98, sjmoon}@ee.knu.ac.kr
[3] Information Security Institute, Queensland Univ. of Technology, GPO Box 2434, Brisbane, QLD, 4001, Australia
boyd@isrc.qut.edu.au

Abstract. Despite various solutions to the security problems in an RFID system, most are unable to fully support all the security requirements. Plus, when designing a viable RFID system, account should also be taken of the computational load on the back-end database and restricted capacity of a tag. Accordingly, an efficient RFID protocol is proposed to reduce the computational load on both the back-end database and the tags, while also guaranteeing most security requirements for RFID wireless communication, including untraceability, authentication, and robustness against replay and spoofing attacks. Plus, in the case of desynchronization resulting from communication failure or malicious attack, the proposed scheme can recover synchronization between the database and the tag.

Keywords: RFID system, Mutual authentication, Privacy, Traceability, Desynchronization attack.

1 Introduction

Radio Frequency Identification (RFID) systems, a new form of automatic identification technology involving the use of small devices called RFID tags, are expected to replace optical barcodes due to several important advantages, including a low cost, small size, quick identification, and invisible implementation within objects. An RFID system consists of RFID tags, an RFID reader, and a back-end database. Yet, since the RFID reader communicates with the tags using RF signals, existing RFID protocols still suffer from various weaknesses, including location privacy, authentication, and resynchronization between two entities. One solution to protect tags from attack is mutual authentication between the

* This research was supported by the MIC of Korea, under the ITRC support program supervised by the IITA(IITA-2006-C1090-0603-0026).

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 80–89, 2007.

tag and the reader. Thus, a lightweight authentication protocol is needed that takes account of the tag's design limitations, restricted implementation cost, and back-end server's capacity.

Several studies have already attempted to resolve the authentication problem between the tag and the reader using physical technologies, including the 'Kill command' [12], 'Active jamming' [5], and 'Blocker tag' [5] approaches. Then, in 2004, Weis *et al.* [10,11,12] proposed a hash-lock protocol and randomized hash-lock protocol as cryptographic solutions. In another approach based on a hash function, Henrici and Müller [3] proposed an ID variation protocol. While this protocol is secure against a replay attack, as the identity of a tag is renewed in each session, location privacy is still compromised, since the tag's response remains constant until the next authentication session when desynchronization occurs [9]. Ohkubo *et al.* [8] also proposed a hash chain-based authentication protocol in which the reader sends a query using two different hash functions. However, this scheme is still vulnerable to a replay attack and spoofing attack, and imposes a heavy burden on the back-end database to authenticate the tag. Rhee *et al.* [9] proposed a challenge-response authentication protocol based on a hash function that is robust against a spoofing attack and replay attack, plus location privacy is also guaranteed. However, the computational load on the back-end database is still heavy when authenticating a tag. The RFID mutual authentication scheme based on synchronized secret information presented by Lee *et al.* [6] also requires many computational operations in the back-end database. Thus, in 2005, Lee *et al.* [7] proposed a low-cost RFID authentication scheme in which a tag and the back-end database only perform two one-way hash operations. Yet, this scheme is also vulnerable to a spoofing attack and location-tracing attack when desynchronization occurs. Recently, Dimitriou [2] proposed a lightweight RFID authentication protocol that enforces user privacy and protects against cloning. However, there is no method for recovering synchronization when a state of desynchronization occurs, where one tag blocks any further tag functionality.

Accordingly, this paper proposes a lightweight and resynchronous mutual authentication protocol (LRMAP) for an RFID system. When a desynchronization problem arises between the back-end database and a tag due to communication failure or a malicious attack, the proposed scheme stays robust and recovers the synchronization. In addition, the computational load on the back-end system is efficient, as a different ID searching method is applied according to the state of the previous session. Moreover, the proposed protocol is secure against location tracing, a replay attack, and spoofing attack, plus mutual authentication is guaranteed between the back-end database and an RFID tag.

2 RFID System

2.1 Composition of RFID System

An RFID system typically consists of three elements, such as RFID tags, (*transponders*), the RFID reader (*transceiver*), and back-end database (*Back-end server*).

- **RFID tag.** An RFID tag generally consists of a microchip for computing and a coupling element, such as an antenna, for wireless communication. A passive RFID tag does not possess an on-board power source, but is powered by the electromagnetic waves from the reader. Meanwhile, an active tag contains an on-board power source, such as a battery. In addition, the tags are categorized into several types according to their physical characteristics and application [1].
- **RFID reader.** The RFID reader interrogates the tags through an RF interface, then transmits the collected data to back-end database. The reader can also read and write the tag data. The channel from the reader to a tag, referred to as the *forward* channel, is insecure, as it is based on an air interface. Similarly, the channel from a tag to the reader, known as the *backward* channel, is also insecure.
- **Back-end database.** The back-end database receives data from the reader and provides certain services to a specific tag, such as product and prices information etc. The communication between the reader and the database is considered as a secure channel.

2.2 Security Requirements for RFID System

Since the communication between the reader and a tag is performed using an air interface, the communicated data can easily be tapped by an attacker. Therefore, various requirements are needed for a secure RFID protocol, as identified in previous literature [4,6,10].

- **Eavesdropping.** An attacker can eavesdrop messages between the reader and tags due to wireless communication, then use secret information or useful messages to perform various enhanced attacks, such as a replay attack or spoofing attack. Therefore, an RFID system should be designed to protect against the leakage of secret information.
- **Spoofing.** An adversary sends a malicious query to a targeted tag, then collects the response messages emitted by the tag. Thereafter, the attacker can impersonate the reader using the messages collected from the tag. On the other hand, an adversary can reply to the reader's query by impersonating a tag.
- **Location tracking.** The adversary seeks information on a tag's location track information. Thus, for perfect location privacy, an RFID system should satisfy both indistinguishability and forward security, where the former means that the values emitted by one tag should not be distinguishable from the values emitted by other tags, while the latter means even if an attacker obtains the secret data stored in a tag, the location of the tag can not be traced back using previous known messages, *i.e.*, disclosed data or communication information.
- **Message Interrupt.** The communication messages between the tags and the reader can be interrupted when an attacker tries to block the service. As a result, a message interrupt attack can create a state of desynchronization between the tag and the back-end database, due to an abnormal closing of a session, message blocking, or different ID updating of two entities within one session.

3 Related Work

3.1 *ID* Variation Protocol

Henrici and Müller [3] proposed an ID variation protocol that changes the identity of a tag in each session. Although this protocol is secure against a replay attack, as the ID of a tag is refreshed in each session by a random number, a spoofing attack can be applied, where an attacker impersonates the reader. Meanwhile, for location tracking, the attacker does not transmit the last message of the protocol, then since the tag then thinks that the information is lost, it does not update its ID [9]. As a result, the protocol has a database desynchronization problem. If the ID of a tag is desynchronized, the tag can be easily traced, as one of emitted values of the tag $H(ID)$ will be identical, thereby compromising the location privacy. This is called a ***desynchronization attack*** in which the attacker traces the tag's location using successive desynchronizations.

3.2 Challenge-Response-Based Authentication Protocol

Rhee *et al.* [9] proposed a challenge-response authentication protocol based on a hash function. This scheme is robust against a spoofing attack and replay attack. In addition, location privacy is guaranteed, as the tag transmits a different response in each session using a random number received from the reader. Nonetheless, the scheme is inefficient in terms of the computational load, as the back-end database is required to perform an ID search to find the specific information related to the tag requesting authentication.

3.3 Low-Cost Authentication Protocol: LCAP

Lee *et al.* [7] proposed a low-cost authentication protocol, LCAP, that only involves two one-way hash function operations in a tag, making it quite efficient. Although location privacy is supposedly guaranteed, the scheme is still vulnerable to location tracing, as a tag will respond to the same $H(ID)$ in the case the last message from reader is not received due to a message interrupt. Therefore, this protocol is vulnerable to location tracing using successive desynchronization attacks.

3.4 Lightweight Challenge-Response Protocol

Recently, Dimitriou [2] proposed a lightweight RFID authentication protocol that enforces user privacy and protects against cloning. However, an attacker can still block the final message transmitted from the reader to the tag. In the resulting state of desynchronization, the tag and back-end database update using different keys, thereby blocking any further tag functionality. In addition, an attacker can trace a tag by repeatedly sending a query from the reader. As a tag will respond with the same message $H(ID_i)$ in which ID_i is fixed in a desynchronized session, the tag cannot satisfy indistinguishability.

4 Proposed Authentication Protocol: LRMAP

This section presents the proposed lightweight and resynchronous mutual authentication protocol (LRMAP) for an RFID system.

4.1 Notations

The following notations are used for the entities and computational operations to simplify the description.

T	: RFID tag or transponder
R	: RFID reader or transceiver
DB	: back-end database or back-end server
ID	: identity of a tag, L bits
HID	: hashed value of ID, L bits
PID	: previous identity of a tag used in previous session, L bits
r_R	: random number generated by reader R
r_T	: random number generated by tag T
$Query$	: request generated by R
$SYNC$	: parameter used to check whether both T and DB succeeded in ID updating simultaneously or not, 1 bit
$H()$	: one-way hash function, $H : \{0,1\}^* \rightarrow \{0,1\}^l$
$L(m)$	: left half of input message m
$R(m)$	: right half of input message m
$\|\|$	: concatenation of two inputs
$\stackrel{?}{=}$	: comparison of two inputs

4.2 System Model and Protocol

To define the model of the proposed lightweight and resynchronous mutual authentication protocol (LRMAP), the RFID system consists of three entities, the tag T, reader R, and back-end database DB. T emits $P = H(ID)$ or $P = H(ID\|r_T)$ according to the state of $SYNC$ in response to a query from R. That is, if T does not receive the last message from R due to a communication malfunction or the verification procedure fails due to a malicious attack, the $SYNC$ value is set as 1 and T responds with $P = H(ID\|r_T)$ in the next session. In the case the protocol finishes normally, the $SYNC$ value becomes 0 and T transmits $P = H(ID)$ to R in the next session. DB manages the ID, hashed values HID, and PID for each T in the database field. According to the state of the previous session, *i.e.*, the value P received from T, DB finds ID for the current session or PID used for the previous session by comparing the received P with the HID and PID in the database field. It is assumed that the communication channel between R and DB is secure, while the communication channel between R and T is insecure. Fig. 1 shows the

Database	Reader	Tag
Database field [ID][HID][PID]		Tag field [ID][SYNC]
	$\xrightarrow{Query,\ r_R}$	If($SYNC \stackrel{?}{=} 0$)$P = H(ID)$ else $P = H(ID\|\|r_T)$ $Q = H(ID\|\|r_T\|\|r_R)$
If($P \stackrel{?}{=} HID$)$PID = ID$ else if($P \stackrel{?}{=} H(ID\|\|r_T)$ $PID = ID$ else if($P \stackrel{?}{=} H(PID\|\|r_T)$ $PID = PID$ else halt $Q' = H(PID\|\|r_T\|\|r_R)$	$\xleftarrow{P,L(Q),r_T,r_R}$	$\xleftarrow{P,L(Q),r_T}$ $SYNC = 1$
$L(Q') \stackrel{?}{=} L(Q)$ $ID = H(PID\|\|r_R)$ $HID = H(ID)$	$\xrightarrow{R(Q')}$	$\xrightarrow{R(Q')}$ If($R(Q) \stackrel{?}{=} R(Q')$) { $ID = H(ID\|\|r_R)$ $SYNC = 0$ }

Fig. 1. The proposed lightweight and resynchronous mutual authentication protocol

process of the proposed LRMAP, and the following gives a detailed description of each step:

1. R chooses a random number r_R and broadcasts it to T with a $Query$.
2. T selects a random number r_T and computes P differently according to the state of $SYNC$. That is, if the $SYNC$ value is 0, then $P = H(ID)$, otherwise T computes $P = H(ID\|r_T)$ using r_T generated by itself. It then computes $Q = H(ID\|r_T\|r_R)$ and sets the $SYNC$ field as 1. T transmits $P, L(Q)$ and r_T to R in response to the $Query$, where $L(Q)$ is the left half of Q.
3. R forwards the message $P, L(Q)$ and r_T received from T to DB together with r_R generated by itself in step 1.
4. DB firstly compares the received $P = H(ID)$ with the HID values saved in the database. If the values match, DB regards the ID as the identity of T requesting authentication. This is a general case when the previous session is closed normally. If DB cannot find the HID in the first searching case then it secondly computes $H(ID\|r_T)$ value with the received r_T and compares it with P. If the tag's response messages were blocked in the previous session, that is, the $SYNC$ value will be 1 and two IDs in the DB and tag will not be updated, then the DB finds a match with the ID of T in the second searching case. However, if DB cannot find the ID of tag in above two cases, then it thirdly computes $H(PID\|r_T)$ value and compares it with P. The DB finds a match with the PID of T when the reader's last messages were blocked in the previous session, that is, the $SYNC$ value will be 1 and DB will update the ID, yet the tag's ID will not be updated. Unfortunately, if DB cannot find the identity of

T in above three cases, it halts the searching of ID and can order R to query again in order to restart the process from the first step. If DB finds the ID or PID in three searching cases, then it computes $Q' = H(PID\|r_T\|r_R)$[1] and verifies that the following equation is satisfied:

$$L(Q') \stackrel{?}{=} L(Q). \tag{1}$$

If equation (1) is satisfied, DB computes $R(Q')$, transmits it to R, and updates the HID for the next session. That is, it computes $ID = H(PID\|r_R)$ and updates the $HID = H(ID)$.

5. R delivers the message $R(Q')$ received from DB to T.
6. To verify the correctness of $R(Q')$, T tests the following equation:

$$R(Q) \stackrel{?}{=} R(Q'), \tag{2}$$

where $R(Q)$ is the right half of $Q = H(ID\|r_T\|r_R)$ computed by itself in step 1. If equation (2) is correct, T updates the identity as $ID = H(ID\|r_R)$, then sets the $SYNC$ value at 0.

5 Analysis

5.1 Security

The security of the proposed LRMAP was evaluated against the threats described in Section 2.

- **Eavesdropping.** To obtain secret information from a tag, an adversary must be able to guess the ID after collecting the communication messages. However, an adversary cannot extract the ID from the $H(ID)$ or $H(ID||r_T)$ due to the security property of a one-way hash function. Otherwise, the adversary has to compute a correct string $L(Q)$ from a known r_T and r_R, which is also hard due to their one-way property. A replay attack cannot compromise the proposed protocol, as the $H(ID)$ or $H(ID||r_T)$ is refreshed by updating the ID or including a random number r_T in each session. Therefore, the proposed LRMAP can defeat a replay attack due to the freshness of the communication messages.
- **Spoofing.** Here, an adversary collects a tag's response, then tries a spoofing attack to impersonate a legitimate tag. However, an adversary cannot compute the hashed messages P and $L(Q)$ without knowing the ID value. Meanwhile, to impersonate as the reader, an adversary must transmit the correct $R(Q)$. This is also impossible, because an adversary cannot compute Q without knowing the ID. Thus, it is impossible to impersonate a tag or the reader using a spoofing attack.

[1] Since ID is updated into PID after finding ID from HID, $Q' = H(PID\|r_T\|r_R)$ is computed regardless of PID or ID.

- **Location tracking.** The proposed protocol guarantees location privacy by refreshing the ID in the tag and back-end database for each session. After the successful authentication is finished in the previous session, the $SYNC$ value is set at 0. Thus, indistinguishability is satisfied with a one-way hash function in which the input of the previous session is refreshed. In contrast, if the previous session is not closed normally, the $SYNC$ value is set at 1. Here, indistinguishability is also satisfied using a one-way hash function, as the input is refreshed by a random number r_T. That is, the value P transmitted from the tag is not $H(ID)$ but $H(ID\|r_T)$. As regards forward security, this assumes that an attacker can obtain a tag's correct ID at some point. However, no previous ID can be extracted due to the one-way property of a hash function. That is, it is impossible to recover the ID from $H(ID\|r_R)$, making it impossible for an attacker to trace the location of a tag backwards. Unfortunately, this protocol may be impossible to satisfy forward security while successive desynchronizations are occurred. An adversary can collect the communication messages and continuously make last massage $R(Q')$ invalid up to the time obtaining a target secret ID. After obtaining the secret ID of tag, the adversary may make it possible to trace the some past histories of T while ID of tag was not changed because he knows the previous P and r_T. Therefore, LRMAP perfectly satisfies the forward security property from setup time to the latest point occurred a successful authentication.
- **Message Interrupt.** In the first case, it is assumed that an adversary can block the response messages transmitted from a tag, *i.e.*, step 2 of LRMAP. At this point, as the reader does not know of the tag's existence, the $SYNC$ value for the tag is set at 1, plus, if the tag does not receive a response from the reader within a predefined time, the tag sends $H(ID||r_T)$ as a response to a query from the reader in the next session. Nonetheless, the two entities T and DB can still recover the synchronization by finding the current ID in the back-end database. In the second case, if an attacker blocks the last messages transmitted from the reader, the DB already knows of the tag's existence and updates the ID value, while the $SYNC$ value for the tag is set at 1. Therefore, when a tag sends $H(ID||r_T)$ as the response in the next desynchronized session, the two entities can recover the synchronization based on finding the PID in the back-end database. Therefore, LRMAP can be protected against messages loss due to an attacker in a wireless channel.

A security comparison with previous authentication protocols is shown in Table 1. The proposed LRMAP is secure against most attacks presented up to now, including a replay attack, spoofing attack, location tracing attack, and desynchronization attack.

5.2 Efficiency

When evaluating the computational load and storage cost for the two entities, as shown in Table 1, the LRMAP exhibited a remarkable improvement in the computational cost for the DB. Even though the challenge-response-based protocol

Table 1. Comparison of security and efficiency

Protocol	Henrici[3]	Rhee[9]	Lee[7]	Dimitriou[2]	LRMAP
Replay attack	O	O	O	O	O
Spoofing attack	×	O	×	O	O
Indistinguishability	×	O	×	×	O
Forward security	△	×	×	△	△
Resynchronization	O	O	O	×	O
ID refreshment	O	×	O	O	O
Comp.(hash # of DB)	3	$m/2+2$	2	4	3*
Comp.(hash # of tag)	3	2	2	4	3
Storage of DB(bits)	$8L \cdot m$	$L \cdot m$	$6L \cdot m$	$2L \cdot m$	$3L \cdot m$
Storage of tag(bits)	$3L$	$1L$	$1L$	$1L$	$1L+1$

*: $m+3$ to recover the synchronization on average.
O: secure or support, △: partially secure ×: insecure or not support,
m: the number of IDs.

[9] satisfies most security items, except forward security, its critical disadvantage is that the DB is required to performs $m/2+2$ hash operations to authenticate a tag. In contrast, the proposed protocol only requires 3 hash operations in the DB and tag, respectively, even though m is large. In the case of desynchronization, the correct ID or PID can be found based on an average of $m/2+3$ or $m+m/2+3$ hash operations. So we can say that the recovery time in desynchronization state is $m+3$ operations on average. However, since desynchronization of a tag is a special and unusual state, the normal synchronization state only requires 3 hash operations.

With the proposed protocol, the storage size of the DB is $3L * m$, where L is the length of an ID or hashed value and m is the number of IDs. Plus, a tag needs $(L+1)$-bits of memory to store an ID and the $SYNC$ value. The length of the total message transmitted from a tag to the reader is $2.5L$, while that from the reader to a tag is $1.5L$, except for a $Query$. Therefore, the LRMAP is suitable for an RFID systems with limited memory space and computational power.

6 Conclusion

A lightweight and resynchronous mutual authentication protocol (LRMAP) was proposed to protect an RFID system against existing attacks. The proposed protocol guarantees untraceability, authentication, and robustness against replay and spoofing attacks. Furthermore, even though the protocol can fall into a desynchronized state as a result of a malicious attacker, synchronization between the database and a tag can be recovered in the next session. As the regards the computational cost, the LRMAP is designed to reduce the computational load on both the back-end database and the tags. Consequently, the proposed scheme can be used in low-cost RFID systems, as in a normal state, the correct ID is found using a comparison of the transmitted hash message with the hashed values in the DB.

References

1. Auto-ID Center. Draft Protocol Specification for a Class 0 Radio Frequency Identification Tag, February, 2003.
2. T. Dimitriou, A lightweight RFID protocol to protect against traceability and cloning attacks. Security and Privacy for Emerging Areas in Communications Networks-2005. SecureComm 2005, pp. 59-66, Sept., 2005
3. D. Henrici and P. Müller. Hash-based Enhancement of Loaction Privacy for Radio Frequency Identification Devices using Varing Identifiers, In *proceeding of the Second IEEE Annual Conference on Pervasive Computing and Communiations Workshops*, pp. 149-162, IEEE, 2004
4. A. Juels. RFID Security and Privacy: A Research Survey. *RSA Laboratories*, 2005.
5. A. Juels, R. L. Rivest and M. Szydlo. The Blocker Tag: Selective Blocking of RFID Tags for consumer Privacy. In *Proceeding of 10th ACM Conference on Computer and Communications Security'03*, pp. 103-111, 2003.
6. S. Lee, T. Asano and K. Kim. RFID Mutual Authentication Scheme based on Synchronized Secret Information. In *proceedings of the SCIS'06*, 2006.
7. S. Lee, Y. Hwang, D. Lee and J. Lim. Efficient Authentication for Low-cost RFID Systems. *ICCSA'05*, NCS 3480, pp. 619-627, Springer-Verlag, 2005
8. M. Ohkubo, K. Suzuki and S. Kinoshita. Hash-Chain Based Forward-Secure Privacy Protection Scheme for Low-Cost RFID. In *proceedings of the SCIS'04*, pp. 719-724, 2004.
9. K. Rhee, J. Kwak, S. Kim and D. Won. Challenge-Response Based on RFID Authentication Protocol for Distributed Database Envirionment. *SPC'05*, LNCS 3450, Springer-Verlag, 2005.
10. S. E. Sarma, S. A. Weis and D. W. Engels. Radio-Frequency Identification: Security Risks and Challenges. *RSA Laboratories*, Volume 6, No. 1, Spring, 2003.
11. S. A. Weis. Security and Privacy in Radio-Frequency Identification Devices. MS Thesis, MIT, 2003
12. S. A. Weis, S. E. Sarma, R. L. Rivest and D. W. Engles. Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems. *Security in Pervasive Computing'03*, LNCS 2802, Springer-Verlag, 2004.

Novel Process Methodology of Smart Combi Card (SCC) Manufacturing for RFID/USN

JeongJin Kang[1], HongJun Yoo[2], and SungRok Lee[3]

[1] Dong Seoul College, Korea
jjkang@dsc.ac.kr
[2] JT Corp, Korea
junyoo@jtcorp.co.kr
[3] Patent Attorney DK CHOI and Partners, Korea
patentist@dkchoi.com

Abstract. We propose a novel process for manufacturing smart combi card(SCC). The proposed process excludes a milling process for grinding a cavity and conductive adhesive which uses in a conventional process. Direct bonding including welding and soldering can be performed between an antenna coil and RF interface of a COB (Chip On Board) module. This brings the electrical connection to high reliability in order to hold RF functionality of a Smart combi card(SCC) with a terminal good. Hot melt sheet is employed between the COB module and the card body, and it is melt and stiffened through thermal lamination. This results in good adhesion between the COB module and the card body. Thus, the proposed process can provide the smart combi cards having good physical properties against repetitive bending and torsion stresses. The cards manufactured by the proposed process held their resonance frequencies for RF communication stable over 5,000 times of stresses. Therefore it can be applicable to the useful smart card for Radio Frequency Identification(RFID)/Ubiquitous Sensor Network(USN) environments.

Keywords: Smart combi card(SCC), RF card, COB(Chip On Board), RFID/USN.

1 Introduction

Magnetic striped cards have some drawbacks of having small data storage capacity, losing their stored data when exposed in electromagnetic fields, and having a very low degree of security. This leads for the magnetic striped cards to be replaced with smart cards containing an integrated chip module inside of a plastic card body. Moreover, in accordance with EMV (Europay-Master-Visa) 2000, a recent version of EMV Standards, all of the magnetic striped cards used all over the world should be changed to smart cards, also called 'IC cards', having a contact interface for data communication until 2006. There are three types of IC cards in view of data transfer type. They are a contact card having a contact electrode exposed on the pre-defined position of a card body, a contactless

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 90–99, 2007.

IC card having RF interface with an antenna coil inside of a card body, and a Dual Interface IC Card, also called 'Combi Card', having both of the contact and contactless interfaces. The Smart combi card(SCC) can carry out multi-applications such as a credit card, a debit card, an electric purse, payment for transit fare, a residence identification and so on. Under EMV Standards, the SCC will be more popular in the near future.[1-3] For those multi-applications, important are the physical properties of the SCC including electrical reliability and durability against physical or mechanical stresses. However, the combicard manufactured by conventional process does not satisfy such property and has fatal defects that lead to occur gradual malfunction. In this paper the conventional process and its defects are discussed and then an improved process is proposed. Various physical tests are performed for the SCCs manufactured by the conventional and proposed processes

2 Conventional Manufacturing Process for Smart Combi Card(SCC)

The conventional process comprises the following steps:

- **a.** Preparing an antenna sheet and card body lamination by stacking an antenna sheet, printed sheets and overlay sheets, and laminating the sheets with heat,
- **b.** Milling a cavity on a card body for a COB (Chip-On Board) module to be embedded, and
- **c.** Embedding the COB module using conductive adhesive between the COB module and the cavity, and personalization process of individual card.

2.1 Loop Antenna Winding and Lamination Process for Card Body

A loop antenna for a contactless interface is winded on a PVC (polyvinyl chloride) sheet. The loop antenna is commonly made of a copper. The both ends of the antenna should be placed in a predetermined position in order to be connected with a pair of antenna electrodes of the COB module. This sheet is called an 'antenna sheet'.

The antenna sheet is sandwiched between printed sheets. Finally two transparent overlay sheets are added on tops of the printed sheets. This procedure is schematically shown in Fig. 1.

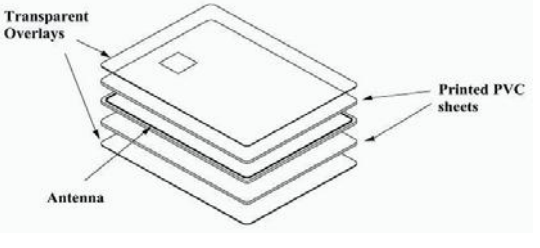


Fig. 1. Laminating sheets for a card body

2.2 Milling Process for Cavity in Which COB Module Is Embedded

This step is a process for cutting out a groove on a predefined position using a milling apparatus to be a cavity into which the COB module is inserted. The position of COB module is strictly defined by international standards including ISO 7816. It is important to control the milling machine so that the antenna contacts are exposed in the cavity as shown in Fig. 2, 3.

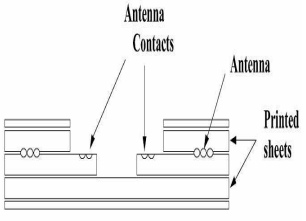


Fig. 2. Forming a cavity for a COB module

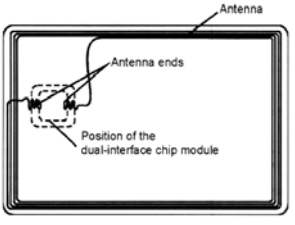


Fig. 3. Exposed antenna ends for being connected with a COB module after milling

2.3 Embedding Process Using Conductive Adhesive and Personalization Process

After the antenna contacts are exposed by milling process, conductive adhesive (or conductive glue) is applied between antenna electrodes of the COB module and the antenna contacts of the antenna in the cavity as shown in Fig. 4. The conductive glue is an adhesive in which conductive particles are dispersed as shown in Fig. 5. When the COB module is attached to the cavity, the conductive particles placed between the RF interface of COB module and the antenna contacts make electrical connection of the COB module and the antenna coil. Cross-sectional view of COB embedding is showen in Fig. 6. Personalization process follows the embedding process to provide finished cards. The personalization changes an anonymous card to an individualized card by attaching a hologram and embossing a user's name, valid dates and so on.

2.4 Fatal Defections of the Conventional Smart Combi Cards

The Smart Combi Cards prepared by such conventional process have fatal defects as follows:

(1) If physical stresses such as bending and torsion are given to the cards continuously and repeatedly, the electrical connection between the COB module and the antenna contacts becomes weak, which leads for RF communication to fail badly. It is called 'gradual malfunction'. The reason why the gradual malfunction occurs is because of weak adhesion of the conductive adhesive. Recently, it is reported that failure rates of 10% to 25% after one month of use are typical in the conventional dual-interface cards. [3]

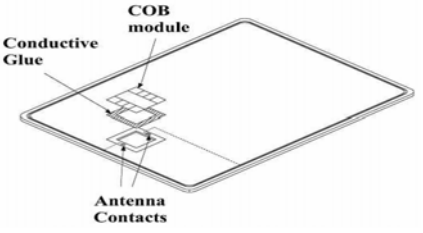


Fig. 4. Embedding COB module with conductive glue

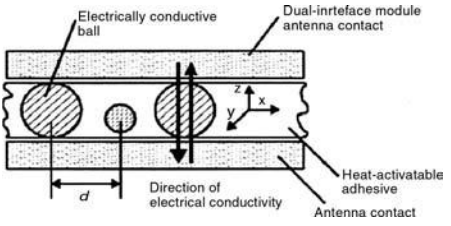


Fig. 5. Principle of electrical connection through conductive glue, an anisotropic conductive heat-activatable adhesive

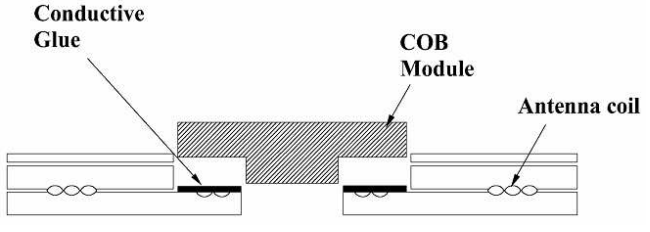


Fig. 6. Cross-sectional view of COB embedding

(2) If the depth of the cavity is not smooth, the electrical connection becomes poor in that milling process exposes the antenna contacts of the card body. The milling machine is rather expensive and the process using the machine is delicate and laborious work.
(3) In order to expand areas contacted with the conductive adhesive, the antenna contacts have the 'S' shape strands as shown in Fig. 7. Unfortunately, the 'S' shaped strands increase resistance of the electrical connection and then electrical reliability becomes poor.
(4) A gap or a crack exists between the COB module and the cavity of the card body. Humidity can permeate into the gap or the crack, and then the finished card is weak for thermal changes.

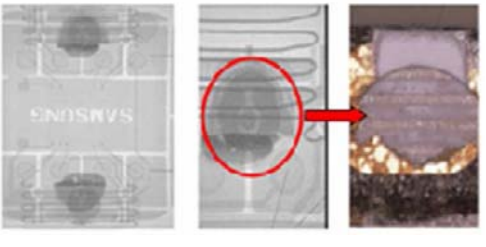


Fig. 7. 'S' shaped strands between the antenna contacts and RF interface of COB module

3 Proposed Process for Manufacturing Smart Combi Card(SCC)

3.1 Overview

The proposed manufacturing process is focused on reinforcing the bonding strength between the COB module and the antenna contacts, and ensuring the electrical reliability for RF communication. This process excludes the milling process of the conventional method. Instead, each of PVC sheets has a punched hole for the COB module to be inserted. That is, punched sheets are stacked on the COB module one another after as shown in Fig. 8. This process enables for the antenna contacts to be bonded to the RF interface of the COB module directly by soldering or welding.[6] The proposed process comprises the following steps:

a. Preparing an inlay sheet comprising an antenna sheet and a core sheet,
b. Pre-attaching the COB module to the punched hole of the inlay sheet with a hot melt sheet, and bonding the RF interface with the antenna, and
c. Applying the fillers to the antenna contacts and the upper side of the COB module, and laminating and thermal pressing the sheets with printed sheets having the same hole for COB module and overlay sheets.

3.2 Preparing Inlay Sheet

An antenna sheet using this process is a PVC sheet which has punched holes for a molding part, an encapsulated epoxy layer, of COB module and antenna contacts. An antenna coil is winded on the punched sheet. Both ends of the antenna coil have ⊏ shape respectively. It is important that the antenna contacts lie in straight lines on the antenna electrodes of the COB module as shown in Fig. 9.

Another PVC sheet having the same shape of the antenna sheet, a core sheet, is laminated on the antenna-winded sheet and then an inlay sheet is prepared. Both sheets are hot-laminated for 0.5 hours at about 70-100.

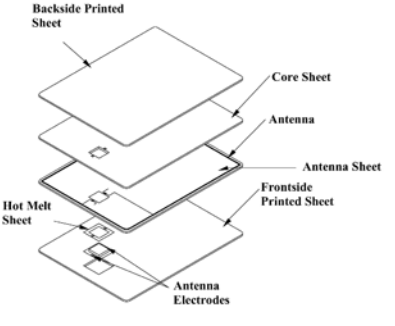


Fig. 8. Schematic view of the proposed process

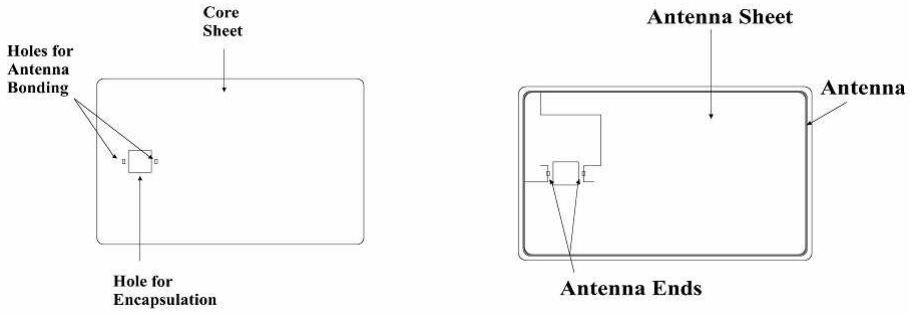


Fig. 9. A core sheet and an antenna sheet for an inlay sheet used in the proposed method

3.3 COB Module Pre-attaching with Hot Melt Sheet and Direct Antenna Bonding

After attaching hot melt tape to the COB module, the COB module is inserted into the hole of the inlay sheet. Although the adhesion of the hot melt sheet is rather insufficient at room temperature, the hot melt sheet provides an excellent adhesive strength when hot-laminated. By this reason, such insertion of the COB module to the hole in this process is called 'pre-attaching'. The pre-attached inlay sheet is laid with its RF interface, i.e. antenna electrodes, upward. A worker can see the antenna connection parts and bond it directly by welding or soldering. In the conventional process, a worker is not able to see them and perform direct bonding because the antenna contacts are covered by the COB module. Thus, an anisotropic conductive adhesive has been inevitably used. Hot melt tape is melt by thermal pressing and then strong adhesion is accomplished in a final lamination process as shown in Fig. 10.

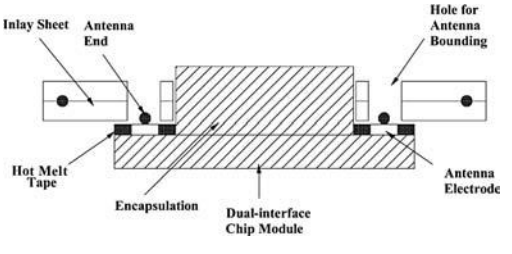


Fig. 10. Pre-attaching COB module to punched hole of inlay sheet with hot melt tape

Fig.11 and Fig.12 show direct bonding between the electrodes of COB module and antenna ends. The bonding method used is high current welding. The enamel coated on the antenna coil is exploded like a chromosome when high current is applied. Fortified electrical connection can be accomplished by this direct bonding, in comparison with the conventional method using conductive adhesive.

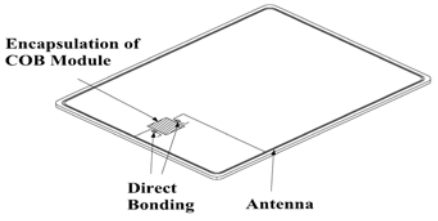


Fig. 11. Direct bonding between ⊏ shaped antenna ends and electrodes of COB module

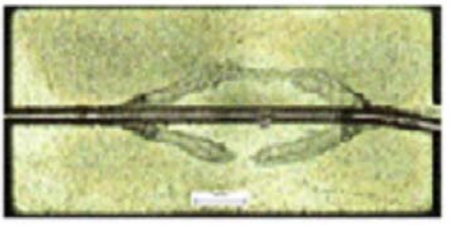


Fig. 12. Microscopic view of bonded antenna

3.4 Applying Fillers and Laminating All Sheets Including Core Sheets

After pre-attaching and bonding process, fillers are applied to gaps formed at the top of the COB module and the antenna connection parts. Various fillers can be used in this process, e.g. a UV curing agent, an instant adhesive, a heat curing agent, or the like. The UV curing agent is stiffened when exposed in UV lights and the heat-curing agent is hardened when the sheets are laminated with heat. With the fillers applied, the inlay sheet is sandwiched between printed sheets. Overlay sheets are also stacked to protect the printing. More core sheets can be stacked optionally on the inlay sheet to compensate the height of the card body. Magnetic stripe can be written on the backside of overlay foil on demand. The stacked sheets are then pressed and laminated with heat at 120-150 degree under the force of 15 kgf for at least 0.5 hour. Each of card bodies is punched and cut out of the laminated sheets. Personalization is applied to the individual card body.

An antenna wire for RF data communication is embedded inside of the card body with a contact electrode. Fabricated Smart combi card(SCC) is shown in Fig. 15.[4-5]

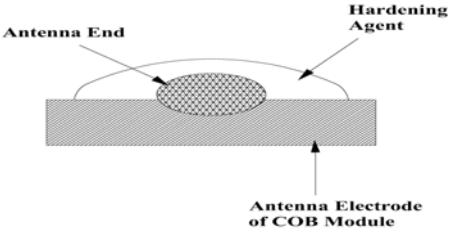


Fig. 13. Cross sectional view of bonded antenna end with hardening agent

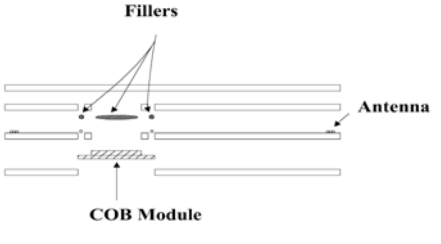


Fig. 14. Applying fillers to antenna connection holes and molding part of IC chip

Fig. 15. Smart combi card(SCC)

3.5 Advantages of Proposed Process

(1) Direct bonding between the antenna ends and the electrodes of the COB module gives highly reliable electrical conductivity and physical connection.
(2) COB module is strongly attached to the card body because hot melt tape is sufficiently heated to be molten and stiffened.
(3) Milling process to form a cavity for the COB module is not necessary any more thanks to using punched sheets. Punching a hole is much easier than milling a cavity.
(4) Humidity cannot exist in a card body because gaps between the antenna connection parts and between a COB module and a card body are filled with fillers.
(5) Backside of a card body, especially a position of a COB module, has good smoothness and appearance.

4 Test Procedures and Results

4.1 Bending and Torsion Test

Physical properties of a smart card are defined in ISO 7816-1. In this test, the test conditions were exceeding those of the ISO standards. Commercially available five smart cards manufactured by the conventional process using conductive adhesive and other five cards prepared by the proposed process were employed to perform bending and torsion test. As shown in Fig. 16, each of the cards was inserted with the contact facing upward and with its long side into the bending device. With this movement the maximum deflection of each card was 2 cm. The card was bent at a rate of 30 bendings per minute. This occurred 250 times. Afterwards, each of the cards was turned with the contact downward and bent 250 times. Subsequently, the card was inserted with its short side into the bending device with the contacts upward and bent again 250 times. Here for the maximum deflection is 1 cm. Next, the card was turned again and bent another 250 times. Lastly, the card was clamped at the short side and twisted around 151with 250 times of twists at a rate of 30 twists per minute as shown in Fig. 17. Thus, altogether a bending and torsion cycle of 1,000 bends and 250 twists was completed and repeated all the above steps.

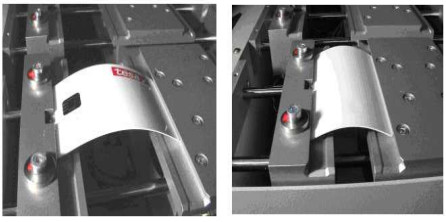


Fig. 16. Bending tests

Fig. 17. Torsion test

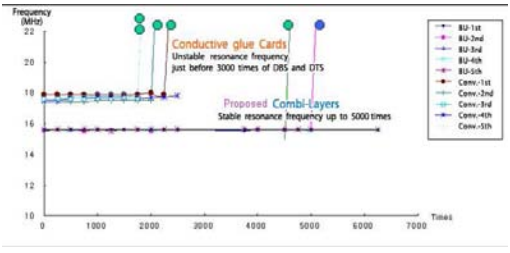


Fig. 18. Measured resonance frequencies

Between the changes after 250 bends or twists the card was tested for electrical functionality. That is, resonance frequencies for RF communication were measured per 250 times of bending and torsion for each card.[6]

As the tests went on, the resonance frequencies of the conventional cards became unstable and finally could not be measured up to 3,000 times. This means the connection between the antenna and the COB module was separated. The separated connection is shown in Fig. 19. The cards manufactured by the proposed process held stable resonance frequencies over 5,000 times of stress.

4.2 Thermal Shock Test

The conventional cards and the subject-processed cards were put in the thermal shock chamber. The temperature of the chamber was adjusted at -15 for 20 minutes and at 50 for 20 minutes respectively. Distances within which the card

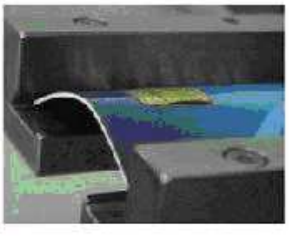


Fig. 19. Disconnection between COB and Antenna (after 2,500 times stressed)

		Conventional Cards		Card prepared by proposed method	
		Before	After	Before	After
At −15℃ for 20 min	Distance (mm)	28±2	Failed	42±2	42±2
	Frequency (MHz)	17.8±0.1	52	15.1±0.1	15.1±0.1
At 50℃ for 20 min	Distance (mm)	Failed	Failed	42±2	42±2
	Frequency (MHz)	Failed	Failed	15.1±0.1	15.1±0.1

Fig. 20. Results of the thermal shock test

reader can read the data of the IC card and resonance frequencies for RF communication were measured. The measured distances and frequencies are shown in below fig. 20. The conventional cards lost their RF functionality after left at -15 for 20 minutes.

The cards prepared by the proposed method had no changes by thermal shock after being left at -15 for 20 minutes and at 50 for further 20 minutes. The conventional cards have a little humidity penetrated into gaps between a COB module and a card body or between antenna connection parts. However, in case of the proposed cards, there are no gaps like those of the conventional cards owing to applying fillers, which result in good electrical reliability under thermal changes.

5 Conclusions

We have discussed the conventional process and the proposed process for Smart Combi Cards(SCCs). As described above, the SCCs should have electrical reliability against physical or mechanical stresses to perform multi-applications. The proposed process can provide SCCs that can avoid gradual malfunction occurred in a conventional SCCs. Therefore it can be applicable to the useful Smart Combi Card (SCC) for Radio Frequency Identification(RFID)/ Ubiquitous Sensor Network(USN) environments.

References

1. Y. Haghiri, et. al., Smart Card Manufacturing, Wiley and Sons, 2002.
2. Seungho Tak, Let's Smart Card, Sungandang, 2004.
3. Card Technology, Feb. 2004, p23.
4. Hongjun Yoo, et. al., Method of Manufacturing a Dual-interface IC cards by Laminating a Plurality of Foils, Korean Patent No. 515,001.
5. Hongjun Yoo, et. al., Method of Manufacturing a Dual-interface IC cards by Laminating a Plurality of Foils, Korean Patent Laid Open No. 2004-49981 and 2004-65402
6. http://www.jtcorp.co.kr.
7. http://www.kdnsmartec.co.kr
8. http://www.esmartec.co.kr

Product Control System Using RFID Tag Information and Data Mining

Cheonshik Kim[1], San-Yep Nam[2], Duk-Je Park[3], Injung Park[3], and Taek-Young Hyun[4]

[1] Digital Media Engineering, Anyang University
mipsan@anyang.ac.kr
[2] Dept. of Information and Communicstion, Kook Je College
R1337@unitel.co.kr
[3] Dept. of Electronic Eng. Dankook University in Korea
fly21c@airport.co.kr,digitallab@kornet.net
[4] UM technology
cto@umtech.co.kr

Abstract. In this paper, we suggest a method that applies RFID tag information and a data mining technology to a manufacturing execution system (MES) for efficient process control. The MES is an efficient process control method for many enterprises. But, the MES is not an analysis technique for process control. Therefore, we will supplement a data mining technology and RFID tag information to generate a more efficient process control system. In order to accomplish this, we designed and implemented an efficient product control system and adapted it to a TFT LCD production line using RFID tag information and data mining. As a result, the method proposed solved defects in parts and problems of personnel expenses.

Keywords: RFID, Data Mining, MES, LCD line.

1 Introduction

Information systems in the modern process and manufacturing sector bridge several layers, from the boardroom to the shop floor. At one end of this spectrum the ubiquitous ERP systems exist that have become essential to today's IT-enabled enterprises. At the opposite end of the spectrum, sensors, actuators and other field devices are found that are equally vital for ensuring perfect process control. Between these extremes a diverse range of systems with varying degrees of interconnection, fineness and cohesion exist[1].

Information systems span a wide range of data, processing power and time scales. The objectives and abilities of each of these systems are different, yet there is a clear need for them to operate in sync with each other. Any disconnect among them leads to inefficient operations, higher costs and lower quality, which ultimately translates into lower profits. Therefore, it is vital that there should be tight integration and perfect communication between all systems.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 100–109, 2007.

A clear need exists for a set of systems that seamlessly bridge this gap. The manufacturing execution system (MES) fills this need. The MES controls the operations that enable realization of the plans, close the execution gap by providing links among shop floor instrumentation, control hardware, planning and control systems, process engineering, production execution, the sales force and customers.

The MES has multiple advantages. Nevertheless, there is no function regarding analysis techniques concerning the manufacturing process in the MES. Therefore, we designed and implemented the MES for the manufacturing process of TFT LCDs. Also, as reported in this manuscript, we investigated and analyzed the defects of LCDs that are produced in the manufacturing process using a data mining technology.

2 Manufacturing Execution System

A Manufacturing Execution System (MES) is a system that companies can use to measure and control production activities with the aim of increasing productivity and improving quality. The ISA has defined standards regarding the structuring of MES and its integration in a larger company-wide IT architecture. MES fits in between ERP and process automation level. MES gets production order and those are scheduled by ERP and that is also not in detail. Material requirement planning does the scheduling at the ERP level. MES collects the production order and does a detail scheduling for a small period.

The industry wants to know how to reduce the manufacturing cycle time, improve the quality, lower the cost and get more profit. Since MES can provide real time monitor and integrate with ERP and other information system, the potential utility will appear after the enterprise used MES. MES collected the manufacturing data, and managers can make the strategic decision by it and carry out the decision. It is information presented on-line to the production operator and to the desk of manufacturing management. Under MES implementation, integration with your accounting system, order entry system, inventory system, scheduling system and others become easy. These all become plug and play pieces because the data is sharable.

While the ERP focus is in areas such as finance, HR, etc., new investments focused on improving and optimizing production and logistic resources. An Advanced Planning and Scheduling (APS) system uses advanced programming techniques to improve/optimize production planning and scheduling, to allow the company to achieve pre-defined objectives, such as improvements on delivery performance without raising inventory levels, or maximize plant throughput [2]. Tao et al. [3] proposed an implementation process model for integrating the extensible Markup Language (XML) into an enterprise application, which also meets the inter-organizational data exchange standard of RosettaNet. Farahvash and Boucher [4] introduced an architecture that integrated shop floor agents for scheduling, cell control, transportation, and material management.

3 Design of MES for TFT LCD's Process Control

3.1 System Architecture Design

Product inspection is handled in real-time, because the RFID [5] concept is applied to the MES. If the LCD Panel or AD Board arrives in the examination line, the RFID tag will be printed. Information for parts is also provided because the information is linked with the ERP or SCM. However, because the product is selected from the database in this study, the RFID TAG LABEL is printed.

The system architecture (Fig. 1) was based on the MES component principle. Quality management process management, labor management, data collection and acquisition, dispatching production units and document control of the MES function were modeled.

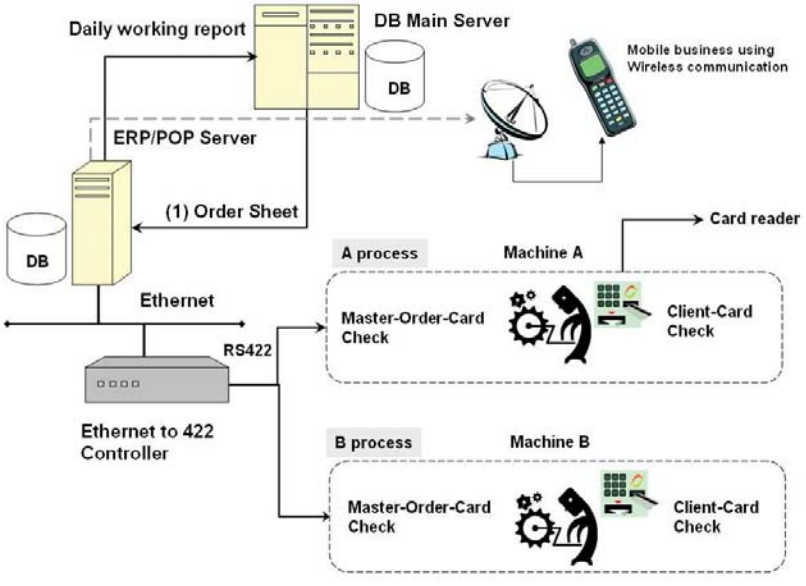


Fig. 1. Architecture of system for the parts

3.2 Data Flow Diagram of the System

The data flow diagram of the part inspection system using RFID in an electronic device manufacturing process, in which parts are produced or arrived at the subcontractors, the part information is inputted, the tag is attached by using an RFID printer, and it is moved to the inspection conveyor is depicted in Fig. 2. An inspector reads the tag information using a RFID reader, and outputs the information for the product concerned on the monitor. An LCD

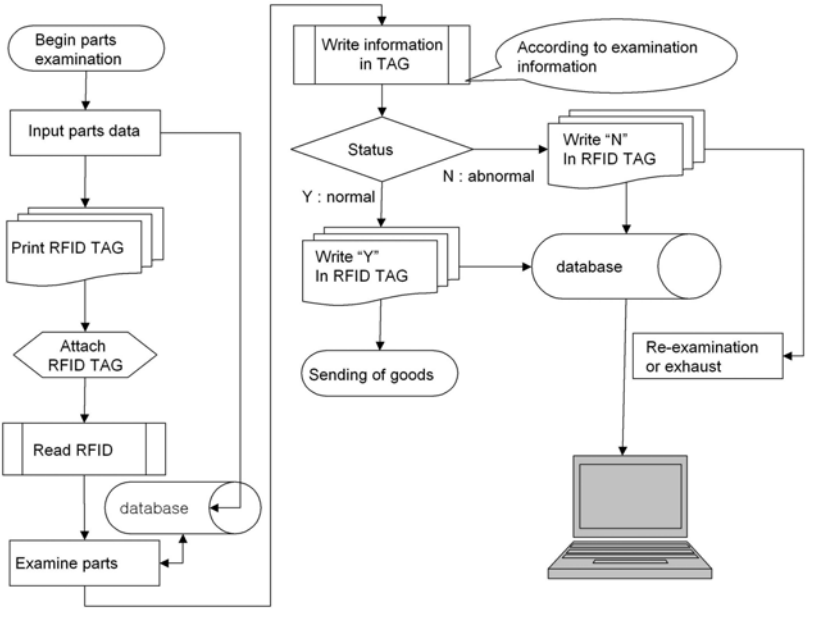


Fig. 2. Data flow diagram

electronic component is conveyed to an appropriate box or a pallet after judgment is made regarding whether to send the part to a shipment conveyor line, re-inspect it, or send it to disposition on the conveyor line due to inferior quality, when tag information is read through the final RFID reader. The system stores inspection data on various parts in the database, and permits them to shown on the monitor of a subcontractor, or an office manager, through the EDI/WEB. The system is configured to link with the existing SCM, and the ERP DB.

The diagram in Fig. 3 illustrates a detailed explanation of processes. When an electronic component arrives at the inspection line via the conveyor belt, the RFID reader automatically reads the appropriate information from the tag and judges whether or not it is a panel inspection or a board inspection.

Fig. 4 shows the main screen of the parts inspection system. The screen is on standby in the state depicted in Fig. 4. The left hand side of the screen configuration automatically indicates the details of the electronic component concerned. On the right hand side, the current date, time, inspector's name and ID, and inspection line are displayed. In the center of the display, which is the core of the program, the inspection items are automatically displayed in line with the electronic component concerned. Therefore, the appropriate inspection items can be easily understood by non-skilled personnel. The main screen of the parts checking system (Fig. 4) is implemented by the MES.

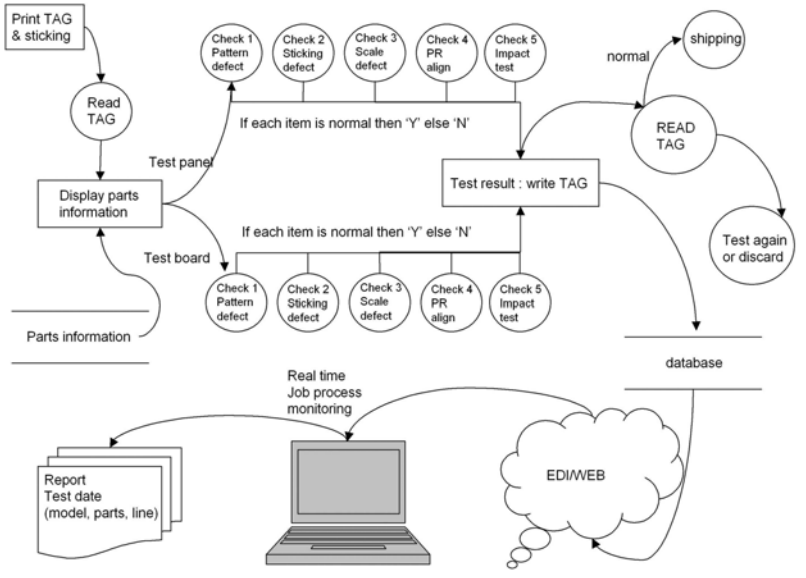


Fig. 3. Data flow diagram for the detail processing

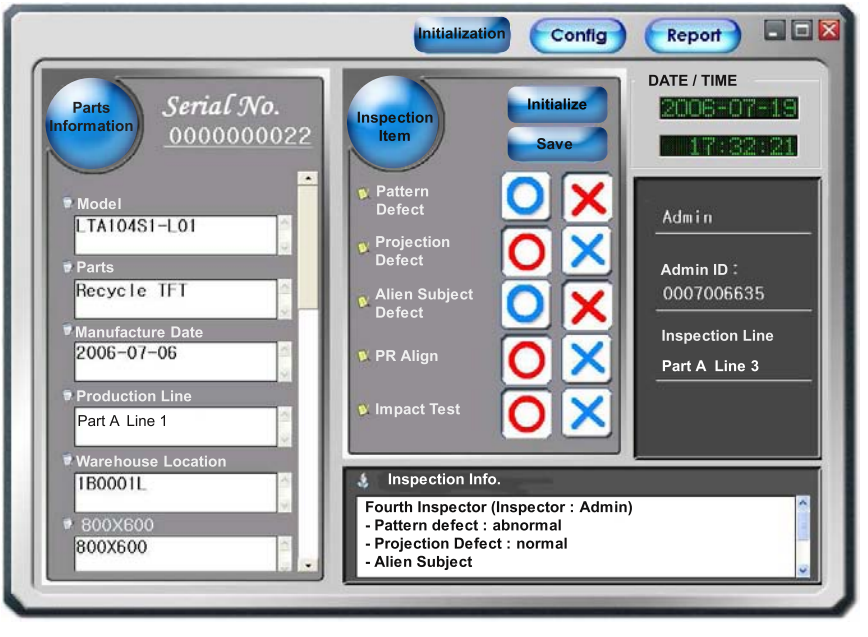


Fig. 4. Main screen for parts inspection

4 Data Mining for Efficient Production Control

In this chapter, we will explain neural networks and C 4.5 algorithms. Also, we apply these algorithms in the manufacturing process for TFT LCDs and suggest methods by which to locate defective parts. The method proposed may contribute to improve the efficiency of the manufacturing process for TFT LCDs. A neural network algorithm is a technology used in estimate modeling. Neural network algorithms are very efficient; however, this type of algorithm has shortcomings that can not explain the estimate sequence. Therefore, we used C4.5 algorithms and supplemented the shortcomings of the neural network algorithm. The C4.5 algorithm appropriately explains the sequence.

4.1 Neural Network Algorithm

Neural network[6] technology uses a multilayered approach that approximates complex mathematical functions to process data. It consists of many processing elements or nodes that work in parallel. Nodes are connected to each other in layers, and layers are interconnected. These nodes are simple mathematical functions; the connections between these nodes, which weight the data transformation from each node and send the information to the next node or output layer, are how neural networks "think." As the complexity of the task increases, the network size increases, and the number of nodes increases rapidly.

To properly train a neural network, the developer feeds the model a variety of real-life examples, called training sets. The data sets normally contain input data and output data. The neural network creates connections and learns patterns based on these input and output data sets. Each pattern creates a unique configuration of network structure with a unique set of connection strengths or weights. A neural network adapts to changing inputs and learns trends from data. A set of examples of the data or images is presented to the neural network, which then weights the connections between nodes based on each training example. Each connection weight builds on previous decision nodes, propagating down to a final decision (Equ. 1).

- Signals passed from each input node are gathered and become a linear combination. That is, the hidden node, L, is expressed as follows if (x1, ..., xp) is the explanatory variable.
- Connection weights for each input are summed, resulting in a unique complex function each time the neural network is trained with a set of inputs and outputs. Successsively summed weights define the algorithm that the neural network uses to make a pattern-matching decision.

$$L = w_1 X_1 + ... + w_p X_p \tag{1}$$

After the neural network reaches a final decision, it compares its answer against an answer provided in the training set. If there is a match, within a

predefined tolerance, the neural network stores these connection weights as successful. If the decision outcome is outside the tolerance, then the neural network cycles through the training set again. A neural network may cycle thousands of times to reach an acceptable tolerance.

Table 1 is used to determine the variable used to forecast the defect in TFT LCDs

Table 1. Factor for TFT LCDs quality decision

Input variable	Meaning of variable	Category
Variable 1	Pattern defect	Y/N
Variable 2	Output voltage	Y/N
Variable 3	Decide projection	Y/N
Variable 4	Terminal R	Y/N
Variable 5	Terminal Y	Y/N
Variable 6	PR Align	Y/N
Variable 7	Terminal W	Y/N
Variable 8	Impact test	YES/NO

4.2 C4.5 Algorithm

Statisticians developed a tree-structured classification of many members known as machine learning. Characteristics of the tree model are described as follows.

```
If A, then B, Else C
```

The decision tree C4.5 algorithm [7] used an entropy standard. Entropy is a concept used to measure randomness in thermodynamics. Suppose that data set, T, depends on Y and is divided into k. Then, the ratio (p1,, pk) of the category can be classified. Therefore, the entropy of T is defined in equation 2.

$$Entrop(T) = \sum_{i=1}^{k} p_i log\ p_i \tag{2}$$

The C4.5 model must find a separation variance that generates the lowest entropy in the entropy test. In this paper, we will determine factors concerning the defects of LCDs using C4.5 algorithms. The factors in Table 1 were used in an experimental design. The sequence of results is illustrated in Fig. 5.

4.3 Analysis of the Proposed Algorithm

Table 2 and Table 3 are results from the manufacturing process that apply to the neural network algorithm and the C4.5 algorithm, respectively. Table 2 and Table 3 present data on precision and recall.

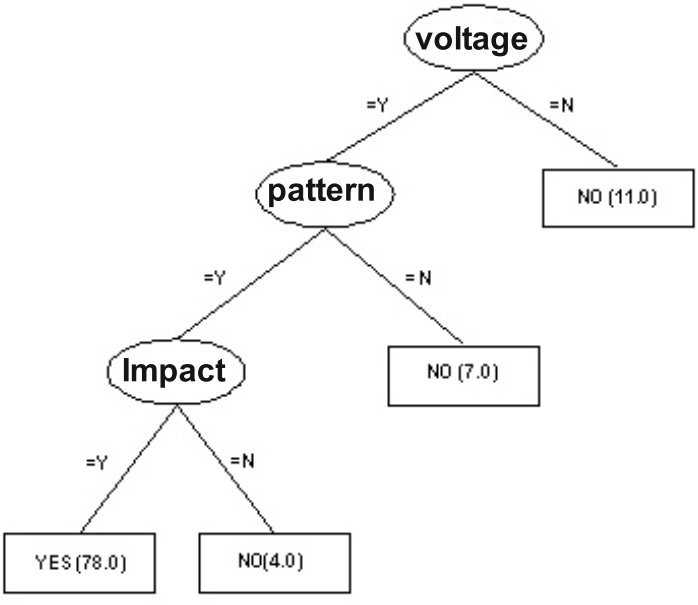


Fig. 5. Application of C4.5 algorithm to LCD line

Table 2. Experiment result of neural network algorithm

Precision	Recall	F-Measure	ROC Area	Class
0.98	0.987	0.985	0.946	YES
0.943	0.943	0.971	0.946	NO

Table 3. Experiment result of neural network algorithm

Precision	Recall	F-Measure	ROC Area	Class
0.97	0.985	0.977	0.983	YES
0.971	0.943	0.957	0.983	NO

The results of analysis are as follows.

- 11 items had a voltage defect in the whole parts production number.
- 7 items had both a pattern defect and a voltage defect in the whole parts production number.
- 4 items simultaneously had a pattern defect, a voltage defect and an impact defect in the whole parts production number.

Therefore, the incidence of voltage defects should be reduced. Also, we determined that a voltage defect causes a pattern defect. Finally, voltage defects in the manufacturing process should be managed specifically. We used RFID

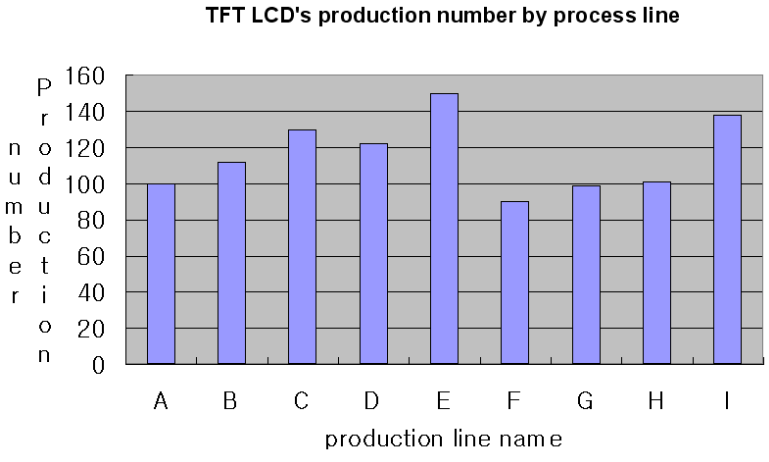


Fig. 6. Statistical analysis of the parts using RFID tag

technology and demonstrated that the process of producing TFT LCDs can generate monitoring in real time. Also, we produced real-time data information by chart to more easily confirm manufacturing information. Fig. 6 illustrates productivity according to each line of TFT LCD production.

5 Conclusions

In this paper, the system proposed was established to actualize an overall inspection system for electronic components or devices using RFID. This study developed a system to inspect electronic components and devices, specifically, LCDs Panels that are the core parts of an LCD monitor store inspection result data in the RFID TAG and the Reader/Writer. The existing system managed the inspection result data by manually attaching stickers containing inspection values. As a result of implementing the system developed in this research, the inspection time in the real parts inspection line was greatly reduced. The system developed consists in a way that the inspection data of multiple types of parts or devices can be displayed in real time to raise the efficiency of the concerned inspector or manager. This system is comprised of a MS-SQL SERVER or MySQL, which is a general purpose database, and can be linked with various ERPs and SCM. This system is forecast to provide many benefits to LCD panel and parts inspection companies. The MES is an excellent system for process control. However, the MES lacks an analysis function concerning information that occurs in process control. As a result, this approach was effective for closely examining the cause of necessary product defects in process control. We expect that the system proposed in this paper will be useful in various fields.

References

1. Manufacturing Execution System - A Concept Node, TATA consultancy services, 2002.
2. Integration of MES with Planning and Scheduling Solutions. Broner Metals Solutions Ltd - Watford. Watford, UK (2004)
3. Yu-Hui Tao, Tzung-Pei Hong, Sheng- sun. An XML implementation process model for enterprise applications *Computer in Industry* 55, 2004.
4. Pooya Farahvash, Thomas O. Boucher. A multi-agent architecture for control of AGV system. robotics and computer-Integrated manufacturing 20,2004.
5. Juels, R. Rivest, M. Szydlo. The Blocker Tag: Selective Blocking of RFID TAG for Consumer Privacy". 10th ACM CCS, 2003.
6. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.1993.
7. Abdi, H. A neural network primer. *Journal of Biological Systems*, 2, 247-281, (1994)

iSCSI Protocol Parameter Optimization for Mobile Appliance Remote Storage System at Smart Home Environment Approach

Shaikh Muhammad Allayear[1], Sung Soon Park[1], and Cheonshik Kim[2]

[1] Dept. of Computer Science and Engineering, Anyang University
allayear@anyang.ac.kr,sspark@aycc.anyang.ac.kr
[2] Major in Digital Media Engineering, Anyang University
mipsan@anyang.ac.kr

Abstract. In Mobile appliance, users have a limited amount of storage availability to them due to their limited size and weight. To relieve problem we developed iSCSI remote storage system, which is an excellent solution for smart home automation too. User can store or access their valuable data to the home server from anywhere, anytime and also get facility to use mass storage space. The iSCSI protocol has emerged as a transport for carrying SCSI block-level access protocol over the ubiquitous TCP protocol. It enables a client's block-level access to the data on remote storage over an existing IP infrastructure. However, the performance of the iSCSI based remote storage system for mobile appli-ances were sharply dropped in wireless networks; especially when we adapt default parameters value suggested in standard for our remote storage system in wireless networks. This paper focuses our experiments, which are performed to investigate the best performance values of iSCSI parameters for iSCSI-based remote storage system, are taken out in CDMA networks in order to realize the access to a remote storage system anytime and anywhere. And after the experiment, we suggest the optimal value of parameters. The experiment results from several test cases show us the best values are not the default values specified in the iSCSI standard.

Keywords: Mobile, SCSI, SAN, CDMA.

1 Introduction

Mobile appliances are going to be used in more area as the time goes by. Smart Home network is one of important platform. Due to huge development in mobile appliance area, people want to automate their home by mobile appliance also. In this paper we focus our experiments investigate performance values of iSCSI parameters of our developed iSCSI protocol based Remote Storage System for mobile appliance [1], for smart home automation. User can access their home server from anywhere, anytime to store their data's and access them. But many efforts are performed to apply traditional wired network environment services such as multimedia and database to mobile appli-ances in wireless network environment. As the amount of contents is increased for the services, the need

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 110–119, 2007.

for storage expansion is increased more and more. Due to their mobility, mobile appliances should be small and they use a flash memory to store the data. However, it is still difficult to store multimedia data such as mpg, mp3, etc and install large software such as database engines [2][3][4]. Moreover, mobile appliances are more vulnerable and fragile than stationary devices, because they can be easily stolen, lost or damaged [1]. It is need to keep their data in a secure and safe space. To alleviate these problems, we developed Remote Storage System for Mobile Ap-pliance [1], a iSCSI based remote storage system, for providing the allocation of a mass storage space to each mobile client through networks. The system offers to its users the possibility of keeping large size of multimedia data and database in a secure and safe space. A number of papers have discussed the performance of iSCSI based remote storage system for mobile appliances. In [2][3][4], they proposed cache server to improve performance in wireless environment. However, the paper does not describe the adap-tation of iSCSI protocol for wireless network. This paper we study iSCSI parameters defined in the standard [5] to investigate the effect and suggests best optimal value of parameters into CDMA network. In Section 2, we introduce iSCSI protocol and parameters. In section 3 we analysis of iSCSI parameters in Wireless environment. Section 4 describes iSCSI Remote Storage Sys-tem architectures. In section 5, we describe our experiment setup procedure, experi-ments methodology, experiments results of parameter optimization, here our experi-ments show the results of adapting the various settings of the iSCSI parameters for our iSCSI based remote storage system in CDMA networks and suggest the best perform-ance sets of parameter values. Finally we conclude this paper in section 6.

2 Backgrounds

2.1 iSCSI Protocol

The iSCSI protocol [5][6] is a mapping of the SCSI remote invocation procedure model [7] onto the TCP/IP protocol suite. The iSCSI specifications have recently been approved as an Internet Engineering Task Force (IETF) standard [5]. The strength of iSCSI stems from the fact that it builds on well-established technologies like SCSI, TCP/IP and Ethernet. Native SCSI, as used in direct attached storage, is both a proto-col and a physical transport. iSCSI is a session-based protocol in which the transport is provided by TCP/IP. iSCSI uses a client-server architecture in which the client is called the iSCSI initiator and the server is called the iSCSI target. The iSCSI protocol architecture model is as shown in figure 1. iSCSI establishes a communication session between the initiator and target. The session may consist of one or more TCP connec-tions. SCSI Command Descriptor Blocks (CDB) are passed from the SCSI layer to the iSCSI transport layer. The iSCSI transport layer encapsulates the SCSI CDB in an iSCSI Protocol Data Unit (PDU) and forwards it to TCP. On a receive, the iSCSI transport layer extracts the CDB from the iSCSI PDU, received from the TCP layer and forwards the CDB to the SCSI layer.

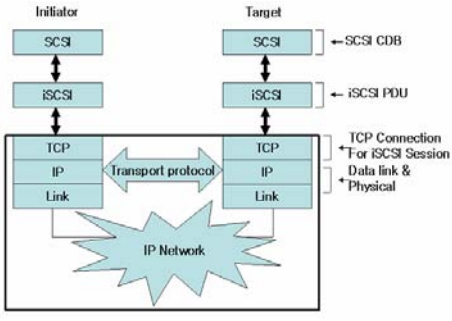


Fig. 1. iSCSI (Internet Small Computer System Interface) Protocol Architecture

2.2 iSCSI Parameters

The values of iSCSI parameters can be determined during login phase and full feature phase. Each iSCSI connection begins with login phase. The initiator and target can negotiate iSCSI parameters to improve performance during login phase. After that, iSCSI enters the full feature phase, during which iSCSI commands and data are ex-changed over the established iSCSI connections. There are two classes of iSCSI pa-rameters. One of them is related to iSCSI read operation while the other is related to iSCSI write operation.

iSCSI read operation parameters. There are three parameters, which are related to an iSCSI read operation.Number of sectors per command. Most SCSI disks define a sector size of 512 bytes, and require I/O operations to be in multiples of a sector. The iSCSI initiator is usually required to declare a limit on the number of sectors in a single SCSI I/O operation. It is the one of important parameters for iSCSI read operation, since the value of this parameter limit the request size of an iSCSI read command. The target can continu-ously send out the data requested from an iSCSI read command using Data-In PDUs. However, it is not directly related to iSCSI write operation. Though an iSCSI write command request the data transmission from initiator to target, R2T mechanism control the transmission size of the data associated with the command. MaxRecvDataSegmentLength (MRDSL) of initiator. All iSCSI PDUs have one or more header segments and, optionally, a data segment. The Basic Header Segment (BHS) is the first segment in all of the iSCSI PDUs. The BHS is a fixed-length 48-byte header segment. Additional Header Segment (AHS), a Header-Digest, a Data-Segment, and a Data-Digest may follow it. The initiator declares the maximum data segment length in an iSCSI PDU. Thus, the DataSegmentLength of a Data-In PDU must not exceed MaxRecvDataSegmentLength of the initiator. Phase Collapse. The normally finishes a read command by sending a separate iSCSI response PDU containing the command's status. However, when the status is success, a non-negotiable option allows the target to use a phase collapse in which it sets a bit in the final Data-In PDU and omits the iSCSI response PDU.

iSCSI write operation parameters. There are two parameters, which are related to an iSCSI write operation MaxBurstLength. The initiator and target negotiate maximum SCSI data payload in bytes in a solicited Data-Out iSCSI sequence. A sequence consists of one or more consecutive Data-Out PDUs that end with a Data-Out PDU with the F bit set to one. The end of a sequence of Data-Out PDUs requires the transmission of a R2T PDU by the target back to the initiator before the next data-out sequence can begin. Therefore, the value of the MaxBurstLength parameter limit the total amount of all data segments of all PDUs in a solicited data sequence requested by a R2T PDU. MaxRecv-DataSegmentLength (MRDSL) of target. The target declares the maxi-mum data segment length in an iSCSI PDU. Thus, the DataSegmentLength of a Data-Out PDU must not exceed MaxRecvDataSegmentLength of the target.

3 Motivation

Analysis of iSCSI Parameters in Wireless Connection. The iSCSI Data-In PDUs are passed to the TCP layer, since the iSCSI PDU is a SCSI transport protocol over TCP/IP. When the iSCSI Data-In PDU size is greater than the MSS (maximum segment size) of TCP layer, the PDU will be further fragmented into smaller segments. The iSCSI Data-In PDU is generally larger than the MSS in size. If some segments of parts of an iSCSI Data-In PDU are lost due to the high bit error rate in wireless networks, TCP layer would require re-transmitting the segments. At that time all the other segments of those parts of an iSCSI Data-In PDU must wait for being reassembled into an iSCSI Data-In PDU in TCP buffer. It decreases the per-formance of the system due to the reducing of the available capacity of TCP buffer. The sender can transmit data segments, which are allowed by the receiver using TCP flow control mechanism. In a wireless network with high bit error rate, the more size of iSCSI PDU is increased by the MRDSL (MaxRecvDataSegmentLength) parameter, the more segments would have to wait due to the lost of some segments of parts of an iSCSI PDU in TCP buffer. The performance of the system thus decreases more. In write operation, it has the same results as the read operation in affecting the perform-ance caused by the parameter for target. In addition, a wireless link generally becomes a bottleneck portion in an end-to-end TCP connection because of its narrow band-width, as compared to wired links. Thus a TCP sender's congestion controls are apt to be caused by wireless link congestions. When congestion occurs in wireless link, there are two indications of packet loss, which are a timeout occurring and the receipt of duplicate ACKs. Though TCP can reduce the transmission amount of data segments using congestion avoidance mechanism, it still transmits data segments until detecting congestion in wireless network. If the amount of SCSI data payload which are con-tinuously passed to the TCP layer is increased by increasing the value of Max-BurstLength for write operation or Number of sectors per command for read operation, and the size of an iSCSI PDU is enlarged by increasing the value of MRDSL, the sender

will transmit more segments until recognizing congestion in wireless network. It causes the performance falling to the iSCSI based remote storage system for mobile applications.

4 iSCSI Remote Storage System for Smart Home

4.1 System Architecture

The system architecture of iSCSI Remote Storage system is shown in figure 2. The system is consists of 2 parts: one is the Client or Initiator for mobile appliance, and another is Server or target as storage server.

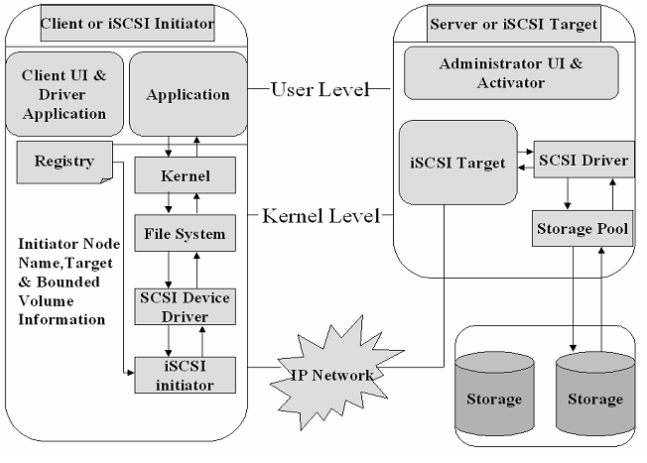


Fig. 2. iSCSI (Internet Small Computer System Interface) Protocol Architecture

4.2 Client for Remote Storage System

The User Interface (UI) for client can activate the iSCSI initiator and pass the in-formation of target, initiator and bounded volume via registry. iSCSI Initiator can create and manage the virtual volume to access the remote storage server with mobile appliance. (i.e. Windows CE based PDA) Also, it can receive the request of user File I/O from SCSI subsystem. After that, it transfer the command to SCSI based Storage Server (Linux based Storage Server) through TCP/IP connection

4.3 Server for Remote Storage System

The User Interface (UI) for administrator can activate the iSCSI target and pass the target name, volume size and name etc. The iSCSI target processes I/O operation according to iSCSI Initiator's requests.

5 Performance Evaluations

5.1 Experiments Environment Setup

Our experiment setup consists of a PDA and Storage server connected on to Internet with CDMA 2000 1xEVD0 network. The initiator module embedded in PDA based on Windows CE can transmit iSCSI command and data to the target of Storage server for I/O. We set the experimental environment as Table 1. The Storage server is connected with RAID subsystem, which has the storage capacity of 1TB (Terabyte) through FC switch. We use the CDMA 2000 1xEV-D0 module as the wireless interface of PDA. It is commonly used as a wireless module in commercial PDA products. PDA users can access and download packet data from networks at the maximum speed 2 Mbps in CDMA 2000 1xEV-D0. The maximum speed of uploading is 307.2 Kbps. LAN card, which has the speed of 1 Gbps is used as a network interface of Storage server. Mobile client's memory has the capacity of 256MB and is operated on OS, Intel PXA270.

Table 1. Experimental Environment information table between Target (Storage Swerver) and Initiator (PDA)

	Storage Server or Target	PDA or Initiators
OS	Linux 8.0 (kernel version 2.4.18)	Windows CE 4.0
CPU	PIII800Mhz	Intel PXA270
Memory	512MB	256MB
NIC	1Gbps LAN	CDMA 2000 1xEV-D0

5.2 Experimenting Methodology

We Consider throughput as performance metrics, which is the total number of applica-tion level bytes carried over an iSCSI connection divided by the total elapsed time taken by the application, as expressed in Bytes per millisecond (B/ms). We used the system times of PDA, which is based on WinCE in order to measure throughput. Total elapsed time was measured as the time interval from the initial time when the experi-ment program started generating the first byte of data to the time when the last byte of data was confirmed to have been sent (received). The elapsed time therefore includes all data transfers and all read and write commands as well as responses at all levels of the protocol stack. We perform two kinds of experiments. In the first experiment, we generate an I/O stream of 5 megabytes, and then measure throughput. Our request for a 5 megabytes I/O operation is passed through the file system and the SCSI subsystem to the iSCSI initiator as a number of SCSI commands. Then the initiator sends the commands and data to the target device. At that time, we adapt various settings of the iSCSI parameters to investigate the best parameter values for performance in wireless network. Each data point plotted in every graph of this paper was calculated as the average of 5 runs with identical parameter settings. We also perform the same experi-ment again with a different size of I/O stream,

which is a 100 megabytes I/O stream. We perform the second experiment in order to examine the effect of the characteristics of the high bit error rate within the commercial CDMA network and the variable bandwidth of the network. Our experiment program generates read I/O bursts continu-ously for 15 hours, and then measures throughput every 1000 seconds. The second experiment shows that the increase of the MaxRecvDataSegmentLength parameter value cannot always bring performance improvement to iSCSI-based remote storage system in unstable wireless network with high bit error rate and variable bandwidth.

5.3 Experiments Results of Parameter Optimization

As the analysis of section 3, if we increase of the value of MaxRecvDataSegmentLength (MRDSL), MaxBrustLength and Number of sectors per command will reduce extra PDUs transmission and extra processing overhead on each side, it causes the performance improvement to iSCSI bases remote storage system in a stable wired network. But in a wireless network with high bit error rate, the more size of iSCSI PDU is increased by the MRDSL parameter, the more segments would have to wait due to the lost of some segments of parts of an iSCSI PDU in TCP buffer. The per-formance of the system thus decreases more. In write operation, it has the same results as the read operation in affecting the performance caused by the parameter for target. But we may increase the performance and take the decision to setting the MRDSL value because our scheme avoids drastic reduction of transmission rate from TCP congestion control and decrease the high bit error rate in wireless environment. In this subsection, we show the results from the two kinds of experiments. We per-formed the first experiment in order to investigate the best performance values of iSCSI parameters for iSCSI-based remote storage system in CDMA network.

Figure 3 shows the result from the first experiment for 5 megabytes read operation in CDMA network. We first examined the effect of the Number of sectors per com-mand with the parameter MaxRecvDataSegmentLength (MRDSL) from 512bytes to 8Kbytes respectively. The Number of sectors per command is increased from 16 to 2048, which increases the expected command size from 8Kbytes to 1024Kbytes due to a 512 bytes sector size for our system. In section 3, we explained that the increase of the value of MRDSL, MaxBurstLength and Number of sectors per command will reduce extra PDUs transmission and extra processing overhead on each side. It causes the performance improvement to iSCSI-based remote storage system in a stable wired network. Figure 3 illustrates that the increase of the Number of sectors per command causes the performance improvement in the wireless network. It is very similar to the experi-ment results in a wired network. However, there are the performance falling at the Number of sectors per command value of 2048 with MRDSL of 1Kbytes and 4Kbytes. At the MRDSL values of 1Kbytes and 4Kbytes, there are 1In write operation, the Number of sectors per command is kept constant at 1024(512Kbytes) and the MaxBurstLength (MBL) varies from 512Bytes to 256Kbytes. We use the fixed Number of sectors per command, because the value of the parameter MaxBurstLength limits the total amount of all data segments

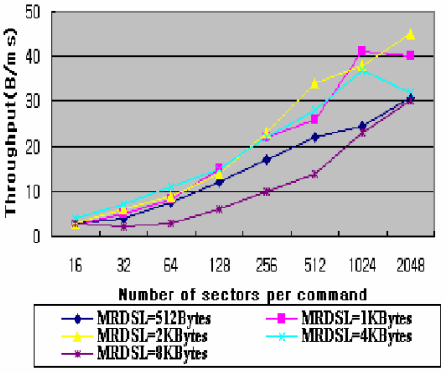


Fig. 3. Effect of Number of sectors per command on throughput for 5 Mbytes read operation in CDMA network

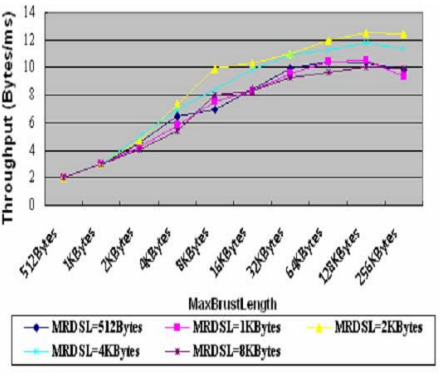


Fig. 4. Effect of Number of sectors per command on throughput for 5 Mbytes write operation in CDMA network

of all PDUs, which were requested by R2T as what we have already explained in section 2. Figure 4 shows that the throughput is increased as MBL increases from 512Bytes to 128Kbtyes in a wireless network, after that there is a slight decrease in throughput when MBL is around 256Kbytes. The same kinds of decrease in throughput happen in read operations too, which is caused by the narrow bandwidth of CDMA network as what we have already explained in section 3.2. The number of iSCSI PDUs which are continuously passed to the TCP layer can be increased by increasing either the value of MaxBurstLength for write operation or Number of sectors per command for read operation, then the sender will transmit more segments until recognizing congestion in wireless network, and thus causes the performance falling of iSCSI based remote storage system for mobile appliances. In a write operation, the result is more obvious because the CDMA 2000 1x EV-DO network has the narrower bandwidth for upload than that for download. Therefore it is not always true that the increase of MBL value would cause the performance improvement. The MBL value of 256Kbytes in standard is not suitable too. When the MRDSL value is 2Kbytes or 4Kbytes, the throughputs are better than that is 512bytes or 1Kbytes or 8Kbytes. At the MBL of 128Kbytes, the throughput of MRDSL of 2Kbytes are 26.

From these we can also see the standard value for wired network is not suitable here for a wireless network. Therefore, we suggest that you use the parameters settings of the MRDSL of 2Kbytes and 4Kbytes with the MBL of 128Kbytes when performing a small file write operation in CDMA network. Figure 5 shows the results from our experiment of 100 megabytes read operation. When MRDSL is at a value of 1Kbyte, the throughput is better than that at the value of 512bytes or 8Kbytes. When the MRDSL value is 1Kbytes, 2Kbytes and 4Kbytes, however, it is difficult to say which one is the best value. In the case of large size of I/O burst, such as 100 megabytes read operation, the best value for performance

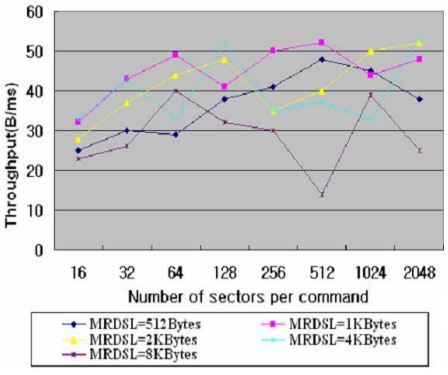


Fig. 5. Effect of Number of sectors per command on throughput for 100 Mbytes read operation in CDMA network

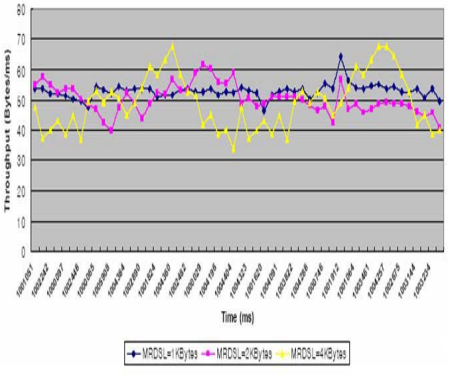


Fig. 6. Effect of MRDSL on throughput about 15 hours in wireless network

is also different from the standard value. When the MRDSL value is 2Kbytes, 4Kbytes and 8Kbytes, the performance is sharply dropped or increased with the increase of Number of sectors per command. At 8Kbytes of MRDSL value, there is a drastic decrease by 64Therefore we suggest that you should set the MRDSL parameter value at 1Kbytes when performing a large file I/O operation while moving in CDMA network.

6 Conclusions

In this paper we developed a remote storage system to solve the problem of storage capacity for mobile appliances and it is useful to use smart home environment. At anytime from anywhere we can access our home storage target server by our initiator mobile appliance. And then we studied and analyzed iSCSI parameters defined in the standard to investigate the effect of those. We performed two kinds of experiments to find out the best performance parameters setting in wireless network. Note that in several cases the best values are not the default values specified in the iSCSI standard. From the experiments, we suggest that you should use the parameters settings of the MRDSL of 1Kbytes, 2Kbytes and 4Kbytes with the Number of sectors per command values of 1024 or 2048 when performing a small file read operation in CDMA net-work. We also suggest that you should use the parameters settings of the MRDSL of 2Kbytes and 4Kbytes with the MBL of 128Kbytes when performing a small file write operation in CDMA network. In the case of large files operation, we found out that the smaller size of iSCSI PDU could be less affected by the characteristics of a high bit error rate within the wireless channel, and a narrow and variable bandwidth of the wireless channels.

References

1. Shaikh Muhammad Allayear, Sung Soon Park: "iSCSI Multi-Connection and Error Recovery Method for Remote Storage System in Mobile Appliance". The 2006 International Conference on Computing Science and Its Application, ICCSA-06, Glasgow, LNCS 3891, pp.641-650, May 2006.
2. D.Kim, M. Ok, M.-s. Park. An Intermediate "Target for Quick-Relay of Remote Stor-age to Mobile Devices". Proc. Of ICCSA, May (2005).
3. Sura Park, Bo-Suk Moon, Myong-Soon Park: Design, Implement and Performance Analysis of the Remote Storage System in Mobile Environment, Proc. ICITA 2004.
4. M. Ok, D. Kim, M.-s. Park, UbiqStor: A Remote Storage Service for Mobile Devices, The Fifth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 04) Singapore, December (2004).
5. J. Satran et al. Internet small computer systems interface (iSCSI). Technical Report RFC3720, Internet Engineering Task Force (IETF), April 2004.
6. Kalman Z. Meth, Julian Satran, Design of the iSCSI Protocol, IEEE/NASA MSST 2003, April (2003).
7. SAM-3 : Information Technology - SCSI Architecture Model 3, Working Draft, T10 Project 1561-D, Revision7, 9, May (2003).
8. V.Jacobson: Congestion avoidance and control, In SIGCOMM 88, August (1988).

A Study on Optimal Fast Handover Scheme in Fast Handover for Mobile IPv6 (FMIPv6) Networks

Byungjoo Park[1], Youn-Hee Han[2], and Haniph Latchman[1]

[1] Department of Electrical and Computer Engineering,
University of Florida, Gainesville, USA
{pbj0625, latchman}@ufl.edu
[2] School of Internet-Media, Korea University of Technology and Education,
Cheonan-Si, Chungnam, 330-708, Korea
yhhan@kut.ac.kr

Abstract. Mobile IPv6 (MIPv6) is protocol for handling routing of IPv6 packets to mobile nodes that have moved away from their home network. However, in MIPv6, the handover process reveals numerous problems manifested by a time-consuming network layer based movement detection and latency in configuring a new care-of address, with confirmation scheme called duplicate address detection (DAD). To reduce the handover latency in standard MIPv6, the fast handover for Mobile IPv6 (FMIPv6) protocol is proposed. To do this, each time a mobile node moves to a new location, it configures and confirms its temporal IP address during layer 2 handover. In this paper, we study the impact of the address configuration and confirmation procedure on the FMIPv6's IP handover latency. A mathematical analysis comparing the various parameters is provided to show the benefits of our scheme over the current procedure like standard MIPv6 and FMIPv6.

1 Introduction

To accommodate the increasing demand of mobility in the Internet, Mobile IPv6 (MIPv6) has been proposed in the IETF [1]. The protocol provides seamless connectivity to MNs when they move from one wireless point of attachment to another in a different subnet. According to the proposal, a mobile node (MN) should generate a new care-of address (CoA) by using IPv6 stateless address auto-configuration whenever it moves to new link. To verify the uniqueness of this CoA, it should run duplicate address detection (DAD) algorithm [2] before assigning the address to its interface. The algorithm determines if the address chosen by an MN is already in use. MN must perform DAD every time it handovers between IPv6 networks and cannot begin communication until DAD completes. According to the current RFC 2462 DAD algorithm, it takes at least 1000ms to detect that there is no duplicate address in the link. After finishing DAD procedure, the MN has to wait for random delays for router solicitation message (RS) and router advertisement message (RA) [2,3].

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 120–129, 2007.

Fast handovers for Mobile IPv6 (FMIPv6) [4], has been proposed in IETF to reduce the handover latency in standard MIPv6. This proposal describes a protocol to replace such new care-of address (NCoA) configuration procedure. It enables MN to quickly detect that it is now moving to a new subnet by providing the new access point (NAP) identifier and receiving the associated subnet prefix information. MN formulates a prospective NCoA, if at all possible, when still present on current subnet. Furthermore, in order to make MN allocates NCoA to its interface immediately after attaching to new subnet, FMIPv6 allows the NCoA confirmation procedure to be executed before or while MN switches its subnet. The scenario in which an MN receives the positive result about the confirmation of its prospective NCoA on the current subnet is called predictive mode. The scenario in which an MN checks the uniqueness of NCoA after MN attaches to a new subnet is called reactive mode. Although MN initiates the NCoA confirmation at an early time on the current subnet, FMIPv6 would fall into reactive mode if MN could not receive the confirmation result on the current subnet. In addition, if the proposed NCoA is rejected during the NCoA confirmation procedure, MN may configure NCoA by itself after movement so that handover latency becomes long. In order to achieve more reduction of handover latency, it is required that predictive mode should occur more frequently than reactive mode. So, it is necessary that the NCoA confirmation should be done promptly and its result should be always successful. However, a proper confirmation method has not been provided.

In this paper, we propose new movement detection, address configuration and confirmation scheme (*NAC*) in FMIPv6 networks that remarkably takes off the DAD procedure from the whole layer-3 handover procedure, thereby reducing layer-3 handover latency.

The reminder of this paper is organized as follows. Section 2 introduces our proposed FMIPv6 with *NAC* scheme. The performance evaluations and comparisons in MIPv6, FMIPv6 and proposed FMIPv6 with *NAC* scheme are shown in section 3. Finally we present the conclusion in section 4.

2 New Address Configuration and Confirmation Algorithm in Fast Handover for Mobile IPv6 (*NAC*)

In this section, we describe our new optimized address configuration and confirmation scheme called *"NAC"* to reduce total handover latency. We can define the handover procedure like movement detection, NCoA configuration and confirmation (DAD) procedure.

2.1 New Fast Movement Detection in FMIPv6

In FMIPv6, the movement detection is based on an indication from a wireless Layer 2 (L2) trigger which informs that MN will soon be handover. First, we assume that the L2 trigger signaling message from NAP includes stored router advertisement (SRA) message based on EAP [5,6]. To begin a fast handover, an

MN sends the router solicitation for proxy (RtSolPr) message to the previous access router (PAR).

The RtSolPr contains the L2 identifier of a target AP which the MN will move to. At this time, PAR starts to map the L2 identifier into proper target new access router (NAR). In response, The MN will receive the proxy router advertisement (PrRtAdv) message from PAR. Based on SRA and PrRtAdv messages, the MN compares the prefix of the SRA message with existing prefixes in the cache. If the prefix is different, the MN starts to configure a prospective NCoA using the IPv6 stateless (or stateful) address auto-configuration method. And then, the MN immediately sends the fast binding update (FBU) message with the prospective NCoA. When PAR receives FBU message, it sends the modified new handover initiation (*NHI*) carrying a newly define 1-bit D-flag, named "*NCoA DAD Request bit (D bit)*" to NAR, which validate the MN's new CoA and initiates the process of establishing a bidirectional tunnel between the PAR and the MN at its NCoA. This *NHI* message contains the "*previous MN's CoA*" and "*previous AR's global address*" to support interoperability with normal nodes by using a bit in the reserved field.

2.2 New NCoA Configuration and DAD Scheme in FMIPv6

To reduce DAD processing delay, we propose new NCoA configuration and DAD scheme using modified neighbor cache in NAR. This modified neighbor cache supports new enhanced lookup algorithm which can reduce DAD processing delay from 1000 ms to 5.28 μ sec. That is, the DAD using lookup algorithm consumes an extremely short amount of time, typically a few micro second units, such as Longest Prefix Matching speeds in routing table.

In the current FMIPv6, there is no specific address confirmation scheme. So, in our paper, we assume that RFC 2462 DAD is also used for the confirmation scheme. If the period of address confirmation procedure is long, then the delivery of handover acknowledgement (HAck) message and fast binding acknowledgement (FBAck) message would be delayed. That is, the MN can not receive FBAck before it disconnects with PAR. This means FMIPv6 could fall into the reactive mode and the MN has to resend FBU message as soon as it attaches to NAR. As a result, at this case, it requires to deliver an additional FBU message, which will be encapsulated in fast neighbor advertisement (FNA), with the consumption of wireless bandwidth. On the other hand, if NAR receives FNA and an encapsulated FBU, and detects that NCoA is duplicated, it must discard the inner FBU and notify this fact of MN. It will cause handover latency to be extended. This kind of case can occur even when the period of confirmation procedure is very short. If the result of confirmation shows that the prospective NCoA is invalid, the MN should itself configure its NCoA and run RFC 2462 DAD after moving to NAR.

However, if the NAR adopts proposed *NAC* scheme, these problems can be obviously removed. After receiving *NHI* message, NAR starts new DAD procedure using a lookup algorithm in modified neighbor cache in NAR. As soon as DAD procedure finishes, the NAR can unicast the HAck message to the PAR.

This HAck could also be modified like the NHI message by adding a 2-bit F-flag to the reserved flag and containing the "*New MAC address*", "*New link-local address*" and "*NCoA DAD Reply option*" in the option field in case of address duplication. We name this new HAck message as '*NHA*'. Table 1 defines the F-flag in proposed FMIPv6.

Table 1. The F-Flag of NHA message

F-Flag	Mean
00	Must change MAC address. (Can not apply in IEEE-802 case)
01	Can allocate a link-local address and a new CoA
10	Must change the link-local address allocated into the Alternative Address
10	Can not use

After the PAR receives *NHA* message from NAR, it sends FBAck message carrying the *NHA* message's "*NCoA DAD Reply option*" to MN. As soon as the MN receives the FBAck, it configures the address specified in the NCoA DAD Reply option into its interface. In the proposal, it takes a very short time to configure and confirm NCoA such as 5.28 μ sec in worst case. Also, in the proposed scheme, NCoA does not become invalid, since the unique NCoA is provided by NAR.

2.3 Lookup Procedure for Fast DAD Procedure in FMIPv6

We denote t_{LU} as the address lookup delay, which is the time required to check an MN's MAC address for movement detection and DAD in the Patricia Trie search. Accordingly the address lookup delay (t_{LU}) is given as:

$$t_{LU} = t_{DAC} \cdot N \tag{1}$$

Where t_{DAC} is the delay for access and comparison operations in RAM and N is the number of lookups in Patricia Trie. This Patricia Trie has the worst performance in line per minute (LPM). We use this algorithm in order to show the lookup time of the worst performance. Under the present circumstance, since a memory access requires from 60 to 100 nsec [7] and a comparison requires 10nsec in DRAM [8], we can use the value of t_{DAC} as 70 and 110nsec. In the Patricia Trie case, since lookups require accessing memory 48 times in the worst case, the N value is 48. Hence, t_{LU} is 3.36 μ sec and 5.28 μ sec and the calculated lookup delay is very small.

We describe the analysis method by using the queuing system. We assume that arrival packets are stored in the buffer and processed by the FIFO policy. We also suppose that the packet interarrival times can be modeled by a poisson process. Then, we use an M/G/1 queuing model to calculate the average performance of the MAC address in lookup algorithm. We denote λ_p as the *NHI* packet arrival rate at the AR. An average of lookup processing time ($E[t_{LU}]$) is determined

according to the corresponding neighbor cache lookup delays and the probability density of addresses determined by the memory access times. We define the traffic intensity φ :

$$\varphi = \lambda_p \cdot E[t_{LU}] \tag{2}$$

The traffic intensity φ is the quantity that governs the stability of the system. Let us introduce *LU* as the lookup delay, which is defined as the time duration from when an *NHI* packet arrives at the AR to when an *NHA* message is forwarded to the output link. By applying the M/G/1 queuing model, the mean lookup processing delay is derived by

$$E[LU] = E[W] \cdot E[t_{LU}] \tag{3}$$

Where $E[W]$ is the expected mean waiting time of a packet in queue. Using the Pollaczek-Khinchin (P-K) formula, the mean waiting time is derived by

$$E[W] = \left(\frac{\lambda_p \cdot E[t_{LU}]^2}{2(1-\varphi)} \right) = E[t_{LU}] \cdot \left(\frac{\varphi}{1-\varphi} \right) \tag{4}$$

Where C_B^2 denotes the squared coefficient of variation of the processing time. An important observation is that, clearly, the mean waiting time only depends upon the first two moments of the lookup processing time.

3 Performance Analysis

In this section, we will calculate the handover latency per movement for each protocol. Handover latency is defined for a receiving MN as the time that elapses between the disconnection with the previous attachment of point and the arrival of the first packet after the MN moves to NAR. We use a simple model for the data packet traffic, although the self similar nature of it has been noticed. Our packet traffic model has two layers namely session and packets. During a session, several packets are generated by a CN at an arbitrary rate and they reach an MN at the same rate. We assume that the session duration time has the exponential distribution with mean $E[t_o] = 1/\lambda_o$.

3.1 Network System Model and Mobility Model

We assume that a homogeneous network of which all wireless AP areas in a subnet domain have the same shape and size. First, we can define some parameters used for performance analysis. Let t_s and t_p be i.i.d. random variables representing the subnet domain residence time and the AP area residence time, respectively. Let $f_s(t)$ and $f_p(t)$ be the density function of t_s and t_p, respectively. In our paper, we suppose that an MN visit k AP areas in a subnet domain for a period t_s^k. During t_s^k, the MN resides at AP area i for a period t_i.

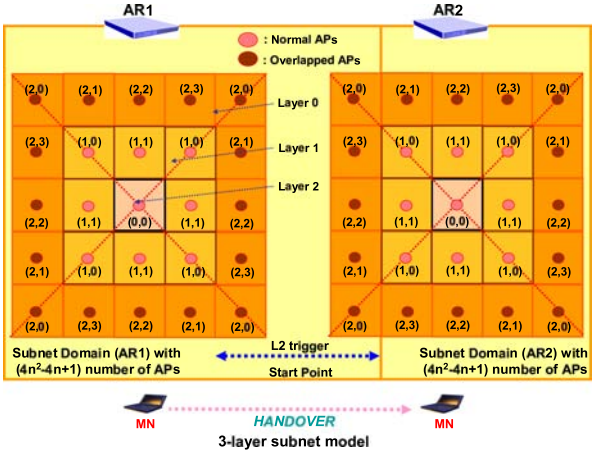


Fig. 1. 3-Layer Subnet Area Structure

Then, $t_s^k = t_1 + t_2 + t_3 + \cdots + t_{k-1} + t_k$ has the following density function

$$f_s^{(k)}(t) = \int_{t_1=0}^{t} \int_{t_2=0}^{t-t_1} \cdots \int_{t_{k-1}=0}^{t-t_1-\cdots-t_{k-2}} f_p(t_1) f_p(t_2) f_p(t_3) \cdots f_p(t_{k-1}) f_p(t-t_1-\cdots-t_{k-1}) dt_{k-1} \cdots dt_2 dt_1. \quad (5)$$

Using the Laplace transform convolution, we can determine the Lasplace transform for $f_s^{(k)}(t)$ as follows:

$$f_s^{(k)*}(s) = [f_p^*(s)]^k. \quad (6)$$

where $f_p^*(s)$ is the Laplace transform of $f_p(t)$.

We describe a two-dimensional random walk model for mesh planes in order to compute the subnet domain residence time density function. Our model is similar to reference [9] and considers a regular AP area/subnet domain overlay structure. We assume that an MN resides in an AP area for a period and moves to one of its four neighbors with the same probability, i.e. with probability 1/4. A subnet is referred to as a *n-layer* domain if it overlays with $N = 4n^2 - 4n + 1$ AP areas.

Fig. 1 shows the *3-layer* subnet domain architecture in which each of the 25 small squares and the entire square represents each of the AP areas and one subnet domain area, respectively. The AP area at the center of the subnet is called the *layer 0 AP area*. The AP areas that surround *layer x-1* AP areas are called *layer x* AP areas.

There are $8x$ AP areas in *layer x* except exactly one AP area which is in *layer 0*. An *n-layer* subnet overlays AP areas from *layer 0* to *layer n-1*. Particularly the AP areas that surround the *layer n-1* AP areas are referred to as boundary neighbors, which are outside of the subnet. According to the equal

moving probability assumption, we classify the AP areas in a subnet domain into several AP area types. An AP area type is of the form $< x, y >$, where x indicates that the AP area is in *layer* x and y represents the $y+1$st type in *layer* x. AP areas of the same type have the same traffic flow pattern because they are at the symmetrical positions on the mesh domain. For example, in Fig. 1, the AP type $< 1, 1 >$, $< 2, 1 >$ represent that this AP is in ring 1 and ring 2 and it is the AP of 2nd type in ring 1 and ring 2, respectively.

In the random walk model, a state (x, y) represents that the MN is in one of the AP areas of type $< x, y >$. The absorbing state (n, j) represents that an MN moves out of the subnet from state $(n-1, j)$, where $0 \leq j \leq 2n-3$. We assume that the AP area residence time of an MN has a Gamma distribution with mean $1/\lambda_p$ $(=E[t_p])$ and variance ν. The Gamma distribution is selected for its flexibility and generality. The Laplace transform of a Gamma distribution is

$$f_p^*(s) = \left(\frac{\gamma\lambda_p}{s+\gamma\lambda_p}\right)^{\gamma}, \qquad where\ \gamma = \frac{1}{\nu\lambda_p^2}. \tag{7}$$

Also, we can get the Laplace transform $f_s^*(s)$ of $f_s(t)$ and its expected subnet domain residence time $E[t_s]$ from [9]. For an MN, in the end, the probabilities $\Pi_p(i)$ and $\Pi_s(j)$ that the MN moves across i AP areas and j subnets during a session duration, can be derived as follows [10]:

$$\Pi_p(i) = \begin{cases} 1 - \frac{E[t_o]}{E[t_p]}(1 - f_p^*(\frac{1}{E[t_o]})) & , i = 0 \\ \\ \frac{E[t_o]}{E[t_p]}(1 - f_p^*(\frac{1}{E[t_o]}))^2 (f_p^*(\frac{1}{E[t_o]})^{i-1} & , i > 0 \end{cases} \tag{8}$$

$$\Pi_s(j) = \begin{cases} 1 - \frac{E[t_o]}{E[t_s]}(1 - f_s^*(\frac{1}{E[t_o]})) & , j = 0 \\ \\ \frac{E[t_o]}{E[t_s]}(1 - f_s^*(\frac{1}{E[t_o]}))^2 (f_s^*(\frac{1}{E[t_o]})^{j-1} & , j > 0 \end{cases} \tag{9}$$

3.2 Handover Latency Comparisons

At first, we introduce distance parameters used for handover latency functions. t_{WD} is the wireless component of the delay for a new AP re-association and authentication latency (MN's switching delay between APs). t_{RS} and t_{RA} are the transmission delays for the RS/RA messages in standard MIPv6, $(t_{RS} + t_{RA} = 2t_R)$. $^*t_{RD}$ is the random delay for RS, RA defined as the RFC 3775 ($^*t_{RD}$=$t_{RD\_RS} + t_{RD\_RA}$).

t_{BU} and t_{BAck} are the transmission delays for $BU/BAck$ messages respectively $(t_{BU} + t_{BAck} = 2t_B)$. t_{Packet} is the packet transmission delay from CN to MN. t_{DAD} is the DAD processing delay defined as the RFC 2462. t_{LU} is the lookup delay for DAD. ζ is the weighting factor of packet tunneling. ψ is the total delay between the time to exchange FBU/FBAck and the time of disconnection (Link-Down) with the current AP. t_{RS_FNA} and t_{RA_NAAck} are the transmission delays for RS with Fast Neighbor Advertisement and RA with Neighbor Advertisement Acknowledgment $(t_{RS\_FNA} + t_{RA\_NAAck} = 2t_{FR})$. $t_{NHI/HI}$ and

$t_{NHA/HAck}$ are the transmission delays for new NHI/NHA messages in proposed FMIPv6 with NAC and regular HI/HAck messages in standard FMIPv6 for address confirmation and to setup the tunnel between PAR and NAR. t_{Packet_MN} is the packet forwarding delay between MN and NAR. t_{Packet_PN} is the buffered packets forwarding delay from PAR to NAR. Using such parameters, for the standard MIPv6, FMIPv6 and proposed MIPv6 with NAC, the total handover latency per session duration is defined as follows:

$$T_{MIPv6} = \sum_{i=0}^{\infty}\{i\Pi_p(i)\cdot t_{WD}\} + \sum_{j=0}^{\infty}\{j\Pi_s(j)\cdot(2t_R +^* t_{RD} + t_{DAD} + 2t_B + t_{Packet})\} \tag{10}$$

In FMIPv6, MN sends FBU to PAR prior to disconnection with PAR. At this time, the handover procedure of FMIPv6 is divided into two independent procedures; H_I, the procedure to be executed by MN itself with PAR and NAR, and H_{II}, the procedure to be executed by both PAR and NAR to establish the bidirectional tunnel. The two separated procedures will combine into one when NAR receives FNA from MN after MN's subnet movement. We first assume that NAR has already received at least HI from PAR, when it receives FNA from MN. Before the two procedures H_I and H_{II} combine into one, the completion times of each procedure are defined as follows:

$$T_{H_I} = \psi + t_{WD} + t_{RS\_FNA} + t_{RA\_NAAck} \tag{11}$$

$$T_{H_{II}} = (t_{NHI/HI} + t_{NHA/HAck} + \zeta \cdot t_{Packet\_PN}) + t_{LU/DAD} \tag{12}$$

If H_{II} finishes before the completion of H_I(that is, $T_{H_I} > T_{H_{II}}$), NAR has buffered the packets tunneled from PAR and forwards them to MN when it receives FNA. if not, NAR waits the packets which will be tunneled from PAR when it receives FNA. At the latter case, NAR have to wait the completion of the address confirmation procedure. After announcing its attachment to NAR and receiving the tunneled packets, MN sends binding update messages with its new CoA to HA, and to CNs consecutively. In FMIPv6, the total handover latency per a session time are defined in Eq. 13. In our paper, the SRA message and L2 information are triggered together with an association response message. We assume that $t_{NHI/HI}$, $t_{NHA/HAck}$, t_{RS_FNA} and t_{RA_NAAck} have the same value in transmission time. Also, t_{BU} and t_{BAck}have the same value in transmission time. In our proposed scheme, t_{LU} is the most important factor determining the performance.

$$T_{FMIPv6} = \sum_{i=0}^{\infty}\{i\Pi_p(i)\cdot t_{WD}\} + \sum_{j=0}^{\infty}\{j\Pi_s(j)\cdot(MAX\{T_{H_I}, T_{H_{II}}\} + \zeta\cdot t_{Packet\_MN} - t_{WD} - \psi)\} \tag{13}$$

And, from the above function with $t_{LU} = 3.36\ \mu$ sec and $5.28\ \mu$ sec in Eq.12, we can get the handover latency for the enhanced FMIPv6 equipped with NAC scheme.

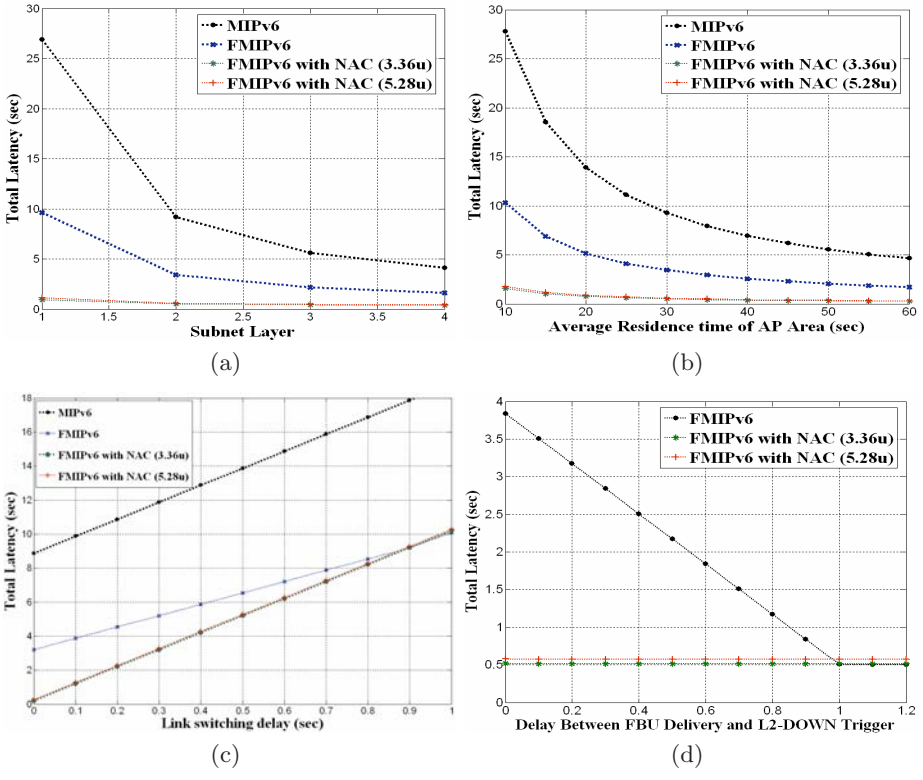


Fig. 2. Total Handover Latency Comparison

3.3 Numerical Results

For examinations, the following fixed parameters are used: $t_{RS/RA} = 0.015$, $^*t_{RD} = 1.5$, $t_{DAD} = 1$, $t_{BU/BAck} = 0.065$, $t_{NHI/HIandNHA/HAck} = 0.01$, ν=1.0, $\zeta = 1.2$, λ_0=0.0033 (session duration is 300sec) and $t_{Packet/Packet\_MN/Packet\_PN}$ = 0.065/0.015/0.01. As the target of investigation, we select the following changeable parameters and their default values: n=2 (subnet layer is 2), λ_p(mean of AP area residence time is 30 sec.), t_{WD} = 0.03sec., and $\psi = 0.1$. While we select one parameter and change its value, the remaining parameters values are set to their default values during the following investigation. Fig. 2 explains the total handover latency per session duration with respect to each changeable parameter. From the figures, we can know that proposed FMIPv6 with *NAC* handover latency are considerably reduced the address configuration and confirmation process.

Fig. 2 (a) shows the total handover latency of each protocol with respect to the subnet layer. It shows that the reduction of latency becomes high when a subnet contains many AP areas. The figure show that proposed FMIPv6 with *NAC* is under little influence of such system deployment. Fig. 2 (b) shows that the handover process occupies much time within the whole session duration when

MN moves across AP areas and subnets more frequently. Fig. 2 (c) shows the relationship between the handover latency and the delay of link switching in a session duration. When the switching delay (t_{WD}) becomes high, all protocols' handover latency become high, too. When the link switching delay is 1, the procedure H_I becomes the dominant factor of handover latency. Fig. 2 (d) shows that FMIPv6's handover latency becomes low if MN sends FBU to PAR more early before it disconnects with PAR. If FBU can be delivered to PAR as soon as possible, NAR receives HI early in FMIPv6 handover process.

4 Conclusion

In this paper, we have introduced the proposed FMIPv6 with *NAC*. The use of a modified neighbor cache with look up algorithm has merits, such as a faster DAD checking speed, which solves the short-comings of normal DAD when a router has more than two links. We also can obtain alternative addresses by managing addresses in the network. In the numerical analysis, we developed packet traffic, system and mobility models. Based on the numerical results, we can see that the major benefits of our scheme are to remarkably reduce CoA configuration and confirmation latency concerned in any seamless handover schemes, and preventing address collision from occurring provided there is no packet loss.

References

1. D.Johnson, C. Perkins, J. Arkko, "Mobility Support in IPv6", RFC 3775, June 2004.
2. S. Thomson, T. Narten, "IPv6 Stateless Address Auto-configuration", RFC 2462, Dec. 1998.
3. Narten, T., Nordmark, E. and W. Simpson, "Neighbor Discovery for IP version 6 (IPv6)", RFC 2461, December 1998.
4. Koodli, R., "Fast Handovers for Mobile IPv6", RFC 4068, July 2005.
5. Byungjoo. Park, Y-H. Han, H. A. Latchmann, "EAP: New Fast Handover Scheme based on Enhanced Access Point in Mobile IP Networks", International Journal of Computer Science and Network Security, Vol. 6, No.9, pp. 69-75, Sep. 2006.
6. JinHyoeck Choi, DongYun Shin, "Router Advertisement Cashing in Access Point (AP) for Fast Router Discovery", draft-jinchoi-mobileip-frd-00.txt, June 2002.
7. V. Srinivasan, G. Varghese, "Fast Address Lookups Using Controlled Prefix Expansion", ACM Transactions on Computer System, Vol.17, Feb.1999.
8. R. Kawabe, S. Ata, M. Murata, "On Performance Prediction of Address Lookup Algorithms of IP Routers through Simulation and Analysis Techniques", IEEE International Conference on Communications 2002 (ICC 2002).
9. I. F. Akyildiz, Y.B. Lin, W. R. Lai, and R. J. Chen, "A new Random Walk Model for PCS Networks," IEEE JSAC, Vol.18, No.7, pp.1254-1260, July 2000. 10.
10. Y. H. Han, "Hierarchical Location Chacing Scheme for mobility Managament", Dept. of Computer Science and Engineering, Korea University, Dec. 2001.

DCAR: Dynamic Congestion Aware Routing Protocol in Mobile Ad Hoc Networks

Young-Duk Kim, Sang-Heon Lee, and Dong-Ha Lee

Daegu Gyeongbuk Institute of Science and Technology (DGIST)
Deoksan-Dong 110, Jung-Gu, Daegu, 700-742, Korea
{ydkim, pobbylee, dhlee}@dgist.org

Abstract. In mobile ad hoc networks, most of on demand routing protocols such as DSR and AODV do not deal with traffic load during the route discovery procedure. To achieve load balancing in networks, many protocols have been proposed. However, existing load balancing schemes do not consider the remaining available buffer size of the interface queue, which still results in buffer overflows by congestion in a certain node which has the least available buffer size in the route. To solve this problem, we propose a load balancing protocol called Dynamic Congestion Aware Routing Protocol (DCAR) which monitors the remaining buffer length of all nodes in routes and excludes a certain congested node during the route discovery procedure. We also propose two buffer threshold values to select an optimal route selection metric between the traffic load and the minimum hop count. Through simulation study, we compare DCAR with other on demand routing protocols and show that the proposed protocol is more efficient when a network is heavily loaded.

Keywords: Ad hoc networks, Routing protocols, Load balancing.

1 Introduction

A mobile ad hoc network (MANET) is a self-configuring network of mobile hosts connected by wireless links without fixed infrastructure such as base station. In MANETs hosts are free to move randomly, and thus network topologies may change rapidly and unpredictably. Devising an efficient routing protocols for MANETs has been a challenging issue and DSDV (Destination Sequence Distance Vector) [1], DSR (Dynamic Source Routing) [2], AODV (Ad-hoc On-demand Distance Vector) [3] are such protocols to tackle the issue.

Recently, the requirement for real time and multimedia data traffic continues growing. In this situation, the occurrence of congestion is inevitable in MANETs due to limited bandwidth. Furthermore, by the route cache mechanism in the existing protocols, the route reply from intermediate node during the route discovery procedure leads to traffic concentration on a certain node. When a node is congested, several problems such as packet loss by buffer overflows, long end-to-end delay of data packets, poor packet delivery ratio, and high control packet

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 130–139, 2007.

overhead to the reinitiate the route discovery procedure can occur. In addition, the congested node consumes more energy to route packets, which may result in network partitions.

In this paper, we propose the DCAR (Dynamic Congestion Aware Routing Protocol) which ties to distribute traffic load and avoid congested nodes during the route discovery procedure. DCAR monitors number of packets in an interface queue and defines traffic load as the minimum available buffer length among the nodes in the route. By avoiding the node with minimum available buffer length in the route, we can achieve load balancing, and improve performance in terms of packet delivery ratio and end-to-end delay, etc.

The rest of this paper is organized as follows. In Section II, we review two protocols DSR, DLAR [4]. In Section III and IV, we illustrate the motivation and detail operation of our proposed protocol. Performance evaluation by simulations is presented in Section IV. Finally, concluding remarks are given in Section VI.

2 Related Works

2.1 DSR (Dynamic Source Routing Protocol)

DSR is an on demand routing style protocol for ad hoc networks. Every source node knows a complete route to a destination and maintains a route cache containing the source routes that it is aware of. Each node updates the entries in the route cache if there is a better route, when it learns about a new one. Two main mechanisms of DSR are route discovery and route maintenance.

The route discovery procedure is initiated in an on-demand basis when a source node requires a route to a destination for routing. At first, if there is no route available in the route cache, the source broadcasts a Route Request (RREQ) packet which is flooded throughout the entire network. Each RREQ packet contains a record of listing the address of each intermediate node as well as initiator (source) and target identifier (destination) of the RREQ. If a node receiving the RREQ packet is the destination or an intermediate node having a path to the destination node in its cache, it can reply to the RREQ by sending a Route Reply (RREP) packet which contains the route information between the source and the destination. When the source node receives this RREP, it stores this route in its route cache for sending subsequent data packets to this destination.

In route maintenance procedure, when a node detects that its descendant node in the route is unreachable either by no packet receipt confirmation from the descendant node or no link level acknowledgement in the link layer, it sends a Route Error (RERR) packet to the source node. The RERR packet contains addresses of two end nodes of the broken link. During the propagation of the RERR packet to the source, every intermediate node in the route as well as the source removes the broken route entry in its own route cache and the source invoke the route discovery process again to construct a new route.

2.2 DLAR (Dynamic Load Aware Routing Protocol)

DLAR [4] is a DSR based load balancing routing protocol that uses the traffic load information of the intermediate nodes as the main route selection criterion. Similar to DSR, DLAR is also an on-demand routing protocol and has two main mechanisms of route discovery and route maintenance. Figure 1 illustrates the protocol operation of DLAR for route selection.

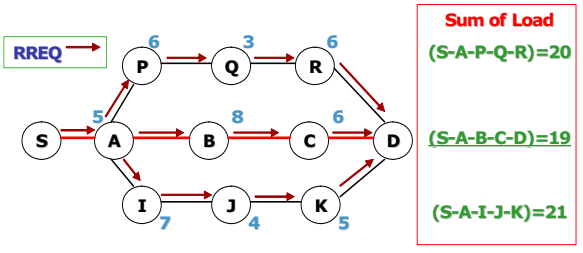


Fig. 1. Operation of DLAR

In route discovery procedure of DLAR, the source node S broadcasts the RREQ packet to its neighbors to find a route. When an intermediate node receives the RREQ packet, it sums and attaches its own load information, then rebroadcast the RREQ packet. The load information of the node is defined as the number of packets that is currently buffered in its interface queue. All nodes in the network monitor this load information. Unlike to DSR, an intermediate node does not send a RREP packet on behalf of the destination in order to deliver fresh entire load information of the route to the destination. The destination node D can receive multiple RREQ packets from different routes for some amount of time. After receiving RREQ packets, D selects a best route presumed to be the one having the least load and sends a RREP packet to the source node via the reverse path. In the figure, the route S-A-B-C-D is chosen because the route has the least sum (19).

In the route maintenance procedure of DLAR, intermediate nodes that are in an active data session periodically piggyback their load information on data packets to report the load status of the active path. If the active path is believed to be congested, the source node reinitiates the route discovery procedure and finds an alternative route. When the intermediate node finds a broken link, it sends a RERR packet to the source node and the source node restarts the route discovery procedure. Elements of the figure described in the caption should be set in italics, in parentheses, as shown in this.

2.3 Other Routing Protocols with Load Balancing

There are other routing protocols that consider load balancing as the primary route selection criterion. However, their protocol operations are similar to that of

DLAR or DSR. Thus, we only present their main differences without describing the protocol operations. In LBAR (Load-Balanced Ad hoc Routing) [5], the network load is defined as total number of active routes passing through the node and its neighbors. During the route discovery procedure, load information on all paths from the source to a destination is forwarded to the destination node. In TSA (Traffic-Size Aware Routing) [6], the network load is defined as traffic sizes of routes, which is presented in bytes, not in number of packets because the packet sizes may vary. In MCL (Routing Protocol with Minimum Contention Time and Load Balancing) [7], the network load is defined as the number of neighbors which content with a source node. In CRP (Congestion-adaptive Routing Protocol) [8], although the number of packets currently buffered in interface is also defined as network load, the congestion is classified into three statuses, which are red (very likely congested), yellow (likely congested), and green (far from congested). If a node is aware of congestion symptom, it finds a bypass route which will be used instead of the congested route.

3 Motivation

As discussed in the previous section, DLAR is a load-balancing protocol which establishes a route with minimum load. However DLAR only monitors the number of packets buffered in a node's interface and monitoring the number of buffered packets does not directly reflect the situation of network congestion. Figure 2 illustrates this problem.

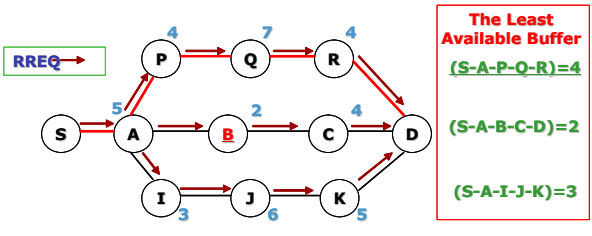


Fig. 2. Operation of DCAR

The Figure 2 is the same topology as Figure 1 except that it additionally includes the number of remaining packets in each node's interface queue. The maximum size of each buffer is assumed to be 10. When the number of currently buffered packets are used a primary key for selecting a route, like DLAR, the destination node D selects the route [S-A-B-C-D] which has the least sum. However, if we look at the remaining available buffer size, node B in the route selected by DLAR is most likely to be congested because its remaining buffer size is only 2. When a node does not have enough space to accommodate data packets originated from the new route, the routes including the node should

be excluded from the route selection. In Figure 2, the route containing node B, which is selected as the best route by DLAR, should be avoided. Another problem of DLAR is that it does not consider the minimum hop count metric significantly. In DLAR, a destination node uses the hop count to select a route only when two or more routes have the even load sums. Lastly, we must consider a case when the buffer size of each node varies, because the packet processing capacity of each node is different from another. In such a case, DLAR can not measure the exact traffic load in every node.

The problems addressed above clearly motivates us to devise a new protocol that considers the minimum available buffer size as one of the primary route selection criteria to avoid the most congested nodes and to achieve load balancing in ad hoc networks.

4 Proposed Protocol

In this section, we present the proposed protocol, referred to as DCAR (Dynamic Congestion Aware Routing Protocol), to improve the performance by avoiding the congested nodes during the route discovery procedure in mobile ad hoc networks.

4.1 Route Discovery and Selection Procedure

DCAR is an on-demand routing algorithm and assumes that every node in the network is aware of its own traffic load by monitoring the available buffer size of its interface. When a source node wants to send data packets, the source starts a route discovery process by broadcasting a RREQ packet to the entire network. To find the most congested node in the discovered routes, we define, Q_{min}, the minimum available buffer size among the nodes in the route. Each RREQ packet includes a unique identifier and Q_{min} fields. In the proposed protocol, if an intermediate node receives duplicate RREQ packets that have bigger Q_{min} than the previous one, it can rebroadcast the RREQ packets because the new route consists of less congested nodes. Otherwise, it drops the duplicated ones. When the intermediate node receives the first RREQ packet, it compares Q_{min} in the received RREQ with its own traffic load, represented by the available buffer size. If the traffic load of intermediate node is smaller than received Q_{min}, the node replaces it with its own information and floods the RREQ packet.

As shown in Figure 2, the route discovery procedure of the proposed protocol can be described as follows. The source node S floods a RREQ packet to find a route to the destination node D. When node A receives the RREQ packet, it updates Q_{min} with 5 and rebroadcasts the packet. Then the next node P receives the RREQ and compares Q_{min} (=5) with its own remaining buffer size (=4). Since Q_{min} in the RREQ packet is greater than node P's remaining buffer size, it replaces Q_{min} with its remaining buffer size (=4).

After the same operation is done in node Q and R, the destination node D finally receives the RREQ packet containing Q_{min} of 4 through the route [S-A-P-Q-R-D]. Node D also receives RREQ packets from other routes: the route

[S-A-B-C-D] having Q_{min} of 2 and [S-A-I-J-K-D] having Q_{min} of 3. Once the first RREQ packet has arrived at node D, it sends a RREP packet to node S by using the reverse path. If node D receives a duplicate RREQ packet with bigger Q_{min}, it immediately sends the RREP packet again to node S to change the active route with less congested nodes. Otherwise, it simply drops the duplicate RREQ packets.

When node D selects an optimal route, it considers the minimum hop count as well as the traffic load. The detail of the route selection algorithm is described in the following section.

During the route discovery procedure, our protocol does not allow intermediate nodes to send the RREP packet using its own route cache, because all RREQ packets have to be delivered to the destination to check the congestion status of the entire route. If the intermediate nodes can send the RREP packet, the route obtained from the route cache may be stale, especially when the nodes are highly mobile. Thus, by prohibiting intermediate nodes from sending the RREP packet, we can obtain fresh route information.

4.2 Route Selection Algorithm

When the destination node receives multiple RREQ packets, the route selection algorithm is used to choose an optimal route. The main operation is to select the route with biggest Q_{min} value among the received RREQ packets. However by selecting a route with only load information, the route length may be long, which result in high delivery latency. So we define two thresholds which can find out whether the route should be selected by the load information or the hop count metric. The first threshold is Max-Threshold (T_{max}) which defines congestion criteria in a node. For example, when T_{max} is 30, we believe that Q_{min} with more than 30 is not congestion environment. Thus the destination node selects the route with minimum hop count metric. The second threshold is Diff-Threshold (T_{diff}) which is a numerical difference between Q_{min} values of two routes. For example, if T_{diff} is 5 and the difference between two routes is less than 5, we believe that the two load information is almost same. Thus the destination node chooses the route with shortest distance.

4.3 Route Maintenance

Route maintenance procedure in DCAR is similar to DSR. If a node detects link breakdown, it sends a Route Error (RERR) packet to the source node along the active path. When a node receives the RERR packet, it removes this broken link from its route cache and performs a packet salvaging process, which attempts to salvage the data packet rather than dropping it. In the packet salvaging process, the node sending a RERR packet searches its own Route Cache for a route from itself to the destination node. If the source node receives the RERR packet from its neighbor node, it will restart the route discovery process to find an alternative route to the destination node.

5 Performance Evaluation

5.1 Simulation Environment

To evaluate the performance of the proposed protocol, we used the ns-2 simulator (version 2.28) [9] with the IEEE 802.11b DCF using RTS/CTS. There are 50 mobile nodes that are assumed to be randomly placed in a 1500m x 300m rectangle network area. All mobile nodes moved freely at the given maximum speed of 10m/s with the pause time of 0 during the simulation time of 300 seconds. The radio propagation range for a node is set to 250m. 20 data connections are established with 5 different packet rates of 5, 10, 15, 20, and 25 to represent different network traffic load. Each pair of source and destination nodes of a connection is randomly selected without duplicate sources. Each source generates constant bit rate (CBR) traffic with packet size of 512 bytes. The maximum buffer size of each node's interface is set to 50 and 3 different buffer Max-threshold values of 45, 20, 10 and 3 different Diff-threshold values of 5, 3, and 2 are used for the simulation study.

5.2 Simulation Result

Figure 3 shows the averaged number of dropped packets in a node's interface queue by buffer overflows. As shown in the figure, DCAR provides less buffer overflows because during the route discovery procedure DCAR can avoid congested nodes and can achieve load balancing in the network while the other protocols have frequent packet drops by buffer overflows, which eventually leads to route breakdowns.

Figure 4 shows the packet delivery ratios of DCAR, DLAR and DSR as a function of traffic load. The delivery ratio of DCAR is better than those of DLAR and DSR due to less frequent buffer overflows. Although DLAR also can avoid the congested routes, the performance of DCAR is better because DLAR

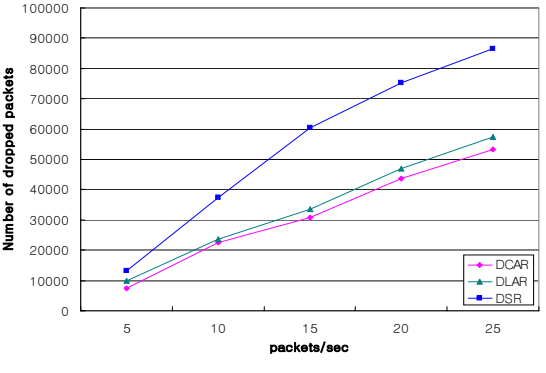


Fig. 3. Number of dropped packets

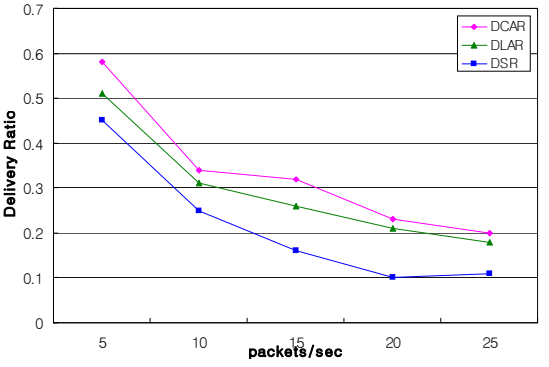


Fig. 4. Packet delivery ratio

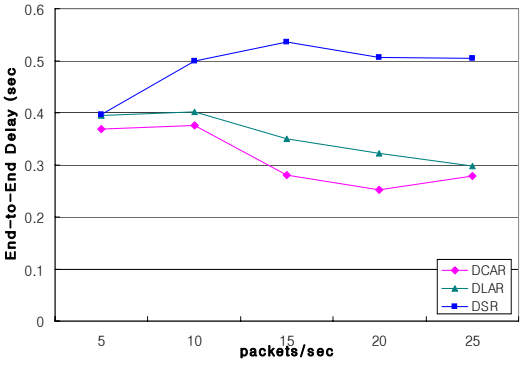


Fig. 5. End-to-end delay

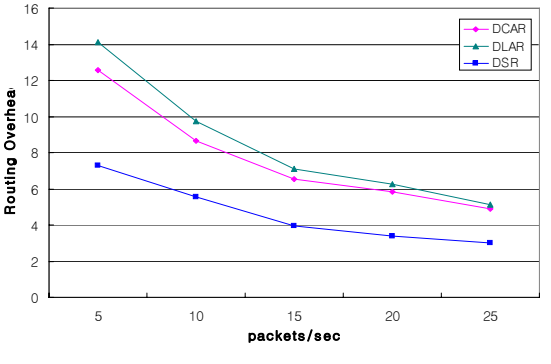


Fig. 6. Normalized routing overhead

does not know the most congested nodes in routes. However, when the packet rate is over 25, delivery ratios of all the protocols are saturated because the entire network is congested.

Figure 5 shows the packet end-to-end delay as a function of traffic load. When the network traffic load increases, the end-to-end delay of DSR also increases. However, the delays of DCAR and DLAR decreases because these protocols can avoid congested nodes and congested routes. In DSR, the end-to-end delay decreases when the packet rate is above 15. When the traffic load is high and the intermediate nodes are congested, the RREQ packets are also dropped by buffer overflows, so the congested nodes can not forward RREQ packets as well as data packets to the destination. Thus DSR can avoid the congested nodes automatically during the route discovery procedure. In the figure, when compared to DLAR, we can see that the overall performance of DCAR is improved about 10% in terms of the packet delivery ratio and the end-to-end delay.

Figure 6 shows the normalized routing overhead which is the number of the control packets transmitted per data packet successfully delivered at the destination node. We can see that routing overhead of DCAR is larger than that of DSR because DCAR does not allow an intermediate node to send a RREP packet using its own route cache. Thus all RREQ packets are delivered to the destination node by flooding, which results in increased number of control packets during the route discovery process. This is same reason why DLAR has also high control packet overhead. However, the overhead of DLAR is a little bit higher than DCAR because DLAR has more frequent buffer overflows as shown in Figure 3. And we can see that as the traffic load increases, there are more buffer overflows, which leads the control packet overhead to decrease by dropping RREQ packets.

Finally, Table I and Table II show the comparison of the performance with different buffer threshold values (Max-threshold and Diff-threshold) of DCAR in order to find the most efficient route. Although it is not easy to select the optimal values, we can see that the buffer threshold value affects the protocol's performance by setting differently. In both scenarios of different packet rates, we can find that DCAR shows the best performance when T_{max} is 20 and T_{diff} is 3, which are approximately correspond to 50% and 5% of the total buffer size, respectively.

Table 1. Various threshold values of DCAR with 5 packets/sec

Threshold		5 packets/sec		
T_{max}	T_{diff}	Delivery Ratio	End-to-End Delay	Overflow Dropped
45	5	0.54	3.89	9424
20	3	0.58	3.69	7518
10	2	0.58	3.7	7602

Table 2. Various threshold values of DCAR with 20 packets/sec

Threshold		20 packets/sec		
T_{max}	T_{diff}	Delivery Ratio	End-to-End Delay	Overflow Dropped
45	5	0.21	3.28	46117
20	3	0.23	2.52	44502
10	2	0.23	2.94	45513

6 Conclusion

In mobile ad hoc networks, congestion can lead to performance degradation such as many packet losses by buffer overflows and long end-to-end delay. However, existing load balancing protocols do not consider the available buffer size in node's interface queue. That is, they do not consider a certain congested node. In this paper, we have proposed DCAR (Dynamic Congestion Aware Routing Protocol) which can monitor the most congested node in route and can avoid it during the route discovery procedure because the RREQ packet of DCAR contains the minimum available buffer size among the nodes in a discovered route. We also defined two buffer thresholds to choose the route selection metric between the traffic load and the minimum hop count. Simulation study shows that DCAR shows a good performance in terms of packet delivery ratio, end-to-end delay, routing overhead when a network is heavily loaded.

References

1. C. E. Perkins and P. Bhagwat, "Highly Dynamic Destination Sequenced Distance-vector Routing for Mobile Computers," Comp. Commun. Rev., October. 1994
2. David B. Johnson, David A. Maltz, Yih-Chun Hu, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)," Internet Draft, IETF Mobile Ad hoc Networks (MANET) Working Group
3. C. E. Perkins and E. Royer, "Ad-hoc on-demand Distance Vector Routing," Proc.2nd IEEE Wksp. Mobile Comp. Sys. App., February 1999
4. Sung-Ju Lee and Mario Gerla, "Dynamic Load-Aware Routing in Ad hoc networks," Proceedings of IEEE ICC 2001
5. H. Hassanein and A. Zhou, "Routing with Load Balancing in Wireless Ad hoc Networks," ACM MSWiM, July 2001
6. Abdulranhman H. Altalhi and Golden G. Richard, III, "Load-Balanced Routing through Virtual Paths," IPCCC 2004
7. Bong Chan Kim, Jae Young Lee, Hwang Soo Lee and Joong Soo Ma, "An Ad-hoc Routing Protocol with Minimum Contention Time and Load Balancing," IEEE Global Telecommunications Conference, 2003
8. Tran, D.A. and Raghavendra, H., "Routing with congestion awareness and adaptivity in mobile ad hoc networks," IEEE WCNC 2005
9. S.McCanne and S. Floyd, "NS network simulator," URL: http://www.isi.edu/nsnam/ns

Anonymous Secure Communication in Wireless Mobile Ad-Hoc Networks

Sk. Md. Mizanur Rahman[1], Atsuo Inomata[2], Takeshi Okamoto[1], Masahiro Mambo[1], and Eiji Okamoto[1]

[1] Graduate School of Systems and Information Engineering, University of Tsukuiba, Japan
[2] Japan Science and Technology Agency, Tokyo, Japan

Abstract. The main characteristic of a mobile ad-hoc network is its infrastructure-less, highly dynamic topology, which is subject to malicious traffic analysis. Malicious intermediate nodes in wireless mobile ad-hoc networks are a threat concerning security as well as anonymity of exchanged information. To protect anonymity and achieve security of nodes in mobile ad-hoc networks, an anonymous on-demand routing protocol, termed RIOMO, is proposed. For this purpose, pseudo IDs of the nodes are generated considering Pairing-based Cryptography. Nodes can generate their own pseudo IDs independently. As a result RIOMO reduces pseudo IDs maintenance costs. Only trust-worthy nodes are allowed to take part in routing to discover a route. To ensure trustiness each node has to make authentication to its neighbors through an anonymous authentication process. Thus RIOMO safely communicates between nodes without disclosing node identities; it also provides different desirable anonymous properties such as identity privacy, location privacy, route anonymity, and robustness against several attacks.

Keywords: Ad-hoc network, Anonymity, Routing, Pairing-Based Cryptography, Security.

1 Introduction

Conventional wireless mobile communications are normally supported by a fixed wire/ wireless infrastructure. In contrast, mobile ad-hoc networks, MANETs do not use any fixed infrastructure. So, the shared wireless medium MANETs, introduces opportunities for passive eavesdropping on data communications. Thus traffic analysis is one of the most subtle and unsolved security attacks against MANETs. By definition, it is an attack such that an adversary observes network traffic and infers sensitive information of the applications and/or the underlying system [1].

Anonymity and/or privacy is an important criteria for securing ad-hoc network communication. Anonymity ensures that a user may use a resource or service without disclosing the user's identity. Thus anonymity requires that other users or subjects are unable to determine the identity of a user bound to a subject or operation [2]. If anonymity is the stronger the less is known about the

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 140–149, 2007.

linking to a subject. As a result, adversaries fail to make correlation between the eavesdropped traffic information and the actual network traffic patterns. Thus traffic analysis attack can be efficiently defeated. In this paper an anonymous on-demand routing protocol, called RIOMO, is proposed. In RIOMO, every node can generate its own pseudo IDs dynamically and independently based-on pairing-based cryptography, without making communication with the system administrator. Thus pseudo IDs maintenance cost is reduced compared to the previous proposed method namely MASK by Zhang et al., [3]. A route is discovered without disclosing the nodes IDs for successful communication.

The remainging of this paper is organized as follows. In section 2, preliminaries are described. In section 3, RIOMO architecture and design are given. In section 4 RIOMO protocol is described. In section 5 anonymity achievements and security analysis are given. Finally, section 6 describes conclusions and future works.

1.1 Related Work and Our Contributions

The proposed protocol RIOMO is exclusively based on pairing-based cryptographic properties. There is also another approach of anonymous communication based on pairing-based cryptography proposed by Zhang et al., [3], called MASK. In MASK, system administrator generates a large set of pseudo IDs for every node, thus every node has a fixed pseudo ID set and it should be large enough, otherwise there is a chance of finding pseudonym linking by the intruders. To keep strong anonymity in MASK, every node should have to manage an extremely large enough number of pseudo IDs set provided by the system administrator, which is costly for ad-hoc network communication in terms of extra task for nodes namely IDs maintenance cost. In this paper we explicitly show, by using only one pseudo ID taking from system administrator, nodes can generate their own pseudo IDs, independently and dynamically. It is the first approach to achieve anonymity by using only one pseudo ID taking from the system administrator in ad-hoc network. With pairing based IBE properties and random number nodes can generate their own pseudo IDs dynamically, which also provide strong security properties.

There are some other proposals [4,5,6,7] taking care of privacy. In [4], a secure dynamic distributed routing algorithm (denoted as SDDR in this paper) for ad hoc wireless networks is proposed based on the onion routing protocol [5]. The anonymity-related properties achieved in this algorithm include weak location privacy and route anonymity. However, it ignores one important part of privacy in mobile ad-hoc networks, namely identity anonymity, and it cannot provide strong location privacy.

In ref.[6], Kong et al. design an Anonymous On-Demand Routing (ANODR) based on topology. Similar to Hordes [7], ANODR also applies multicast/ broadcast to improve recipient anonymity. ANODR is an on-demand protocol, and is based on trapdoor information in the broadcast. These features are not discussed in regards to Hordes' [7] multicast mechanism.

Compared to ref.[4], ANODR is more efficient than SDDR at the data-transmission stage. However, similar to SDDR in [4], ANODR does not provide identity anonymity and strong location privacy. RIOMO and other two protocols are described in Table 1 with respect to the Anonymity and security related properties. For anonymity related properties ✓: indicates property is achieved, and blank indicates property is not achived, *: indicates identity privacy of source and destination, **: indicates identity privacy of forwarding nodes in route. For security related properties ✔: indicates attack is protected and blank indicates not protected. Detailed discussions of these properties are given in Section 5.

Table 1. Comparison of anonymity and security related properties

Anonymity and security properties	Routing protocol		
	SDDR	ANODR	RIOMO (proposed)
Identity privacy*	✓		✓
Identity privacy**		✓	✓
Weak location privacy	✓	✓	✓
Strong location privacy			✓
Route anonymity		✓	✓
DoS attacks			✔
Wormhole attacks	✔	✔	✔
Rushing attacks	✔	✔	✔

2 Preliminaries

In this section, we just describe some preliminaries and mathematical properties which are useful to understand our proposed protocol.

2.1 Bilinear Maps

Let G_1 be an additive group and G_2 be a multiplicative group of the same prime order q. Let P be an arbitrary generator of G_1. *(aP denotes P added to itself a times).* Assume that discrete logarithm (DL) problem is hard in both G_1 and G_2. We can think G_1 as a group of points on an elliptic curve over F_q, and G_2 as a subgroup of the multiplicative group of a finite field F_{q^k} for some $k \in Z_q^*$. A mapping $\tilde{e} : G_1 \times G_1 \rightarrow G_2$, satisfying the following properties is called a cryptographic bilinear map.

- *Bilinearity:* $\tilde{e}(aP, bQ) = \tilde{e}(P, Q)^{ab}$ for all $P, Q \in G_1$ and $a, b \in Z_q^*$. This can be restated in the following way. For $P, Q, R \in G_1, \tilde{e}(P + Q, R) = \tilde{e}(P, R)\tilde{e}(Q, R)$ and $\tilde{e}(P, Q + R) = \tilde{e}(P, Q)\tilde{e}(P, R)$.
- *Non-degeneracy:* If P is a generator of G_1, then $\tilde{e}(P, P)$ is a generator of G_2. In other words, $\tilde{e}(P, P) \neq 1$.
- *Computable:* A mapping is efficiently computable if $\tilde{e}(P, P)$ can be computed in polynomial-time for all $P, Q \in G_1$.

Modified Weil Pairing [8]and Tate Pairing [9]are examples of cryptographic bilinear maps.

2.2 Diffie-Hellman Problems

With the group G_1 described in section 2.1, we can define the following hard cryptographic problem applicable to our proposed scheme.

- *Discrete Logarithm (DL) Problem:* Given $P, Q \in G_1$, find an integer n such that $P = nQ$ whenever such integer exists.
- *Computational Diffie-Hellman (CDH) Problem:* Given a triple $(P, aP, bP) \in G_1$ for $a, b \in Z_q^*$, find the element abP.
- *Decision Diffie-Hellman (DDH) problem:* Given a quadruple $(P, aP, bP,$-$cP) \in G_1$ for $a, b, c \in Z_q^*$, decide whether $c = ab \bmod q$ or not.
- *Bilinear Diffie-Hellman (BDH) Problem:* Given a quadruple (P, aP, bP, cP)-$\in G_1$ for some $a, b, c \in Z_q^*$, compute $\tilde{e}(P, P)^{abc}$.

Groups where the *CDH* problem is hard but *DDH* problem is easy are called GAP Diffie-Hellman (*GDH*) groups. Details about *GDH* groups can be found in [10].

3 RIOMO Architecture and Design

In RIOMO, system administrator does not take part in routing rather it has the following tasks during the boot strap of the network.

- Determines two groups G_1, G_2, of the same prime order q. We view G_1 as an additive group and G_2 as a multiplicative group as discussed in section 2.1.
- Determines bilinear map $g : G_1 \times G_1 \rightarrow G_2$, collision resistant cryptographic hash functions H_1 and H_2, where $H_1 : \{0,1\}^* \rightarrow G_1$, a mapping from arbitrary-length strings to points in G_1 and $H_2 : \{0,1\}^* \rightarrow \{0,1\}^\mu$, a mapping from arbitrary-length strings to μ-bit fixed length output.
- Generates system's secret $\omega \in Z_q^*$, where $Z_q^* = \{y | 1 \leq y \leq q-1\}$. Any one in the network does not know ω except system administrator. System administrator also uses this secret to generate the secret point of the non-adversary nodes.

Thus the system parameters $< G_1, G_2, g, H_1, H_2 >$ are known to the non-adversary nodes. System administrator also provides the following parameters for nodes, regarding their IDs and secret points.

- Provides each node, a secret point SP_R, with respect to the node's real ID ID_R, which is defined as $SP_R = \omega H_1(ID_R)$. The Source and the destination use their corresponding secret point in the route discovery phase to authenticate each other. For a given set of $< ID_R, SP_R >$ no one can determine the system secret ω as we discussed in section 2.1 and 2.2.

– Provides each node a pseudo ID IDP_i, and their corresponding secret point SPP_i, which is defined as $SPP_i = \omega H_1(IDP_i)$; if $i \neq j$ then $IDP_i \neq IDP_j$ as well as $SPP_i \neq SPP_j$. For a given set of $< IDP_i, SPP_i >$ no one can determine the system secret ω.

With the above information any node can generate its own *pseudo* IDs and the corresponding secret points randomly in every session in communication. Let's check for a node, namely K; K has received its pseudo ID IDP_K and the corresponding secret point $SPP_K = \omega H_1(IDP_K)$ from the system administrator. Now, K is able to generate its own pseudo ID $ID_{PK} = R_K H_1(IDP_K)$, and the corresponding secret point $SP_{PK} = R_K SPP_K = R_K \omega H_1(IDP_K) = \omega R_K$-$H_1(IDP_K) = \omega ID_{PK}$, where R_K is a random generated by K; this relation also holds the previous cited property in section 2.1 and 2.2, that is no one can determine the system secret ω for a given set of pseudo ID and the corresponding secret point , $< ID_{PK}, SP_{PK} >$. Thus a node can generate its own pseudo IDs and corresponding secret points when it is needed.

4 RIOMO Protocol

4.1 Anonymous Neighbor Authentication

When a node wants to join in the network or moves to a new place, it has to authenticate within its neighbor nodes. Say, Alice has received her pseudo ID IDP_A, and the corresponding secret point $SPP_A = \omega H_1(IDP_A)$, i.e.,$< IDP_A, SPP_A >$ from the system administrator. She can join in the network by authenticating within her neighbor nodes or if she moves another place in the network different from her current place, she also needs to authenticate her within her neighbor, to avoid a *target oriented* attack. If Alice wants to change her pseudo ID different from her current pseudo ID without moving her place, she also needs to authenticate her current pseudo ID within her neighbor. For this purpose she generates pseudo ID $ID_{PA} = R_A H_1(IDP_A)$, and corresponding secret point $SP_{PA} = R_A SPP_A = R_A \omega H_1(IDP_A) = \omega R_A H_1(IDP_A) = \omega ID_{PA}$, where R_A is a random generated by Alice; she also generates a random R_{RA} which is used to generate verification codes Ver_0^* and Ver_1. Alice broadcasts her pseudo ID ID_{PA}, and random R_{RA} within her neighbor region. One of her neighbor, let's say Bob, makes a response with his pseudo ID ID_{PB}, and generated random R_{RB} and verification code Ver_0 as shown in Figure 1. If Alice is a valid node then $Ver_0^* = Ver_0$, and $Ver_1^* = Ver_1$ holds, thus she can be a member and she is identified as ID_{PA}, within her neighbor. Alice and Bob use their session key $K_{AB} = g(SP_{PA}, ID_{PB}) = g(ID_{PA}, ID_{PB})^{\omega}$ and $K_{BA} = g(SP_{PB}, ID_{PA}) = g(ID_{PB}, ID_{PA})^{\omega}$; thus $K_{AB} = K_{BA}$ corresponding their pseudo IDs, ID_{PA} and ID_{PB} respectively. No one within Alices neighbor can recognize her as Alice because she is using her pseudo ID and she is changing her pseudo ID time to time. Thus the nodes can hide their IDs in the network and always seem new to each other. Any adversary node can not be a member within its neighbor, because it has to pass the verification process "$?(Ver_1^* = Ver_1)$"

which is not possible to generate without the knowledge of the system secret. Similar way all nodes in the network can authenticate anonymously within their neighbors and generate their corresponding session key. Thus nodes in the network maintain their neighbor table with their pseudo IDs and corresponding session key.

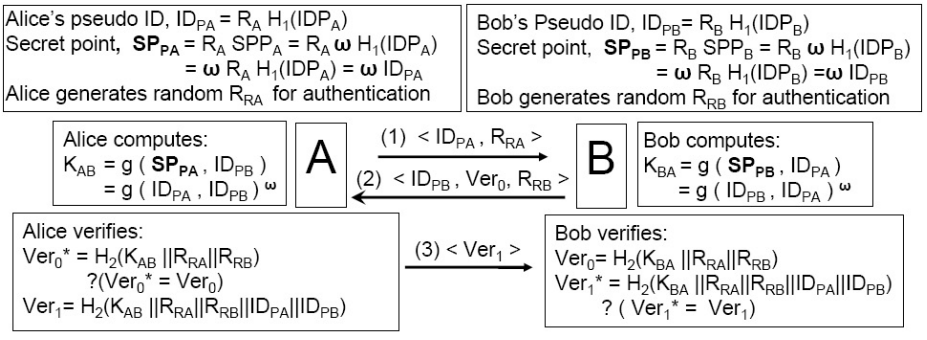


Fig. 1. Anonymous neighbor authentication process for two neighbor nodes Alice and Bob

4.2 Control Packets

RIMIO uses route request packet RRQ, and route reply packet RRP, to find a route in the network. To discover a route and to receive a response it uses RRQ and RRP respectively.

Route Requeat Packet *RRQ*:

ID_{PSE}	$RRQSeqNO$	ID_S	ID_D

ID_{PSE}: Sender pseudo ID ID_{PSE}, it is the pseudo ID of the current sender. When sender broadcasts a RRQ packet it puts its own pseudo ID in this field. Thus $ID_{PSE} \neq ID_S$, but when the source is a sender then $ID_{PSE} = ID_{PSO} \neq ID_S$, here ID_S is the source's real ID and ID_{PSO} is the source's pseudo ID which we discussed in section 3.

RRQSeqNO: Route request sequence number is used for identifying each route-request and corresponding route-reply packet from each other. It is generated by the source uniquely when source wants to communicate with a destination. $RRSeqNO = H(ID_{PSO}||Time)$, where, H is a collision resistant hash function known to all non adversary nodes in the network, ID_{PSO} is a pseudo ID of the source, and Time is the calendar time when source generates RRQ packet. This field remains unchanged for the corresponding RRP generated by the destination.

ID_S: Source's ID ID_S, it is the source's real ID. Source generates a route request packet and puts its real ID in this field, and pseudo ID ID_{PSO}, in ID_{PSE} field; thus for source $ID_{PSE} = ID_{PSO}$ but $ID_{PSE} \neq ID_S$. It is used by the destination to make a sign in route reply packet.

ID_D: Destination's ID ID_D, it is the destination's real ID.
Route Reply Packet RRP:

ID_{PSE}	ID_{PRE}	$RRQSeqNO$	$Sign_D$

ID_{PRE} : Receiver's pseudo ID; on the path from the destination to the source when RRP packet travels ID_{RPE} defines the next node who receives RRP packet.

$SignD$: Destination's Sign; when destination replies to source through intermediate nodes, it generates a sign, so that no one can forge. $Sign_D$ =-$H_2(K_{DS}||RRQSeqNO)$, where K_{DS} is a session key between the source and the destination, and generated by the destination as $K_{DS} = g(\omega H_1(ID_D)$,-$H_1(ID_S)) = g(H_1(ID_D), H_1(ID_S))^{\omega}$.

Destination also uses its session key K_{DS}, to decrypt data, which sent by the source encrypted with source's session key K_{SD}, where $K_{SD} = g(\omega H_1(ID_S)$,-$H_1(ID_D)) = g(H_1(ID_S), H_1(ID_D))^{\omega}$.

4.3 Route Discovery and Route Reply

On route discovery and route response procedures nodes maintain their corresponding tables. When a node receives a RRQ packet it broadcasts within its neighbor and when it receives a RRP packet, it sends the RRP corresponding to the receiver. RIOMO is described in terms of its functionalities which are described below.

Route Discovery. Every node in the network maintains its neighbor table with their pseudo IDs and corresponding session keys. When a source wants to communicate with a destination it generates a RRQ and broadcasts this RRQ within its neighbor to find a route, thus RIOMO is an on-demand routing protocol. By receiving a RRQ, a node checks ID_D and $RRQSeqNO$, of the RRQ and makes the following decisions:

- If the node is the destination i.e., ID_D matches with its real ID then it do the following tasks:
  - It keeps $< RRQSeqNO, ID_{PSE} >$ in its routing table; this ID_{PSE} becomes ID_{PRE} for RRP, generated by the destination. By replacing destination's own pseudo ID in the ID_{PSE} field of RRQ, it broadcasts RRQ, within its neighbor. The purpose of this extra broadcast is to make attackers fool.
  - It generates a RRP with its own pseudo ID ID_{PSE}, receiver's pseudo ID ID_{PRE} already discussed above, makes a sign $Sign_D$ discussed in section 4.2 and sends to the receiver. Notice that $RRQSeqNO$ is unchanged.
- If the node is not the destination and $RRQSeqNO$is new, it keeps $RRQSeqNO$, corresponding pseudo ID ID_{PSE} in its routing table, this information $< RRQSeqNO, ID_{PSE} >$ is used by the node in the route reply procedure; this ID_{PSE} becomes a receiver pseudo ID ID_{PRE} in the route reply procedure. The node becomes a new sender and it puts its own pseudo ID in the ID_{PSE} field of the RRQ and this RRQ within its region.

Route Reply. It is just a reverse path traverse of a *RRP* explored by a *RRQ*. When a *RRQ* reaches to the destination it generates a *RRP* and forwards it in the reverse path as we discussed above. If a node receives a *RRP*, it checks *RRQSeqNO* in its routing table then updates receiver's pseudo ID ID_{PRE}, with an appropriate ID_{PSE} (i.e., from whom it receives the corresponding *RRQ* with the same *RRQSeqNO*), and sends in the reverse path. If source receives a *RRP* it generates $Sign_S = H_2(K_{SD}||RRQSeqNO)$ and verify $Sign_D$. If $Sign_S = Sign_D$ the source sends data in the explored path by encrypting with its session key K_{SD}.

4.4 Working Procedure in Brief

1. Nodes make authentication of their neighbor nodes and maintain their neighbor table. Thus only the trusted nodes can take part in authentication.
2. On Route discovery phase, source generates a *RRQ* and sends within its neighbor. If the destination is not within its neighbor then neighbor nodes become new sender. By replacing their own pseudo IDs broadcast within their own neighbor region. They also maintain this information in routing table as we discussed in section 4.3.
3. If the node is the destination it generates a *RRP* and sends in the reverse path as we discussed in section 4.3
4. Receiving *RRP*, source checks the authenticity of the destination, by comparing $Sign_S$ and $Sign_D$. If success then sends data in the explored path. Source and destination will use their corresponding session key for encryption and decryption as discussed in section 4.2 and 4.3.

5 Anonymity Achievement and Security Analysis

When an *RRQ* and *RRP* travel from node to, every node generates a large bit random sequence corresponding to the fields of *RRQ* and *RRP*. By extracting random bits from the fields of the packets, every node pads their own random bit sequence, and replaces their own pseudo IDs to the ID_{PSE} accordingly. Thus the packets appear new when it moves from node to node. Also the fields (except ID_{PSE}, ID_{PRE}) are encrypted with corresponding session keys, thus it is also protected from intruders.

Identity Privacy. In RIOMO the identities of the nodes are represented by their pseudo IDs which are changed by the nodes in each session of communication. Pseudo IDs are also generated by using random numbers, hash functions as we discussed in section 3, also the control packets are encrypted so no one can recognize who is actual source and/or destination in a route request, route reply phase. Thus identity privacy of nodes is achieved in the network.

Location Privacy. If there is extra information added to control packets when the packets are forwarded form node to node; by observing the route request and the route response packets an attacker can estimation about the distance between the source and the destination. Thus, an attacker can set an attack regarding location privacy.

In our scheme, nodes do not know anything about the locations and identities of the other nodes in the network. So, no nodes in the network can determine the distance from them to the source and to the destination; they also do not know about the starting point of a packet traveling in the network. Only in a session the nodes know pseudo IDs of its neighbor region. Thus RIOMO ensures location privacy.

Route Anonymity. Current attacks on route anonymity are based on traffic analysis [11]. The general theory behind these kinds' of attacks is to trace or to find a path in which packets are moving. For these purpose the malicious nodes mainly looks for common information which are not changing in a packet during movements of control packets. As a result, the adversaries can find or to estimate the route from source to the destination. In RIOMO all the control packets appear new to the network, when it travels form node to node. Because every time random bits are extracted and padded during movements of the control packets as we discussed at the beginning of this section. Thus route anonymity is achieved of a path.

DoS. According to the target of attack, multiple adversaries can co-operate or one adversary with enough power can target to a specific node to exhaust the resource of the node. For this purpose the adversaries try to identify a node and set a target to that specific node. In RIOMO identity privacy is achieved; so one can identify a node make a target to attack. Thus DoS can be protected.

Wormhole Attack. In wormhole attack an attacker records a packet in one location of the network and sends it to another location making a tunnel [12] between the attacker's nodes, later packet is retransmitted to the network under its control. Thus there could be a long distance travel for a packet to find a route from the source to the destination. In RIOMO an attacker can not be a trusted member within its neighbor so it can not be an intermediate node in route discovery or route reply phase thus an attacker can not take part in the routing. So the affect of the wormhole attack is not effective in RIOMO.

Rushing Attack. By using the tunnel of wormhole attack an attacker can introduce rushing attack to rush packets. Existing almost all on-demand routing protocols suffers from rushing attack. As RIOMO can prevent wormhole attack so rushing attack is not effective in this protocol also.

6 Conclusions and Future Works

Anonymity is one of the important characteristics in securing a mobile ad-hoc network routing. In this paper an anonymous on-demand routing protocol, called RIOMO, is proposed, for preventing active as well as passive attacks. Nodes in RIOMO take only one pseudo ID from system administrator and generate their own pseudo IDs for anonymous communications. Thus pseudo IDs maintenance cost is reduced compare to the existing protocol. Moreover RIOMO ensures node privacy, route anonymity and location privacy and is robust against

several known attacks. Comparison analysis and security properties are described. Further research is to consider performance analysis as well as implementation in a specific environment.

References

1. Y. Guan, X. Fu, D. Xuan, P. Shenoy, R. Bettati, and W. Zhao, NetCamo: Camouflaging Network Traffic for QoS-Guaranteed Mission Critical Applications, *IEEE Transactions on Systems, Man, and Cybernetics*, 31(4), July 2001, pp.253-265.
2. ISO99 ISO IS 15408, 1999, available at `http://www.commoncriteria.org`
3. Y. Zhang, W.Liu and W.Lou, Anonymous Communications in Mobile Ad Hoc Networks, In *IEEE Infocom 2005*, Miami, USA, March 13-17, 2005. The 24th Annual Conference Sponsored by IEEE Communications Society, available at `http://ece.wpi.edu/~wjlou/publication/INFOCOM05_Zhang.pdf`
4. K. El-Khatib, L. Korba, R. Song, and G. Yee, Secure dynamic distributed routing algorithm for ad hoc wireless networks, In *International Conference on Parallel Processing Workshops (ICPPW03)*, 2003.
5. M. G. Reed, P. F. Syverson, and D. M. Goldschlag, Anonymous connections and onion routing, *IEEE Journal on Selected Areas in Communications, Special Issue on Copyright and Privacy Protection*, 16(4), 1998, pp. 482494.
6. J. Kong and X. Hong, ANODR: ANonymous on demand routing with untraceable routes for mobile ad-hoc networks, In *Fourth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc03)*, 2003, pp. 291302.
7. B.N. Levine, C.Shields, Hordes:a multicast based protocol for anonymity,*Journal of Computer Security* Volume 10, Issue 3, ISSN:0926-227X, 2002, pp. 213-240.
8. D. Boneh, M. Franklin, Identity Based Encryption from the Weil Pairing, *SIAM Computing*, Vol. 32, No. 3, Extended Abstract in Crypto 2001, 2003, pp. 586-615.
9. P. S. L. M. Berreto, H. Y. Kim and M. Scott, Efficient algorithms for pairing-based cryptosystems, *Advances in Cryptology - Crypto2002*, LNCS 2442, Springer-Verlag (2002), pp.354-368.
10. D. Boneh, B. Lynn and H. Shachum, Short signatures from the Weil pairing, *Advances in cryptology ASIACRYPT01*, Lecture Notes in Comput Sci. 2248 (2001), pp.514-532.
11. J.F.Raymond,, Traffic Analysis: Protocols, Attacks, Design Issues and Open Problems, in *Proceedings of PET 01*, Vol. 2009, LNCS, Springer-Verlag, 2001, pp. 10-29.
12. Y.C. Hu, A. Perrig, and D. B. Johnson, Packet leashes: A defense against wormhole attacks in wireless ad hoc networks, In *Proceedings of the Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003)*, 2003.

A DiffServ Management Scheme Considering the Buffer Traffic Rate in Ubiquitous Convergence Network*

Hyojun Lee[1], Mintaig Kim[1], and Byung-Gi Kim[2]

[1] Mobile Telecommunication Research Division, Electronics and Telecommunications Research Institute (ETRI), 161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Korea
{hjlee04, mtkim}@etri.re.kr
[2] School of Computing, Soongsil Univ. 1-1 Sangdo-dong, Dongjak-gu, Seoul 156-743, Korea
bgkim@ssu.ac.kr

Abstract. Ubiquitous convergence network consists of access networks to provide heterogeneous network services for mobile users. In this paper, we propose a hierarchical policy-based architecture model and the policy procedures based on access networks connected with Ubiquitous convergence networks. We also present a dynamic traffic management scheme which uses DiffServ mechanism and SLA for the management of end-to-end QoS in access networks. At the end of this paper, we will analyze the performance of the proposed schemes through computer simulation.

1 Introduction

Ubiquitous convergence network means a linked network system that is able to use not only each mobile access network's services but also other heterogeneous network services by structuring a unified convergence network [1], [2]. Through these convergence networks, mobile users can use a variety of heterogeneous network services as well as existing network services.

The research for ubiquitous convergence network is currently being progressed actively, but it is still at the beginning stage in which network connection models in some parts are just suggested. The noticeable point of the QoS guarantee for the transmission services is that IP transmission techniques are used for network services [13]. To provide stable and reliable support for ubiquitous networks, therefore, end-to-end QoS mechanisms should be supported [3]. By applying policy-based structure, the research for the convergence network construction of heterogeneous network is made progress and three kinds of convergence network models are suggested. However, the suggested models don't place priority to the QoS management structure through policy agreement procedures among heterogeneous access networks. They also don't consider any scheme for traffic control in network [8].

* This work was supported by the Korea Research Foundation Grant (KRF-2004-005-D00147).

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 150–159, 2007.

Two approaches, IntServ and DiffServ, have been proposed at IETF (Internet Engineering Task Force) to provide end-to-end QoS support for existing IP traffic services [5]. On the other hand, as a service support method for end-to-end QoS guarantee, IETF suggested SLA (Service Level Agreement) method, which receives IP services through the agreement between service providers and service users. If SLA is entered into an agreement between service providers and service users, then SLS (Service Level Specification) is decided to provide SLA based services [9]. SLA is a descriptive parameter related to end-to-end QoS support for network services, therefore, a user can get a corresponding service based on the parameter. CADENUS and TEQUILA projects in Europe have actually applied the method that a user can use a service through SLA in a large IP network [4], [6]. Consequently, SLA method should be considered for IP-based service support in ubiquitous network, which is a heterogeneous network set.

In this paper, we propose a policy-based DiffServ QoS management structure referring PBMN (Policy-based Management Network), which is hierarchically structured to construct heterogeneous access network convergence networks, and SLA, which is to control end-to-end QoS. Hierarchical PBMN means that PDP (Policy Decision Point) in core network controls lower PDPs by performing a role of controller through the communication with PDP in each access networks to connect with other heterogeneous access networks [10]. For this, we suggest total ubiquitous network construction and control procedures. Policy-based DiffServ QoS management structure referring SLA is an approach that manages traffic classes by the policy definition information received from PDP of each access network. It also distinguishes dynamically traffic classes by using the policy decision information of PDP, and refers a method that controls traffic classes by setting two critical values of output buffers. If the critical value of each stage is exceeded, the entrance into the output buffer for particular traffic class is limited, so the overload is reduced. Therefore, the services of higher traffic classes are guaranteed.

This paper consists of 5 sections. Section 2 describes the proposed policy-based DiffServ QoS management structure considering SLA and the communication model. Section 3 shows the policy-based DiffServ QoS control mechanism. Section 4 explains the simulation environment and the performance evaluation of the proposed method, and Section 5 describes our conclusion.

2 Policy-Based DiffServ QoS Management Structure Considering SLA

For end-to-end QoS support to heterogeneous access network and for effective traffic control between heterogeneous networks, we suggest DiffServ QoS management scheme considering hierarchical policy-based QoS management scheme, resource status of each access network, and SLA of subscribers [5]. Fig. 1 shows the layout and components of proposed policy-based DiffServ QoS management structure considering SLA.

As you can see in Fig. 1, core network is the center of the structure, and access networks are linked together through GER (Global Edge Router). LER (Local Edge Router) links a domain of access network to another. Core network and each access network consist of PDP (Policy Decision Point), PEP (Policy Enforcement Point), and PR (Policy Repository) to provide policy-based QoS management [10], [7]. PDP decides the operation policy of each network by approaching PR of each network, and transfers it to PEP so that traffics can be controlled by the decided policy. PDP collects network resource status and information, which is necessary for policy decision, analyzes the collected information and policy information of PR, decides the execution, and performs the policy control. Information transmission for the policy control between PDP and PEP uses COPS (Common Open Policy Service) proposed at IETF [11]. COPS, a TCP/IP-based request/reply protocol, is designed to support a variety of clients without protocol change, and provides message-dimensioned secure for authentication and message integrity.

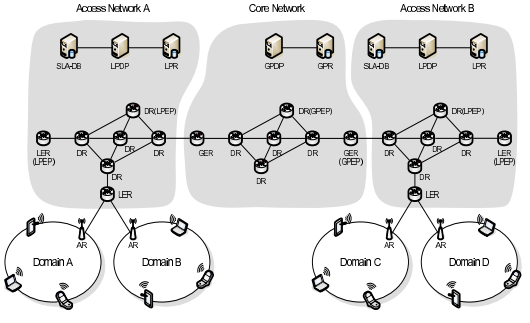


Fig. 1. Policy-based DiffServ QoS management structure considering SLA

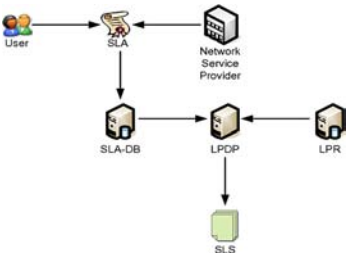


Fig. 2. The relation between SLA and SLS in the proposed structure

SLA-DB is a database system that stores SLA between access network providers and subscribers and performs information storage functions such as detail information for service usage between network providers and subscribers, authentication of subscribers, or service charges. If a access network user asks connection service or particular service, PDP deduces user-level SLS considering network resource status and the subscriber's SLA by approaching SLA-DB and PR, and transmits it to the subscriber to create traffic based on SLS limited by PDP. Fig. 2 describes the relation between SLS, which is assigned between access network providers and users, and SLS [9].

SLS is a detail parameter used for end-to-end support of subscribers, and its priority can be different with the parameters used in each heterogeneous access network. Therefore, to communicate between heterogeneous access networks, a function which can changes SLS parameter to be suitable for each access network is necessary. This function is performed by SLST (SLS Translator) included in GPDP of core network. GER of core network is a gateway for traffic entering to

core network, and performs its functions as an interface between core network and each access network. GER is an edge router of DiffServ, and includes packet tunneling and header change function for communication between heterogeneous networks. Edge routers of core network and each access network classifies traffics into traffic class based on policy decision information which is received from their PDPs, and then the routers perform DSCP field marking of IP header. On the other hand, DR (DiffServ Router) is designed to refer policy decision information and DSCP field and to perform PHB (Per Hop Behavior) for the traffic control by DiffServ mechanism.

In this paper, we propose the method that decides SLS based on SLA of LPDP and transmits service traffics for access network users. For this, mobile terminals should perform subscriber authentication by communication with LPDP if connection is requested, and receive SLS that is available based on SLA. If LPDP receives connection request from a mobile terminal, LPDP requests SLA information of the subscriber. LPDP also deduce SLS for the mobile user by using the resource status of Resource Manager and the policy information which is currently being applied. SLS is transferred to the mobile terminal. If SLS of a mobile user should be changed depending on the resource status or policy change, new SLS is created by the SLS deduction procedure and is sent to the mobile terminal. (a) of Fig. 3 represents message transmission procedure depending on the connection request of a mobile terminal, and (b) shows the message transmission procedure depending on the resource or policy change.

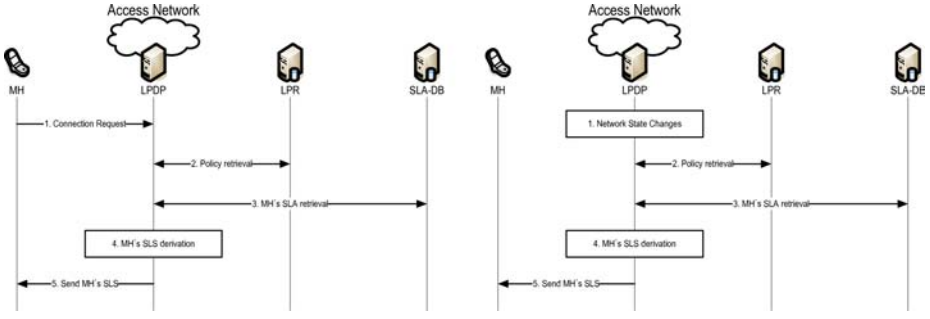


Fig. 3. SLS control procedure of mobile terminals

In the structure proposed by this paper, PEPs of Core network and access networks don't store the status information of mobile terminals. It's an advantage because PEP's overhead for traffic control of mobile terminals can be reduced and mobile user traffic can be controlled by its policy. The core network and each access networks include PDP, PEP, and PR. That is, each network has policy and all rules for policy decision making which are necessary for operation and management of network services, and it means that independent operation at

each network can be guaranteed without consideration of other network's situation. Additionally, end-to-end QoS is guaranteed and the procedures necessary for policy information exchange or negotiation are able to be simplified when core network's GPDP performs its role as a medium in a heterogeneous access network communication, so traffic is created by considering the characteristics and status of corresponding access network. Therefore, our proposed structure in this paper can guarantee the extendibility and the independence of each access network in heterogeneous convergence networks.

3 Policy-Based DiffServ QoS Control Scheme

In this section, we describe the function of ER and DR, which are applied policy-based DiffServ QoS control scheme proposed in this paper to manage dynamic QoS, and the structure for traffic control. We suggest a dynamic traffic control scheme through the traffic reset method or scheduling weight adjustment by network resource management policy.

If a traffic transmitted from a mobile terminal has arrived to ER, then ER classifies the traffic depending on the policy decision information received from PDP, and transmits the information to DR. DR constructs output buffer using received traffic by CBWFQ (Class Based Weighted Fair Queue) and PQ (Priority Queue) depending on the network resource management scheme, and performs PHB [12].

Fig. 4 shows the detail structure of ER in policy-based DiffServ QoS management structure proposed in this paper. ER consists of four components; Classifier, Meter, Marker, and Policy Controller.

Policy controller stores policy decision information received from PDP, and creates Filtering Rule, Traffic Profile, and Marking Rule for function accomplishment of Classifier, Meter, and Marker, and also creates threshold value for output buffer management. Table 1 describes parameters created by Policy Controller in ER for each component in detail.

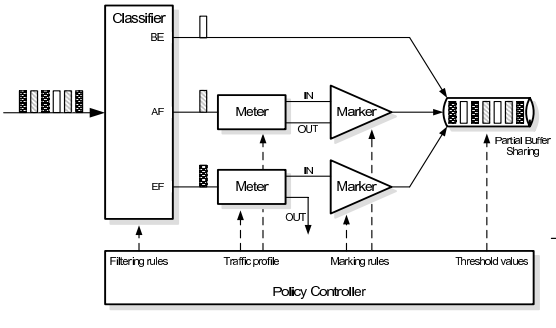


Fig. 4. The structure of policy-based ER

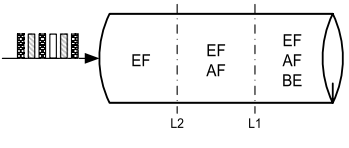


Fig. 5. Partial buffer sharing

Table 1. Detail functions of Policy Controller in ER

Parameter	Detail content
Filtering Rule	Criteria that classifies traffics arrived to ER into traffic classes
Traffic Profile	Characteristics of traffic allowed in each traffic class
Marking Rule	Marking criteria using the metering result
Threshold Value	L1, L2 Weights for construction of output buffer

3.1 The Function and Structure of ER

Classifier applies Filtering Rule of Policy Controller to classify traffics into three traffic classes, EF, AF, and BE, and delivers the information to Meter. Traffics classified into each traffic class are measured by traffic profile measurement. After that, Marker performs DSCP marking by applying Marking Rule depending on the metering result.

If traffics classified into EF class by Classifier satisfy the traffic profile, then Marker performs marking. Otherwise Marker abandons corresponding traffic. If traffics classified into AF class can't satisfy the traffic profile, Marker doesn't perform marking, and change the traffic to BE traffic class. If the traffic profile is satisfied, Marker divides the metering result into AF1, AF2, AF3, and AF4 by Marking Rule. In case of BE class, the control of metering or marking is not performed.

Output buffer uses PSE (Partial Buffer Sharing) method to provide services dynamically by referring traffic priority and overload classified by DiffServ mechanism. Output buffer applies L1, L2 weights provided by Policy Controller. If it is over the weights, particular traffic class is only queued in output buffer, and other traffic classes are abandoned. As you can see in Fig. 5, all traffic classes are allowed to be queued in buffer before the L1 weight, but EF and AF classes are only queued and other traffic classes are all abandoned if it is over of L1 weight. If it is over beyond the L2 weight, only EF class is queued, and other traffic classes are all abandoned.

As we apply PSB method, the overload of ER and DR by the continuous increase of traffics transmitted from mobile terminals can be reduced, and the service rate of higher-level traffic class can be guaranteed by providing traffic services depending on the priority of traffic classes. If lower-level traffic class is continuously abandoned because of the continuous overcrowding from mobile terminals, ER requests policy control to PDP to manage QoS dynamically. Then PDP classifies lower-level traffic class as higher-level by changing Filtering Rule and Traffic Profile.

3.2 The Function and Structure of DiffServ Router

Fig. 6 describes the structure of DiffServ Router. DiffServ Router consists of four components; Classifier, CBWFQ, PQ, and Policy Controller. Policy Controller accomplishes the function that saves policy decision information received from

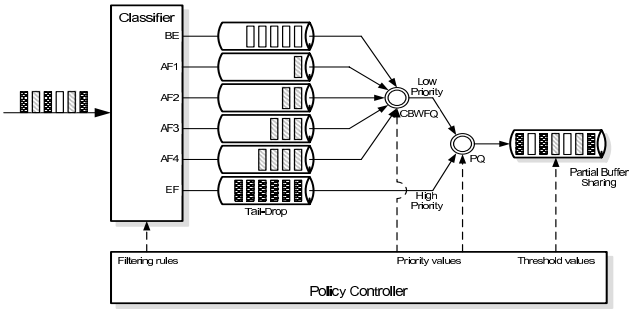


Fig. 6. The structure of polish-based DR

PDP, and creates Filtering Rule for function accomplish of Classifier, CBWFQ, and PQ components, Priority Values, and Threshold Values for output buffer management. The parameters created for each component by Policy Controller are shown in Table 2.

Table 2. Detail functions performed by Policy Controller of DR

Parameter	Detail function
Filtering Rule	Criteria queuing up traffics arrived at DR in traffic class buffer
Priority Value	Priority information applied at CBWFQ and PQ
Threshold Value	L1, L2 weights for output buffer construction

Classifier refers DSCP field of traffics and then delivers it to each class buffer. At that time, Filtering Rule of Policy Controller makes it possible to deliver the information to higher- or lower-level traffic class buffer without DSCP field change for particular traffic class. It's for dynamic traffic control in case that some traffics passing through particular area have a bottle-neck syndrome, or service rate of higher-level class should be increased.

Traffics delivered to each traffic class buffer are queued up in output buffer by CBWFQ and PQ based on the priority information provided from Policy Controller. By applying CBWFQ and PQ in the proposed structure in this paper, traffic services can be increased by the traffic class priority of mobile terminal traffics and QoS of mobile users can be guaranteed.

EF class is designed to be scheduled by PQ without passing CBWFQ, AF and BE classes are designed to be scheduled by CBWFQ and PQ considering priority information. EF class has the lowest disposal rate, and AF has the priority levels and much lower disposal rates in this order: AF1, AF2, AF3, and AF4. BF has the lowest priority level and the highest disposal rate. If the threshold value applied PSB method used at ER is over the L1 or L2 weight, the lower-level traffic classes are abandoned. In case of continuous disposal of particular traffic class, dynamic QoS can be managed by requesting policy control to PDP.

4 Simulation and Performance Analysis

ER distinguishes traffic classes provided to mobile users into three classes; EF, AF, and BE. AF is divided into AF1, AF2, and AF3, so ER transmits these six traffic classes to DR. We produced 100 traffics per every second to evaluate the performance of proposed ER. The traffics consist of EF class 20%, AF class 60%, and BE class 20%. The packet sizes of each traffic class are supposed as EF 350KB, AF 200KB, and BE 100KB. ER can process traffics in speed of 20MB/sec, and we measured the disposal rates (%) of each traffic class for 10 minutes by 5% overcrowding the transmitted traffics per every second.

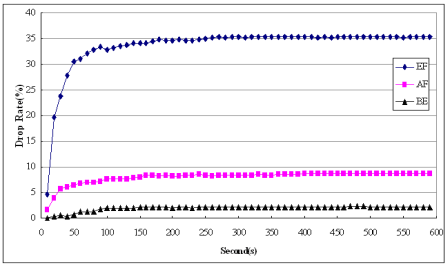


Fig. 7. Disposal rates depending on the traffic classes

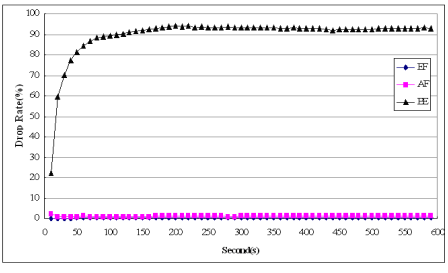
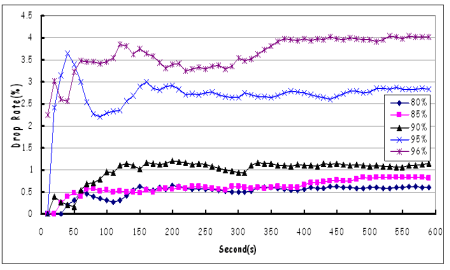


Fig. 8. Disposal rates per traffic classes in the case that L1 weight is set as 80%

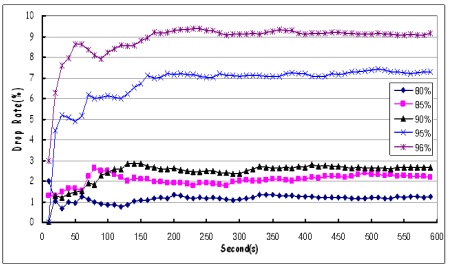
Fig. 7 describes the disposal rates of each traffic class in case of traffic overcrowding if the proposed structure is not applied. We can see that about 3% traffics of BC class is abandoned, but about 35% traffics of EF class and 9% traffics of AF class are abandoned. We analyzed and concluded that EF and AF class traffics, which have much bigger packet size than BE class's, couldn't enter to buffer because of traffic overcrowding, and increased the service lowering by being abandoned.

Fig. 8 shows the result of simulation in which the proposed structure is applied but the situation is same with Fig. 7. When we set L1 weight as 80%, the figure shows disposal rates of each traffic class. As you can see in the figure, the result shows that the disposal rate of BE class which doesn't guarantee its service rate is increased from 3% to 90%. We also can see that the disposal rates of EF and AF classes which guarantee their service rates are seriously increased from 3.5% and 9% to 0.6% and 1.2%. Through the Fig. 7 and 8, we could sure the fact that our ubiquitous convergence model and traffic control structure have effective performance for the service lowering caused by traffic overcrowding. Additionally, if policy control is requested because of continuous BE class disposal shown in Fig. 8, ER or DR can manages traffics dynamically through the weight resetting by requesting policy control to PDP.

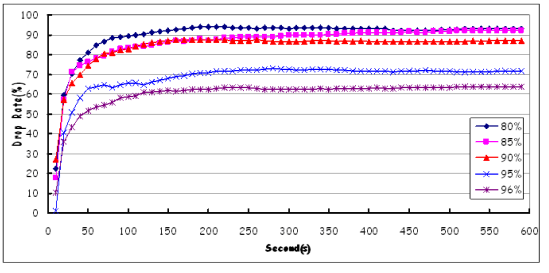
Fig. 9 shows the change of disposal rates of each traffic class when we simulated depending on L1 weight change in the proposed method in this paper. When L1



(a) EF traffic class



(b) AF traffic class



(c) BE traffic class

Fig. 9. Disposal rates of each traffic class depending on L1 weight setting

weight is set as 96%, we can see that the disposal rate of BE traffic class is kept at 60% as you can see in Fig. 9 (c), while the disposal rates of EF and AF classes are kept at 4% and 9% as you can see in Fig. 9 (a) and (b). Weight management of output buffer is an important element for traffic control, and it should be managed dynamically through continuous monitoring. Consistence rate in each class of produced traffics and average size of packet are also important elements to have influence directly on the disposal rate and service rate. These elements are provided resource status information from ER and DR, which performs the role of PEP, to PDP. PDP manages traffics dynamically by the policy decision information referring corresponding information, and provides end-to-end QoS.

5 Conclusion

Ubiquitous network is a convergence network that provides heterogeneous access network services as a form of service to mobile users of each access network. To construct this ubiquitous convergence network, a number of researches are proceeding in the world. However, it is still in the beginning stage and is for the partial work for heterogeneous access networks. This paper doesn't suggest a model for the linkage of particular access network. This paper suggests a model using hierarchical policy-based structure by generalizing access network structure which will

be linked with ubiquitous convergence network, message transmission procedures for policy control, and dynamic buffer management method for traffic control in network components.

If hierarchical policy-based structure and dynamic buffer management method proposed in this paper are applied for access networks, then we expect that network extendibility, end-to-end QoS management between heterogeneous access networks, and management independence of each access network can be guaranteed. In addition, we regard it as a proper model for ubiquitous convergence network construction by providing dynamic traffic services through policy-based DiffServ QoS control method which considers SLA, and by reducing service delay and network overload caused by traffic overcrowding.

References

1. Mario Munoz et, al.: A New Model for Service and Application Convergence in B3G/4G Networks. IEEE Wireless Communication. vol.11 no.5 (2004.10) 6–12
2. Christos Politis et, al.: Cooperative Networks for the Future Wireless World. IEEE Communications Magazine. vol.42 no.9 (2004.9) 70–79
3. Michael L. Needham and Nat Natarajan: QoS in B3G Networks – an Overview. In Proc. of ICCT'03 (2003.4) 1369–1372
4. Eleni Mykoniati et, al.: Admission Control for Providing QoS in Diffserv IP Networks: The TEQUILA Approach. IEEE Communications Magazine vol.41 no.1 (2003.1) 38–44
5. S. Blake et. al.: An Architecture for Differentiated Service. IETF RFC 2475 (1998.12)
6. Giovanni Cortese et, al.: CADENUS: Creation and Deployment of End-User Services in Premium IP Networks. IEEE Communications Magazine. vol.41 no.1 (2003.1) pp.54–60
7. Kicheon Kim: Analysis of Distributed DDQ for QoS Router. ETRI Journal. vol.28 no.1 (2006. 2) pp.31–44
8. Wei Zhuang et, al.: Policy-Based QoS Management Architecture in an Integrated UMTS and WLAN Environment. IEEE Communications Magazine. vol.41 no.11 (2003.11) pp.118–125
9. Emmanuel Marilly et, al.: Service Level Agreements: A Main Challenge for Next Generation Networks. In Proc. of IEEE ECUMN'02 (2002.4) pp.297–304
10. R. Yavakar et, al.: A Framework for Policy-based Admission Control. RFC2735 (2003.1)
11. K. Chan et, al.: COPS Usage for Policy Provisioning. RFC 3084 (2001.3)
12. J. Antonio Garcia-Macias et, al.: Quality of Service and Mobility for the Wireless Internet. Wireless Networks. vol.9 Issue 4 (2003.7) pp.341–352
13. Mahbubul Alam, Ramjee Prasad, John R. Farserotu : Quality of Service among IP-based Heterogeneous Networks. IEEE Personal Communications. vol. 8 no.6 (2001.12) pp.18–24

An Approach to Reliable and Efficient Routing Scheme for TCP Performance Enhancement in Mobile IPv6 Networks

Byungjoo Park[1], Youn-Hee Han[2], and Haniph Latchman[1]

[1] Department of Electrical and Computer Engineering,
University of Florida, Gainesville, USA
{pbj0625, latchman}@ufl.edu
[2] School of Internet-Media, Korea University of Technology and Education,
Cheonan-Si, Chungnam, 330-708, Korea
yhhan@kut.ac.kr

Abstract. In Mobile IPv6, the handover process reveals numerous problems manifested by movement detection, non-optimized time sequencing of handover procedures, latency in configuring a new care of address and binding update to a home agent (HA). These problems may cause packet loss as well as packet disruption. To mitigate such effects, Fast handover for Mobile IPv6 (FMIPv6) has been developed. FMIPv6 can reduce a packet loss using tunnel based handover mechanism which relies on L2 triggers such as transmitting a packet from a previous access router (PAR) to a new access router (NAR). However, this mechanism may result in decreasing the performance of TCP due to the out-of-sequence packets between tunneling packet from the Home Agent, PAR and directly transmitted packet from the correspondent node (CN). In this paper, we propose a new scheme called EF-MIPv6 to prevent packet reordering problem using new snoop mechanism (NS). This new scheme can prevent a sequence reordering of data packet using proposed "MSAD" controlling. Simulation results demonstrate that managing the packet sequence in our proposed scheme greatly increases the overall TCP performance in Mobile IPv6 network.

1 Introduction

Mobile IPv6 is designed to manage the movement of mobile nodes (MNs) between wireless IPv6 networks [1]. This protocol supports transparency above the IP layer, including the maintenance of TCP connections. However, TCP error control is focused on congestion losses and does not distinguish the possibility of temporary time delays due to handovers in wireless, mobile environments. Packet losses during handovers are treated as an indication of network congestion, which causes TCP to take some unnecessary congestion avoiding measures [2].

In MIPv6, to perform packet transmission continuously without disconnection of the layer 3, a mobile node maintains a home IP address (HoA) for identification

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 160–169, 2007.

and a temporal IP address for routing information. When an MN moves to a new subnet it should disconnect with the current access router and connect with the new access router; in addition, the MN should obtain a new temporal address called care-of address (CoA). Next, the MN should register the binding between its new CoA (NCoA) and HoA with it's home agent (HA) and CNs. We also need to know the handover latency, which is collectively defined as the delay incurred during movement detection, the new CoA configuration time, and the binding update time required to start internet service from the new subnet network. During a handover, the packets transmitted from an HA or CN may be lost. Recent work has been aimed at improving the handover performance of Mobile IPv6 in order to support real time and other delay sensitive traffic. a few trials have been developed to solve this problem such as Smooth handover by Route Optimization in Mobile IP [3], Fast Handover for Mobile IPv6 [4], Design and Analysis of the mobile agent preventing out of sequence packets [5], performance improvement by packet buffering [6], and reducing out-of-sequence packets using priority scheduling [7]. Also, the snoop mechanism [9] is designed to improve the performance of TCP while recovering wireless errors locally. The snoop mechanism introduces a module, called snoop agent, at the base station (BS). The agent monitors every packet that passes through the TCP connection in both directions and maintains a cache of TCP packets sent across the link that have not yet been acknowledged by the receiver. The main problem with TCP performance in networks with both wired and wireless links is that packet losses ,which occur because of bit-errors, are mistaken by the TCP sender as being due to network congestion.

We propose an efficient TCP mechanism in FMIPv6 to prevent the mis-ordering problem of packets during a handover using new snoop (NS) mechanism based enhanced fast binding update (EF-BU) message [8]. This is accomplished by adding a new reordering scheme to the base station connected to the NAR and by adding a modified TCP header format at the source device, such as an HA and CN. These functions are performed by two new proposed main routines: NS mechanism for data called *"MSAD"* and the enhanced fast biding update (EF-BU) message procedure. Some modifications that are explained in Section 3 are required to compare the sequence of data packets in the base station and to indicate the arrival of the final packet from the previous access router. The remainder of this paper is organized as follows. We will describe FMIPv6 protocol in Section 2. Section 3 introduces our proposed reordering algorithm (EF-MIPv6) to increase the performance of TCP. The performance evaluations are shown in Section 4. Finally, we present the conclusions in Section 5.

2 Fast Handover for MIPv6 (FMIPv6)

The basic operation of FMIPv6 [4] is depicted in Fig. 1. While an MN is connected to its PAR and is about to move to an NAR, FMIPv6 requires that the MN obtains a new CoA at the NAR while still connected to the PAR.

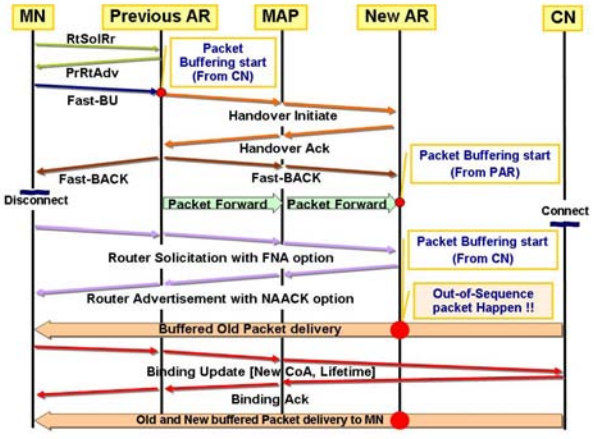


Fig. 1. IETF FMIPv6 Handover Procedure

Furthermore, the MN must send a Binding Update message to its PAR to update its binding cache with the MN's new CoA, and finally the PAR must start forwarding packets, originally destined for the MN, to the NAR.

Either the MN or the PAR may initiate the Fast Handover procedure by using the L2 trigger. The link-layer information indicates that the MN is moving from the current access point (AP) to another; that is, from the PAR to the NAR. If the L2 trigger is received at the MN (Mobile-initiated handover), the MN will initiate an L3 handover by sending a Router Solicitation for Proxy (RtSoPr) message to the PAR. On the other hand, if the L2 trigger is received at the PAR (Network-controlled handover), then the PAR will transmit a Proxy Router Advertisement (PrRtAdv) message to the appropriate MN, without any solicitation message.

The MN obtains a new CoA (NCoA) while still connected to the PAR by means of router advertisements containing network information from the NAR. The PAR validates the MN's new CoA and initiates the process of establishing a bidirectional tunnel between the PAR and the NAR by sending a Handover Initiate (HI) message to the NAR. Then, the NAR verifies that its new CoA can be used on the NAR's link. Also, in response to the HA message, the NAR sets up a host route for the MN's previous CoA (PCoA) and responds with a Handover Acknowledge (HACK) message.

When the MN receives a PrRtAdv message, it should send a Fast Binding Update (F-BU) message, preferably prior to disconnecting its link. When the PAR receives an FBU message, it must verify that the requested handover is accepted by the NAR as indicated in the HACK message status code. Then, the PAR begins forwarding packets intended for the PCoA to the NAR and sends a Fast Binding Acknowledgement (F-BACK) message to the MN. After changing link connectivity with the NAR, the MN and NAR exchange a Router Solicitation (RS)

Message including the Fast neighbor Advertisement (FNA) option and a Router Advertisement message (RA) with the Neighbor Advertisement Acknowledgment (NAACK) option. After the NAR sends a Router Advertisement message with the NAACK option, it starts to deliver buffered packets tunneled from the PAR and buffered packets from the CN directly. Until the CN receives a BU, the packets sent from the CN are tunneled from the PAR to the NAR. After the CN receives a BU, the CN directly delivers the packets to the MN. Consequently, if the distance between the CN and NAR is shorter than the tunneled distance from the CN to the NAR via the PAR, the MN may receive out-of-sequence packets. Fig. 2 shows the out-of-sequence packet problem.

After the PAR receives an F-BU message, packets four through eight are tunneled from the PAR to the NAR and buffered until the NAR receives an RS message with FNA from the MN. When the CN receives a BU from the MN, the CN sends packet nine to ten directly to the NAR. These packets are also buffered in the NAR until the NAR sends an RA with NAACK option to the MN.

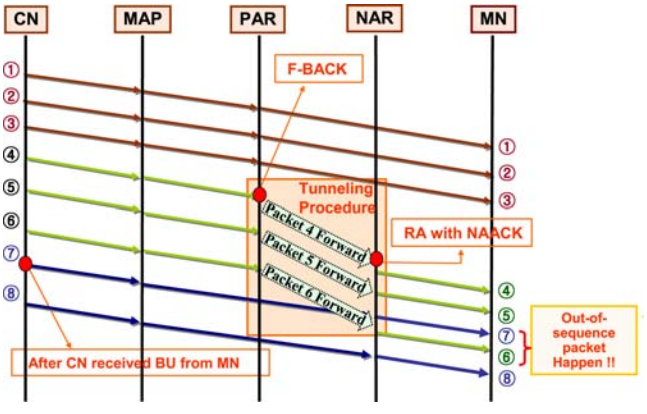


Fig. 2. Out-of-sequence packets in FMIPv6

Consequently, if the distance between the CN and NAR is shorter than the tunneled distance from the CN to the NAR via the PAR, buffered packets in the NAR would be out of sequence due to the packet delay time incurred by tunneling. So when an MN receives mis-ordered packets, the use of TCP in the MN creates a duplicate ACK (DACK) for packets seven and eight in accordance with its congestion control procedure.

3 Reliable and Enhanced Reordering Algorithm (EF-MIPv6)

In this section, we propose a new reordering algorithm to improve TCP performance in Fast Mobile IPv6 networks by eliminating out-of-sequence packets

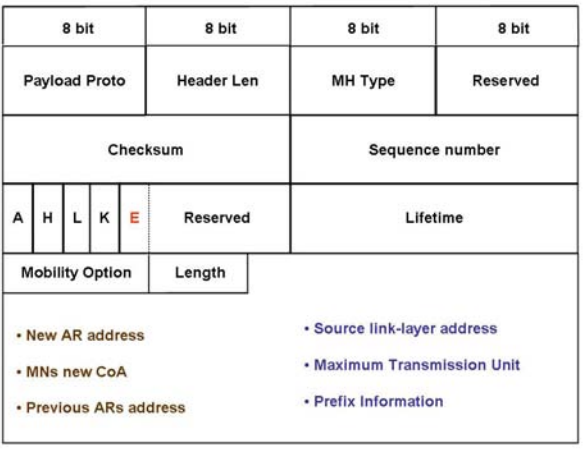


Fig. 3. Enhanced Fast Binding Update (*EF-BU*) Message

during handover. It is based on the NS mechanism which prevents DACK and controls the TCP packet data sequence in new access point (AP). The modified access point (NAP) with NS agent consists of a NAP controller, NAP buffer, and sequence checker. In movement detection, an MN is aware of performing handover to another AP because of channel maintenance or L3 handover. The MN performs a scan to see APs through probes. In proposed scheme, after process of establishing a bidirectional tunnel from the PAR to the NAR is made, the PAR sends new Enhanced Fast Binding Update (EF-BU) [8] message to the CN as soon as an MN start moving so that the number of packets which need to be forwarded from PAR to NAR is decreased.

3.1 New Enhanced Fast Binding Update Message (EF-BU)

In proposed scheme, after process of establishing a bidirectional tunnel from the PAR to the NAR is made, the PAR sends new Enhanced Fast Binding Update (*EF-BU*) message to the CN as soon as an MN start moving so that the number of packets which need to be forwarded from PAR to NAR is decreased. That is, as soon as setting up tunnel between the PAR and the NAR, the PAR send *EF-BU* quickly to the CN. This *EF-BU* message can be modified by adding a 2-bit E-flag to the reserved flag and including the "New AR address" and "MNs New CoA" as options in the option field. Fig.3 show the formats of the *EF-BU* message. Table 1 defines the E-bits. When a CN receives an *EF-BU* message the CN has to be operated by E-bits.

The description for each message exchange in proposed *EF-MIPv6* is as follows:

1) PAR sends EF-BU message after finishing tunneling-path between PAR and NAR.

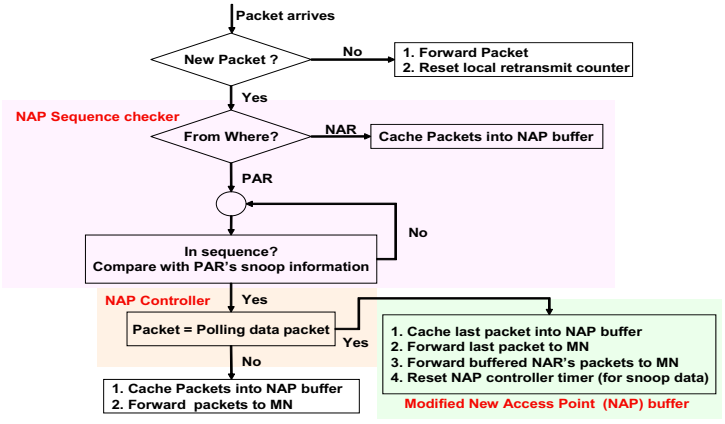


Fig. 4. MSAD Procedure in NAP

2) The CN sends EF-BUACK message to the PAR.
3) The CN send modified TCP data packet after setting MLP flag to "1".
4) The PAR send F-BACK message to the MN and NAR at the same time.
5) The PAR buffer packet addressed to previous CoA and start to forward buffered packets to the NAR.
6) The NAR start to check received TCP data packet MLP flag.
7) The MN sends router solicitation message to the NAR.
8) The NAR sends router advertisement message to the MN.
9) The CN send packets to the MN addressed to new CoA.
10) The NAR buffer packets addressed to new Co until getting the tunneled packet with MLP flag "1"
11) After receiving last tunneled packet with MLP flag "1", the NAR deliver buffered packet which came from the CN directly.

Table 1. The E-Flag of EF-BU message

E-Flag	Mean
00	Can not apply in IEEE-802 case
01	Can send data packet to the MNs new CoA
10	Must send data packet to the MNs old CoA.
10	Must use standard BU message from an MN

3.2 New Enhanced Reordering Algorithm for TCP Data with Polling Scheme

Fig. 4 shows the flowchart of the NS mechanism for data (MSAD) to perform the proposed scheme. Firstly, during a handover in Fast Mobile IPv6, when a PAR

receives an F-BU message, the PAR sends an HI message with the PAR's NS information in order to control the TCP packet sequence. As soon as receiving Hack message from NAR, the PAR send EF-BU message to CN. After the PAR sends an F-BACK message to both the NAR and the MN, buffered packets in the PAR are delivered to the NAR. When the CN receives a EF-BU message, it sends the rest of the packets to the NAR directly. Also, the CN sends the last packet with a modified TCP header to the PAR. The modified last packet is called the MLP. In order to distinguish the last packet amongst all the received packets in the NAR, the TCP packet can be modified by adding a 1 bit MLP flag to the reserved field in the TCP header. Fig. 5 shows modified TCP packet format.

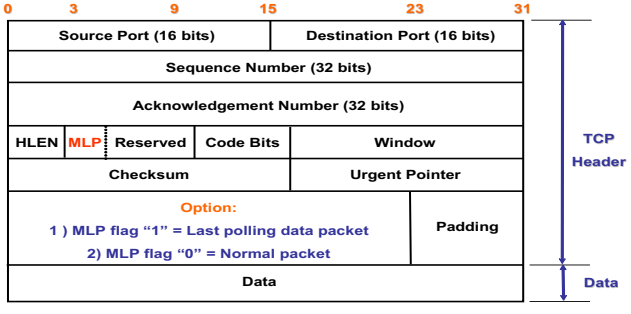


Fig. 5. Modified TCP Packet Header Format

When the MLP flag is "0", the packet acts as a normal packet. However, if the MLP flag is "1", the packet acts as a polling data packet from the PAR. That is, after the CN receives a binding update message from an MN, it simultaneously sends the last data packet and a polling data packet to the MN. The polling data packet is a control message to the MN to signal that no more tunneled packets exist. If the PAR receives this polling message, it can remove the MN's information. After sending a polling data packet, the CN can send a new data packet to the NAR without the tunneling mechanism. First, the NAP sequence checker uses the PAR's NS information, previously sent in a handover initiate (HI) message, to determine if the received packet is from the PAR or the NAR. In other words, when the PAR sends an HI message to the NAR, the HI message includes NS information about the PAR. Then, if the received packet has arrived in sequence, the MNC controller starts checking for the MLP flag.

The MLP flag is important to distinguish between packets delivered by tunneling from the PAR and packets delivered directly from the CN, without tunneling. Until the NAR receives the MLP with the flag bit set to "1", the packets delivered directly from the CN are buffered at the NAP buffer. Fig. 6 shows packet transmission during a handover in EF-MIPv6 networks.

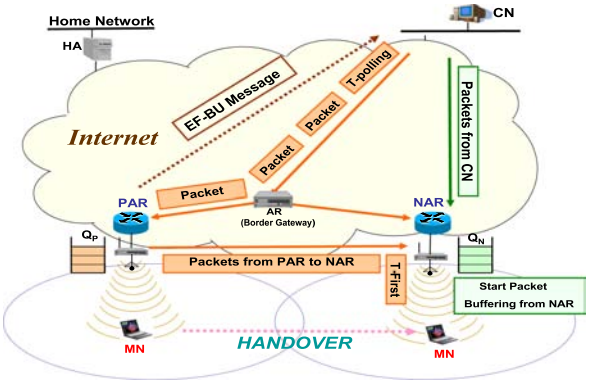


Fig. 6. The packet transmission during handover in EF-MIPv6

As shown in Fig.6, after the CN receives the EF-BU message from the PAR, the packets sent from the CN to NAR directly are cached in the NAP's buffer. Therefore, we define the data packet buffering time (T_{BT}) as the time needed to finish the packet transfer from the PAR to the NAR. T_{BT} is represented by

$$T_{BT} = |T_{First-packet} - T_{Polling-data-packet}| \quad (1)$$

Where $T_{First-packet}$ and $T_{Polling-data-packet}$ are the time when the first packet is delivered and the time when the polling data packet is delivered from the PAR to the NAR via tunneling, respectively; thus, during T_{BT}, only the packet that was received from the PAR by tunneling is forwarded to the MN. At this time, the NAP controller calculates the waiting time, T_{BT}, until the polling data packet arrives. In our paper, we assume the buffer size in the NAP is enough to cache the received packets directly from the CN during T_{BT}. After T_{BT} expires, the NAP buffer starts delivering buffered packets to the MN continuously. Moreover, to prevent packet overflow in the NAP buffer, the NAR sends control messages periodically for notifying buffer states to the CN or HA. Using these message, the CN or MN can control the data traffic. The buffer in the NAR is constructed with non priority First-In-First-Out.

4 Simulation Results

We evaluate the performance of our proposed scheme using Network Simulator (NS-2). Based on the standard NS-2 distribution version ns-allinone2.1b6, the simulation code used for the experiments was designed on top of the INRIA/Motorola MIPv6 code. We have extended the code with two main modules: a reordering algorithm for data and enhanced fast binding update message procedure. Some modifications have been done to the original release in order to extend the code to work with more than one mobile node. In TCP code, we used TCP-NEWRENO. Bulk data transfer (by FTP) is connected between the

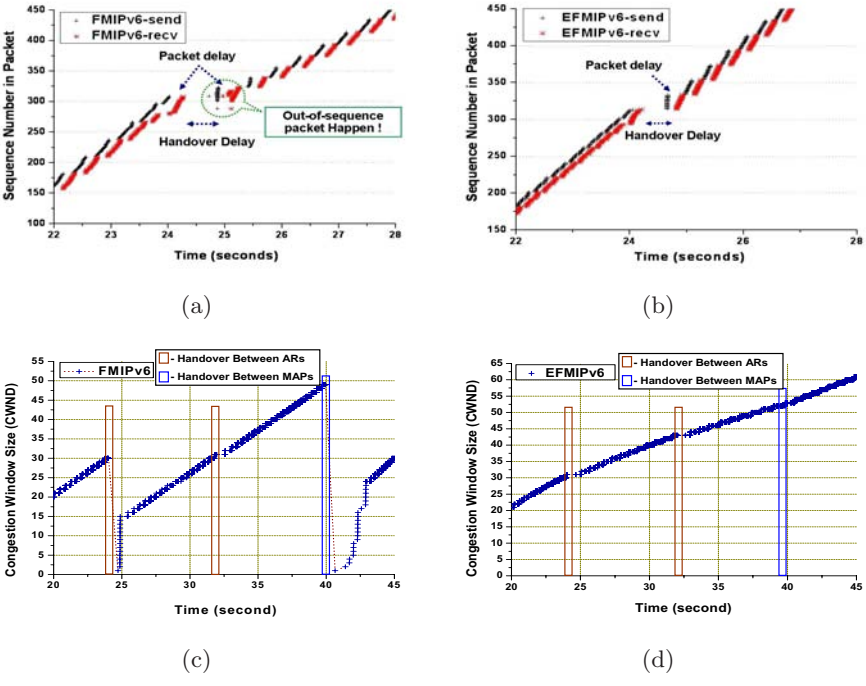


Fig. 7. The effect of handover between ARs: (a) FMIPv6 and (b) EF-MIPv6. TCP congestion window size in packet between ARs after handover: (c) FMIPv6 and (d) EF-MIPv6.

CN and the MN. The throughput of TCP is measured by the sequence number of packets successfully received by the MN. In our simulation, the buffer size is predetermined to be sufficient to cache the received packets directly from the CN to ensure the buffer does not cause packet overflow. Fig. 7 shows the received sequence number (SN) of TCP data with respect to the simulation time in FMIPv6 and Proposed EF-MIPv6, respectively.

Fig. 7 (a) shows the packet transmission associated with buffering in the NAR and PAR during a handover in FMIPv6. Although packet loss does not happen, the received packet sequence is changed in an MN due to the out-of-sequence packet problem caused by the packets received by tunneling and the packets received directly from the CN. Consequently, the out-of-sequence packet problem results in sending a DACK to the CN leading to a drop in TCP performance. Fig. 7 (c) shows the CWND in FMIPv6 between ARs. During a handover between ARs packet loss did not happen. However, after the sender received the same ACK three times from an MN, the sender transmits a delayed packet caused by tunneling between the PAR and NAR. These packet retransmissions reduce the CWND, which causes a large amount of data packets to wait for a larger CWND.

In Fig. 7 (b) and (d), although packet delay occurred, the receiver accepts the packet normally, avoiding packet loss and the out-of-sequence packet problem so retransmission is not required. The EF-MIPv6 scheme can tolerate a slight packet delay and still deliver the packet during fast handover. Therefore, the value of the sender's CWND is maintained and the performance of TCP is improved.

5 Conclusion

This paper has introduced our proposed Fast Mobile IPv6 with reordering algorithm for handovers. We have also analyzed the impact of handovers between ARs for out-of-sequence packets under Mobile IPv6 in a fast handover environment. In this paper, we showed that EF-MIPv6 can improve TCP performance and prevent the out-of-sequence packet problem in existing Mobile IPv6 network.

Acknowledgment

This work was supported in part by the Winitec com. "NGN" Research project and in part by the University of Florida LIST Research Laboratory project of the MIPv6 Technology, 2006.

References

1. D.Johnson, C. Perkins, J. Arkko, "Mobility Support in IPv6", RFC 3775, June 2004.
2. V. Tsaotissidis, "Open Issues on TCP for Mobile Computing", Jouinal of Wireless Communication and Mobile Computing, Vol.01.2, 2002.
3. C.Perkins and D.Johnson, "Route optimization in Mobile IP", Internet Draft, IETF, Sep. 2001.
4. Koodli, R., "Fast Handovers for Mobile IPv6", RFC 4068, July 2005.
5. D. Lee, C. Oh, S. Lee, J. Park, and K. Kim, "Design and analysis of the mobile agent preventing out-of-sequence", Proc. ICOIN, Jan. 1999.
6. D. S. Eom, M. Sugano, M. Murata, and H. Miyahara, "Performance Improvement by packet buffering in mobile IP based networks", IEICE Trans. Communication, vol. E83-B, no. 11, pp.2501-2512, Nov. 2000.
7. D. Lee, G. Hwang, "Performance enhancement of Mobile IP by reducing out-of-sequence packets using priority scheduling", IEICE Trans. Communication, vol. E85-B, no. 8, pp.1442-1446, Aug.2002.
8. Byungjoo Park, H. Latchman, "An Approach to Efficient and Reliable Media Streaming Scheme", in. Proc. of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (ISBMSB), April 3-6, 2006.
9. H. Balakrishnan, S. Seshan, and R. H. Katz, "Improving reliable transport and handoff performance in cellular wireless networks,", ACM Wireless Networks, vol. 1, no. 4, pp. 469-481, Dec., 1995.

Ant Colony Optimization for Satellite Customer Assignment

S.S. Kim[1], H.J. Kim[2], V. Mani[3], and C.H. Kim[4]

[1] Department of Industrial Engineering
Kangwon National University, Chunchon 200-701, Korea
[2] CIST
Korea University, Seoul 136-701, Korea
[3] Department of Aerospace Engineering,
Indian Institute of Science, Bangalore, 560-012, India
[4] Ainsolution Co.
Seoul, Korea

Abstract. This paper considers the meta-heuristic method of ant colony optimization to the problem of assigning customers to satellite channels. It is shown in an earlier study that finding an optimal allocation of customers to satellite channels is a difficult combinatorial optimization problem and is NP-complete. Hence, we propose an ant colony system (ACS) with strategies of ranking and Max-Min ant system (MMAS) for an effective search of the best/optimal assignment of customers to satellite channels under a dynamic environment. Our simulation results show that this methodology is successful in finding an assignment of customers to satellite channels. Three strategies, ACS with only ranking, ACS with only MMAS, and ACS with both ranking and MMAS are considered. A comparison of these strategies are presented to show the performance of each strategy.

1 Introduction

An excellent discussion on bandwidth resource issues related to satellite-based communication is presented in [12]. In this study [12], the problem of satellite customer assignment (see Figure 1) is considered and an integer programming formulation is presented for the solution to this combinatorial optimization problem. It is also shown in [12] that the optimal assignment of customers to channels has real and observable costs and benefits, both in terms of dollars and customer ratings. This is basically an efficient resource utilization problem. In satellite communication, efficient resource utilization is one of the important problems that has received considerable attention [4,8]. A detailed overview of the scheduling problems that arise in satellite communication is described in [11]. The well-known generalized assignment problem (GAP) is known to be NP-complete combinatorial optimization problem and has received a lot of attention in the literature [3], and the satellite customer assignment problem has a lot of similarities with GAP. The GAP involves finding the minimum cost assignment of N jobs to M machines (agents) such that each job is exactly assigned

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 170–179, 2007.

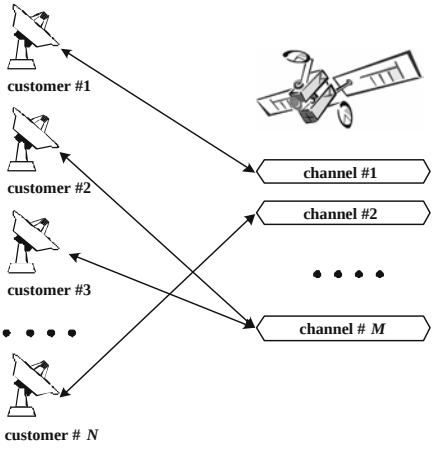


Fig. 1. Concept of customer assignment to satellite channels

to only one machine, subject to machine's available capacity. Another problem in this context well studied is flow shop scheduling [10].

The ant colony optimization paradigm was inspired by the behavior of real ants. In nature the real ants have an ability to find the shortest paths from the nest to food sources. In an ant colony, the medium that is used for information communication among individual ants regarding paths is a chemical substance called pheromone. A moving ant deposits a constant amount of pheromone on the ground (path). Another ant, when it encounters a pheromone trail, has to decide whether to follow it or not. If it follows the trail, the ant's own pheromone reinforces the existing trail. The pheromone also evaporate over time.

The well-known Traveling Salesman Problem (TSP) was the first combinatorial optimization considered for solution using Ant Colony Optimization (ACO) [6], and was published under the name Ant System (AS). In this study, the artificial ants build new solutions stochastically. For building new solutions a combination of heuristic information and an artificial pheromone trails are used by the artificial ants. This pheromone trail is reinforced according to the quality of solutions built by the ants. The AS was able to find optimal solutions for some smaller TSP problems. This study has generated a lot of interest among researchers and this AS has been applied to a variety of combinatorial optimization problems [1,7]. A detailed description of ant behavior relating to ACO is available in [5,14].

We propose the meta-heuristic method of ant colony optimization to the problem of assigning customers to satellite channels. We present ant colony system (ACS) with strategies of ranking and Max-Min ant system (MMAS) for an effective search of the best/optimal assignment of customers to satellite channels under a dynamic environment. Our simulation results show that this methodology is successful in finding an assignment of customers to satellite channels. Three strategies, ACS with only ranking, ACS with only MMAS, and ACS with

both ranking and MMAS are considered. A comparison of these strategies are presented to show the performance of each strategy.

2 Satellite Customer Assignment Problem

Problem Formulation: For ease of understanding, we follow the same notation used in an earlier study [12]. Let there be I customers to be assigned to one of the K channels. As in [12], we assume the following data is available:

- SBW_k : satellite bandwidth available in channel k
- SP_k : satellite power available in channel k
- CBW_i : bandwidth required by customer i
- CP_i : power required by customer i

The decision variable is x_{ik}. The decision variable x_{ik} is 1 if customer i is assigned to channel k, and $x_{ik} = 0$ otherwise. The satellite customer assignment problem is

$$Minimize \quad \sum_{k=1}^{K} \left| \frac{\sum_{i=1}^{I} CBW_i x_{ik}}{SBW_k} - \frac{\sum_{i=1}^{I} CP_i x_{ik}}{SP_k} \right|. \tag{1}$$

The constraints are:

$$\sum_{i=1}^{I} CBW_i \times x_{ik} \leq SBW_k, \quad \forall\ k = 1, 2, ..., K, \tag{2}$$

$$\sum_{i=1}^{I} CP_i \times x_{ik} \leq SP_k, \quad \forall\ k = 1, 2, ..., K, \tag{3}$$

$$\sum_{j=1}^{K} x_{ik} = 1, \quad \forall\ \ i = 1, 2, ..., I, \tag{4}$$

where x_{ik} is either 0 or 1. This objective function is used in an earlier study [12]. The objective function in Equation (1) minimizes the total deviation of fraction of bandwidth utilized from fraction of power utilized. The constraint in Inequality (2) represents that the capacity restriction of available bandwidth, and the constraint in Inequality (3) takes into account the capacity restrictions of available power. Constraint in Inequality (4) ensures that each customer is assigned to only one channel. From the above, we see that, this satellite customer assignment problem is similar to the generalized assignment problem and hence it is NP-complete. We can also see that the satellite customer assignment problem has a close relationship with bin-packing problem (BPP) which is a generalized assignment problem. BPP has been studied in the framework of ant

colony optimization in [9]. Consider an example with 5 customers (I=5) and 3 channels (K=3). Let one solution be {3 1 3 2 1}. The meaning of this is:

Customers 1 and 3 are assigned to channel 3.
Customers 2 and 5 are assigned to channel 1.
Customer 4 is assigned to channel 2.

This solution {3 1 3 2 1} is a valid solution only if the constraints (2) and (3) are satisfied.

3 Ant Colony Optimization

In this section, we explain the methodology of ant colony optimization to our satellite customer assignment problem. Ant systems uses artificial ants to construct a solution from the scratch. A solution is constructed based on a combination of heuristic information and artificial pheromone trails are used by the artificial ants. At each step an individual ant assigns an unassigned customer i to a channel k with a probability p_{ik}. The probability p_{ik} is given by

$$p_{ik} = \frac{(\tau_{ik})^{\alpha} * (\eta_{ik})^{\beta}}{\sum_{k \in K}\{(\tau_{ik})^{\alpha} * (\eta_{ik})^{\beta}\}}. \tag{5}$$

In Equation (5), K is the set of channel numbers that customer i can be assigned, α is the weighting factor of pheromone and β is the weighting factor of attractiveness. The attractiveness η_{ik} is given as

$$\eta_{ik} = \frac{1}{|U - V|}, \tag{6}$$

where

$$U = \frac{SCBW_k + CBW_i}{SBW_k},\ V = \frac{SCP_k + CP_i}{SP_k}. \tag{7}$$

It is possible that before assigning the customer i to channel k, some other customers are already assigned to the channel k. Hence, $SCBW_k$ is the sum of CBW of already assigned customers to channel k. Similarly, SCP_k is the sum of CP of already assigned customers to channel k. While obtaining the attractiveness the constraints in Inequalities (3) and (4) may not be satisfied. In that case the values of attractiveness are

$$\eta_{ik} = 0 \text{ if } U > 1 \text{ or } V > 1.$$

After the run, the ants update the pheromone information τ_{ik} as

$$\tau_{ik}(t+1) = (1 - \rho) * \tau_{ik}(t). \tag{8}$$

This update procedure uses an evaporation rate ρ in order to reduce the effect of past experience and to explore new and alternatives solutions. Here, ρ is the evaporation rate between time t and time $(t+1)$, and $0 < \rho < 1$. The local pheromone update is given by

$$\tau_{ik}(t+1) \;=\; \tau_{ik}(t) + \Delta\tau_{ik}^{j}, \tag{9}$$

where

$$\Delta\tau_{ik}^{j} = \begin{cases} Q/L_j & \text{if } M(i,k) \in S_j \\ 0 & \text{otherwise} \end{cases}$$

S_j is the set of ant movements for the j-th ant, L_j is the evaluation value of the j-th ant and Q is the pheromone update constant. $M(i,k)$ is the movement of an ant for assigning customer i to the channel k. In other words, $\Delta\tau_{ik}^{j}$ is the amount of pheromone and deposits on the assigning customer i to the channel k.

Ranking strategy: We now present the ranking pheromone update strategy [2]. In this strategy, the pheromone update is done as

$$\tau_{ik}(t+1) \;=\; \tau_{ik}(t) + \frac{(w+1-r)}{2}\Delta\tau_{ik}^{r}. \tag{10}$$

where

$$\Delta\tau_{ik}^{r} = \begin{cases} Q/L_r & \text{if } M(i,k) \in S_r \\ 0 & \text{otherwise} \end{cases}$$

The r is the r-th best ant, S_r is the set of ant movements for the r-th ant, L_r is the evaluation value of the r-th ant and w is the w-th best ants for ranking strategy. The global pheromone update is done as

$$\tau_{ik}(t+1) \;=\; \tau_{ik}(t) + \sigma \times \Delta\tau_{ik}^{*}. \tag{11}$$

where

$$\Delta\tau_{ik}^{*} = \begin{cases} Q/L_{go} & \text{if } M(i,k) \in S_{go} \\ 0 & \text{otherwise} \end{cases}$$

S_{go} is the set of ant movements for the global optimal ant, L_{go} is the evaluation value of the global optimal ant, and σ is a positive constant for weighting factor of the elitist one.

Max-Min Strategy: Max-Min strategy introduced in ant system known as MMAS is an improvement over original ant system [13]. This MMAS strategy introduces upper and lower bounds to the values of the pheromone trails. The allowed range of the pheromone trail strength is limited in the following interval:

$$\tau_{min} \leq \tau_{ij} \leq \tau_{max} \text{ for all } \tau_{ij} \tag{12}$$

A flowchart for ant colony optimization is given in Figure 2.

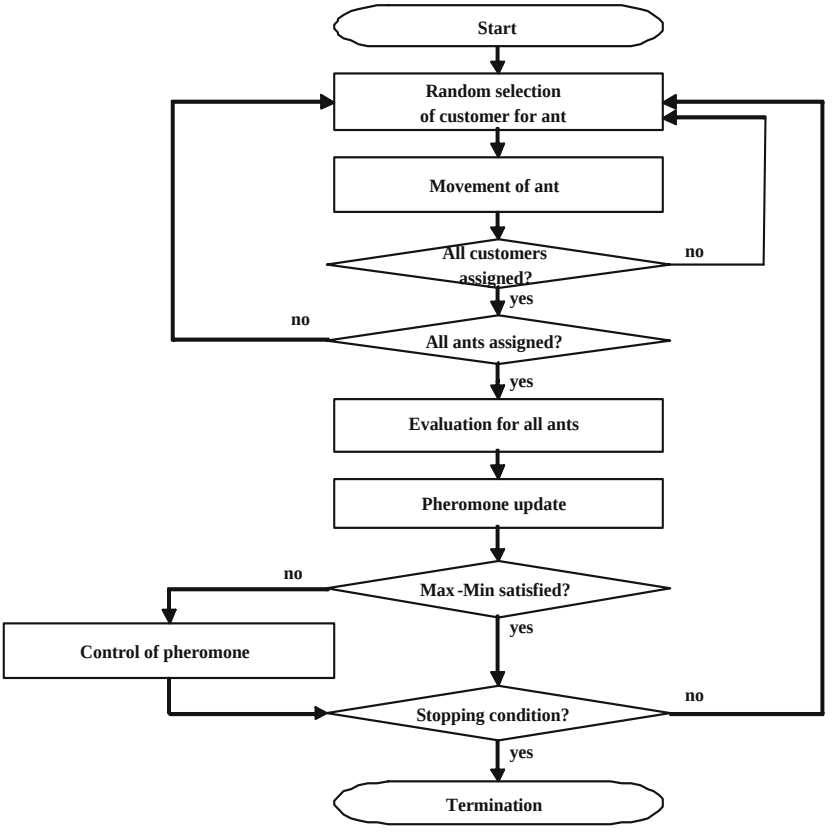


Fig. 2. Flowchart of the algorithm

4 Simulation Results

The ant colony optimization approach to search for the best assignment of customers to satellite channels is tested with some test problems. Two sets of test problems were generated. The first set of problems were generated with 5 customers and 3 channels. The second set of test problems were generated with 20 customers and 10 channels.

For the first set of problems (I=5, and K=3), the customer bandwidth requirements and power requirements are kept as: $CBW_1 = 5$, $CBW_2 = 4$, $CBW_3 = 6$, $CBW_4 = 7$, and $CBW_5 = 3$. The customer power requirements are: $CP_1 = 7$, $CP_2 = 9$, $CP_3 = 8$, $CP_4 = 6$, and $CP_5 = 5$. The number of possible solutions are 3^5. The values of satellite bandwidth available in channels (SBW) and the satellite power available (SP) used in our study are given in Table 1.

Next, we consider a 20 customer (I=20) and 10 channel (K=10) problem. This is really a tough problem because the number of possible solutions are 10^{20}. Hence, the ant colony optimization approach will be very much useful.

Table 1. Satellite bandwidth (SBW_k) and power (SP_k) for the channels

Problem	SBW_1	SBW_2	SBW_3	SP_1	SP_2	SP_3
1.1	30	35	40	40	45	50
1.2	9	11	9	21	17	11
1.3	18	11	19	21	21	21

The numerical values of satellite bandwidth available in channels (SBW) and the satellite power available (SP) used in our study are given in Table 2.

Table 2. Satellite bandwidth (SBW_k) and power (SP_k) for the channels

Problem	SBW_1	SBW_2	SBW_3	SBW_4	SBW_5	SBW_6	SBW_7	SBW_8	SBW_9	SBW_{10}
2.1	22	13	15	20	15	15	15	22	19	13
2.2	20	22	16	17	23	14	14	15	21	14
Problem	SP_1	SP_2	SP_3	SP_4	SP_5	SP_6	SP_7	SP_8	SP_9	SP_{10}
2.1	31	19	24	26	25	24	18	28	23	29
2.2	33	18	30	26	28	31	28	19	33	18

For the problem 2.1, the values of CBW_i for the customers from 1 to 20 are 5, 5, 5, 5, 4, 4, 3, 7, 6, 5, 6, 3, 6, 7, 4, 3, 6, 6, 4, and 4, respectively. The values of SP_i for customers from 1 to 20 are 5, 8, 9, 5, 7, 6, 8, 6, 7, 8, 6, 7, 9, 9, 7, 7, 5, 7, 5, and 9, respectively.

For the problem 2.2, the values of CBW_i for each customer are: 6, 7, 7, 7, 4, 7, 4, 6, 6, 4, 3, 3, 6, 3, 6, 6, 4, 3, 7, and 6, respectively. The values of SP_i are 8, 6, 9, 8, 6, 7, 6, 9, 8, 9, 9, 9, 8, 7, 5, 8, 6, 7, 7, and 7, respectively.

Parameter determination for ACO: Some definitions are given below. One-ant cycle is the cycle for evaluation and pheromone update for each ant in the colony. Generation is a period for the evaluations and pheromone updates for an ant colony of one generation. The following values are used in our simulations: Evaporation rate is 0.5, initial pheromone is 0.01, Q is 0.005, Max-Min is 0.01 and 1.0, w is 10, and σ is 7.

In our simulations, a total number of 40,000 cycles based on number of ants and generations are considered. The results are given in Table 3. In Table 3, the second column corresponds to 20 ants and the generations are 2,000 and hence a total of 40,000 cycles are considered. In the same manner we considered the number of ants 40, 100, 160 and 200 with corresponding generations so that the total number cycles is 40,000. The objective function value for these cycles are given in Table 3 for the test problems.

Now we will compare the performance of three strategies, namely, ACO-rank (ACS with only ranking), MMAS (ACS with only MMAS), and ACO-rank+MMAS (ACS with both ranking and MMAS). We have considered only problems 2.1 and 2.2 because they are hard problems. We can see from Tables

Table 3. ACO results for test problems

Problem	$20 \times 2,000$	$40 \times 1,000$	100×400	160×250	200×200
1.1	0.041666	0.041666	0.041666	0.041666	0.041666
1.2	0.046108	0.046108	0.046108	0.046108	0.046108
2.1	0.013274	0.013284	0.013717	0.013350	0.015171
2.2	0.051543	0.048220	0.052316	0.060514	0.059127

4 and 5 that the minimum values of the objective function obtained both in the MMAS and the ACOrank+MMAS are zero which is the optimal value. But, the computation time increases when these strategies are included.

Table 4. ACO performance comparisons of the three strategies with CPLEX: Problem 2.1

Algorithm	Average	Minimum	Maximum	Deviation	Comp. time (sec)
ACOrank	0.035884	0.007229	0.159086	0.016547	2.88
MMAS	0.008309	0.00	0.024526	0.004735	26.44
ACOrank+MMAS	0.007139	0.00	0.022416	0.004005	29.41
CPLEX	-	0.00	-	-	298.56

We now compare the computation time of the ACOrank+MMAS with the CPLEX optimization. Here the stopping criterion used for ACOrank+MMAS is as follows: The objective function value is zero or when no improvement during 1,000 cycles. We can see from Tables 4 and 5 that the computation time required in our approach is very much smaller in comparison with CPLEX.

Table 5. ACO performance comparisons of the three strategies with CPLEX: Problem 2.2

Algorithm	Average	Minimum	Maximum	Deviation	Comp. time (sec)
ACOrank	0.078833	0.046062	0.103765	0.012594	7.58
MMAS	0.046936	0.028040	0.070021	0.008930	26.39
ACOrank+MMAS	0.040877	0.025973	0.068094	0.007430	27.32
CPLEX	-	0.0129	-	-	887.78

We have also observed that the test Problem 2.1 does not have a unique optimal schedule. The schedule we obtain for Problem 2.1 using CPLEX is

$s_1 = \{8\ 2\ 0\ 3\ 0\ 8\ 3\ 0\ 3\ 4\ 3\ 5\ 0\ 5\ 8\ 4\ 8\ 2\ 4\ 2\}$ and the objective function value is 0.00. The schedule s_1 means the following assignment of customers to channels:

- Customers 3, 5, and 8 are assigned to channel 0.
- Customers 2, 18, and 20 are assigned to channel 2.
- Customers 4, 7, 9, and 11 are assigned to channel 3.
- Customers 10, 16, and 19 are assigned to channel 4.
- Customers 12, and 14 are assigned to channel 5.
- Customers 1, 6, 15, and 17 are assigned to channel 8.

We see that channels 1, 6, 7, and 9 are not used in this assignment (s_1).

The schedule we obtain for the same problem using ACOrank+MMAS is given as $s_2 = \{0\ 2\ 7\ 0\ 5\ 6\ 3\ 3\ 3\ 2\ 6\ 0\ 5\ 4\ 4\ 0\ 7\ 0\ 3\ 4\}$ and the objective function value is zero. The schedule s_2 means the following assignment of customers to channels:

- Customers 1, 4, 12, 16, and 18 are assigned to channel 0.
- Customers 2 and 10 are assigned to channel 2.
- Customers 7, 8, 9, and 19 are assigned to channel 3.
- Customers 14, 15, and 20 are assigned to channel 4.
- Customers 5, and 13 are assigned to channel 5.
- Customers 6, and 11 are assigned to channel 6.
- Customers 3, and 17 are assigned to channel 7.

We see that channels 1, 8, and 9 are not used in this assignment (s_2). Note that the computation times of the proposed methods are far less than CPLEX, which shows the efficiency of the proposed methods.

5 Conclusions

Assigning customers to satellite channels is a difficult combinatorial optimization problem. Hence, the meta-heuristic method of ant colony optimization is presented to the problem of assigning customers to satellite channels. An ant colony system (ACS) with strategies of ranking and Max-Min ant system (MMAS) for an effective search of the best/optimal assignment of customers to satellite channels under a dynamic environment is presented. Our simulation results show that this methodology is successful in finding an assignment of customers to satellite channels. Three strategies, ACS with only ranking, ACS with only MMAS, and ACS with both ranking and MMAS, are considered. A comparison of these strategies are presented to show the performance of each strategy. Numerical examples are presented to show that this approach takes very little computation time to get very-near optimal assignment for this problem in comparison with standard optimization techniques.

Acknowledgement. This work was in part supported by the ITRC under the auspices of Ministry of Information and Communication, Korea.

References

1. E. Bonabeau, M. Dorigo, and G. Theraulez, *Swarm Intelligence: From Natural to Artificial Intellience*, Oxford University Press, New York, NY, USA, 1999.
2. B. Bullnheimer, R. Hartl, and C. Strauss, "An improved ant system algorithm for the vehicle routing problem," *Annals of Operations Research*, vol. 89, pp. 319-328, 1999.
3. D. Cattrysse, and L. N. Van Wassenhove, "A survey of algorithms for the generalized assignment problem," *European Journal of Operational Research*, vol. 60, pp. 260-272, 1992.
4. M. Dell'Amico and S. Martello, "Open shop, satellite communication and a theorem by Egervary (1931)," *Operations Research Letters*, vol. 18, pp. 209-211, 1996.
5. M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artificial Life,* vol. 5, pp. 137-172, 1999.
6. M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: Optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. B.26, 29-41, 1996.
7. M. Dorigo, and T. Stutzle, *The ant colony optimization metaheuristic: algorithms, applications, and advances,* in F. Glover, and G. Kochenberger (Editors), *Handbook of Metaheuristics,* Kluwer Academic Publishers, Norwell, MA, USA, pp. 251-258, 2002.
8. H. Lee, D. H. Ahn, and S. Kim, "Optimal routing in non-geostationary satellite ATM networks with intersatellite link capacity constraints," *Journal of the Operational Research Society*, vol. 54. pp. 401-409, 2003.
9. J. Levine, and F. Ducatelle, "Ant colony optimization and local search for bin packing and cutting stock problems," *Journal of the Operational Research Society,* vol. 55, pp. 705-716, 2004.
10. D. C. Montgomery and L. A. Johnson, *Operations Research in Production Planning Scheduling and Inventory Control*, John Wiely & Sons, 1974.
11. C. Prins, "An overview of scheduling problems arising in satellite communications," *Journal of the Operational Research Society*, vol. 45. pp. 611-623, 1994.
12. C. H. Scott, O. G. Skelton, and E. Rolland, "Tactical and strategic models for satellite customer assignment," *Journal of the Operational Research Society*, vol. 51. pp. 61-71, 2000.
13. T. Stützle, "MAX-MIN Ant system for the quadratic assignment problem," Technical Report AIDA-97-4, FG Intellektik, TU Darmstadt, Germany, 1997.
14. P. Tarasewich and P.R. McMullen, "Swarm intelligence: Power in numbers," *Communications of the ACM,* vol. 45, pp. 62-67, 2002.

Advanced Remote-Subscription Scheme Supporting Cost Effective Multicast Routing for Broadband Ubiquitous Convergence IP-Based Network

Soo-Young Shin[1], Young-Muk Yoon[1], Soo-Hyun Park[1,⋆], Yoon-Ho Seo[2], and Chul-Ung Lee[2]

[1] School of Business IT, Kookmin University
{sy-shin,ymyoon,shpark21}@kookmin.ac.kr
[2] Department of Information and Management Engineering, Korea University, Korea
{yoonhoseo,leecu}@korea.ac.kr

Abstract. Mobile multimedia services such as TV-call or video streaming are gradually becoming popular in the 3^{rd} or more generation mobile network(IMT-2000). IP-based IMT network platform represents an evolution from IMT-2000. The structure of IP-based IMT network as ubiquitous platform is three-layered model : Middleware including Network Control PlatForm (NCPF) and Service Support PlatForm (SSPF), IP-BackBone (IP-BB), access network including sensor network. Mobility Management (MM) architecture in NCPF is proposed for IP-based IMT network in order to manage routing information and location information separately. The generous existing method of multicast control in IP-based IMT network is Remote-subscription. But Remote-subscription has problem that should be reconstructed whole multicast tree when sender in multicast tree moves to another area. To solve this problem, we propose the way to put Multicast-manager in NCPF.

Keywords: Multicast-manager, IP-based IMT Network, Mobility Management, SSPF, NCPF, Multicast Routing.

1 Introduction

Ubiquitous Network, in charge of future communications, should support wideband seamless mobility management, service and furthermore guarantee transmission of large amount of multimedia traffics which has been increasing explosively by the development of wireless accessing technologies. According that ITU-R suggested a guideline all networks, including telecommunications, it should be converted to IP-based networks [1]. NTT DoCoMo proposed IP-based IMT network platform [2][3] as the next generation All-IP mobile network structure taking IP technologies and rapidly increasing multimedia traffics into consideration. This

⋆ Coresponding author.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 180–189, 2007.

platform was designed to have an ability to transmit large amount of multimedia traffic efficiently and accommodate different kinds of wireless access systems. This platform also supports seamless mobility and application services. The structure of IP-based IMT network is categorized into three layers which consists of middleware including NCPF and SSPF, IP-BB and access network including sensor networks [1][4]. IP-based IMT network, as the next generation backbone network, is designed to support mobility management fundamentally which is not the present networks. This mobility management function plays an important role not only making it easy for seamless mobility of terminals or users but also providing various services which are not even imaginable within conventional infrastructure. Mobile multicast is an important research item. At present, there are multicast techniques taking mobility into consideration such as Remote-subscription, Bi-directional tunneling, Mobile Multicast (MoM) and eXplicit Multicast Mobile IPv6 (XMIPv6) for Mobile IPv6 (MIPv6) and so on. Since Bi-directional tunneling, MoM and XMIPv6 use a technique of transmitting packets as a unicast via Home Agent (HA) when they are routing, they cut down the effect of resource integration of multicast. Therefore, Remote-subscription should be used for effective multicast in IP-based IMT network. It has a problem at present, however, that multicast trees of the whole network have to be reconfigured whenever Mobile Node (MN) for sender travels to another area. In this paper, we proposed a solution for the problem by managing the information of multicast group members with a Multicast-manager in NCPF of IP-based IMT network. In Chapter 2, the mobile cast technique used in conventional IP-based ITM network and its problem are described. In Chapter 3, a solution to the problem is described. Based on a series of simulations, performance comparisons of the proposed technique and the conventional mobile multicast technique are described. And final conclusions in Chapter 5.

2 Conventional Mobility Management of IP-Based Network and Multicast Support Method (MIPv4 / MIPv6)

In Mobile IP (MIP) Multicast, there are Remote-subscription and Bi-directional tunneling method which has been proposed by IETF. In case of Remote-subscription, when MN moves to other access networks, multicast group will be reconfigured by a request of MN itself. This method is comparatively simple in its operation and efficient in case that MN's move is not so frequently. In case that MN receives packets, they are delivered through a certain route, which has been optimized based on multicast routing protocol. In case that MN transmits multicast data, a delivery multicast tree is reconfigured and a sending node transmits packets to FA. After that the packets are delivered to a receiving node through a typical multicast delivery route [5]. Figure 1 shows data receiving route in case that moving nodes (Receiver1 and Receiver2) have been moved from its Home Agents (HA1 and HA2 respectively) to Foreign Agent (FA). And at this time, the packets are delivered through an optimized route without passing through HA. In Remote-subscription, routing routes are maintained adequately and relatively

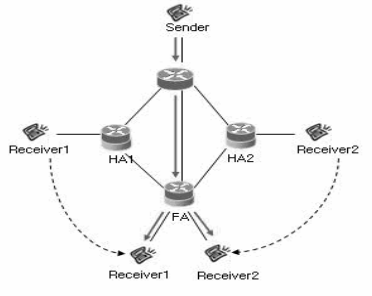


Fig. 1. Remote-subscription

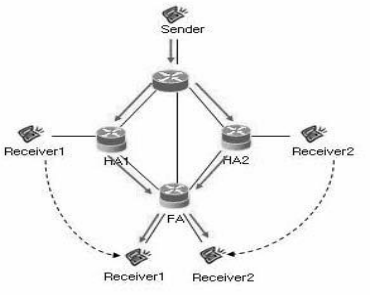


Fig. 2. Bi-directional Tunneling

simple. However, network load will be increased by multicast tree reconfigurations. Bi-directional tunneling is an HA-based multicast method. When MN moves to other access network, the moved MN will send and receive data by tunneling with HA. In case that MN on passage transmits data, MN will send data to HA by using its Home Address in Received Address Field (RAF) of IP header. The transmitted packets are delivered to multicast group members via a certain delivery route. In case that MN receives data, HA will capture multicast packets, which are on the way to MN, and capsulate it into unicast datagram, whose destination is MN's home address, to conduct tunneling to FA [5]. Figure 2 shows the route of data receiving when Receiver1 and Receiver2 is moving from HA1 and HA2 respectively into FA. As for this method, it is possible to send or receive data without any relationship with other group members and maintain compatibility to existing networks. However, the method has shortcomings of non-optimized delivery route and damaged resource integration effects resulted from unicast transmission when tunneling. XMIPv6 originated from MIPv6, has been proposed to provide multicast services. When MN moves to other access network area, MN transmits its data packets by tunneling with HA at first. MN, which has received multicast data, sends Binding Update (BU) message to other MN, which transmitted the data. Then MN, which received the message, comes to be able to communicate directly with multicast group members without passing through HA [6]. Even though this method is able to communicate through optimized route after BU, there are still unresolved problems that it is required to pass through HA at least one time and inefficiency caused by tunneling. Bi-directional tunneling and XMIPv6 are conventional multicast techniques which uses tunneling between HA and MN. Since these conventional techniques use unicast transmission, they decrease resource integration effect. For this reason, Remote-subscription method has been used in IP- based IMT network. In Remote-subscription method, however, overall multicast tree has to be reconfigured whenever sending node moves. Consequently, network traffic increases and problems of message loss are caused. To solve those problems, we propose a method of placing Multicast-manager in NCPF of IP-based IMT

network. In the proposed method, Multicast-manager manages the information of sending/receiving nodes and conducts a series of procedures to improve the efficiency of resource.

3 A Proposal for IP-Based IMT Network Mobile Multicast

3.1 Group Member Management Using Multicast-Manager

In Remote-subscription method, the whole multicast tree has to be reconfigured whenever the sender, which offers multicast service, is moved. Therefore, there are some problems that network traffics are increased and some messages are lost during multicast tree reconfiguration. To resolve the problems, it is proposed to place a Multicast-manager in NCPF of IP-based IMT network. The Multicast-manager has information about sender and receiver, which belongs to Multicast Service Group (MSG), and IP host address (IPha) and IP routing address (IPra) are managed simultaneously to update IPra of MN after MN is moved. Therefore, the efficiency can be improved by changing the single route between sender and Multicast-manager only, not by changing the overall multicast tree of the network after sender, which transmits multicast data, is moved.

3.2 Multicast Operation

In this chapter, the roles of Multicast-manager are described using a designated example. It is assumed that there are one sender (S1) and three receivers (R1, R2, R3) in an arbitrary domain of IP-based IMT network and it is also supposed that any required procedures for multicast services can be conducted using Multicast-manager. Multicast service consists of the following four steps: (1) Initialization of multicast session, (2) Join of MN requiring multicast service, (3) Mobility support procedure when MN moves during multicast service, (4) Ending the multicast service.

3.3 Initialization of Multicast Session

As an example of generating a multicast session, Multicast-manager in NCPF is asked to make a new Multicast session when MN (S1) within the area of Base Station (BS1) wants to be a sender of multicast service. Figure 3 and 4 show the initialization procedure for multicast session by mobile node, S1, and the message flow. 1) S1 sends multicast session address M1 and new session message including IP host address of S1, HaS1, to Multicast-manager which is located in NCPF. 2) Multicast-manager receives IP routing address of S1, Ra1a, from RM. 3) Multicast-manager makes Multicast-manager table having multicast group ID, M1, and stores IPha of S1 (HaS1) and IPra of S1 (Ra1a) to the table. 4) Multicast-manager sends multicast group ID (M1) and IPha / IPra of S1 to AR1, to which S1 belongs. 5) AR1 makes Table for Multicast Sender (TMS) in

cache and stores multicast group ID (M1), IPha of S1 (HaS1) and IPra of S1 (Ra1a) to the table. 6) AR1 sends reply message to S1. By this message, S1 is notified that Multicast-manager table is generated in Multicast-manager and a new multicast session is started.

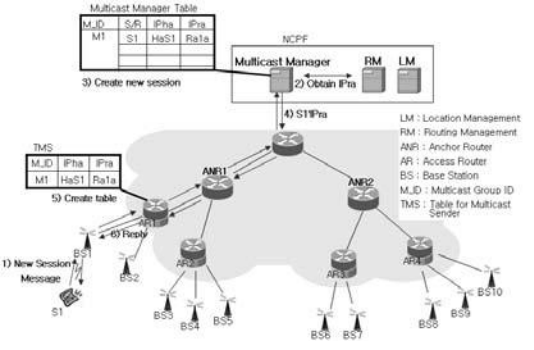


Fig. 3. Initialization of Multicast Session

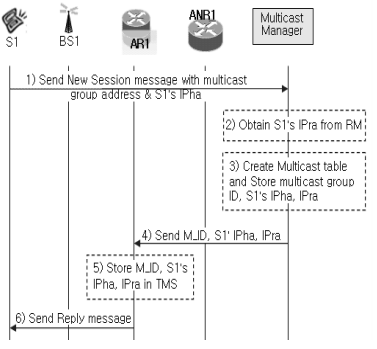


Fig. 4. Message Flow

3.4 Joining Multicast Session

Nodes who want to join with multicast session will go through the following procedures. It is supposed that Multicast-manager already has managing table (MMT : Multicast-manager Table) of M1 which is generated in the preceding procedure. Figure 5 and 6 show the procedure of MN, R1, joining multicast group, M1, and message flow between Multicast-manager and the other network elements. 1) R1 sends join query, which contains multicast group ID (M1), IP host address of R1, HaR1, to Multicast-manager. 2) Multicast-manager obtains IP routing address of R1, Ra2b, from RM. 3) Multicast-manager stores IPha

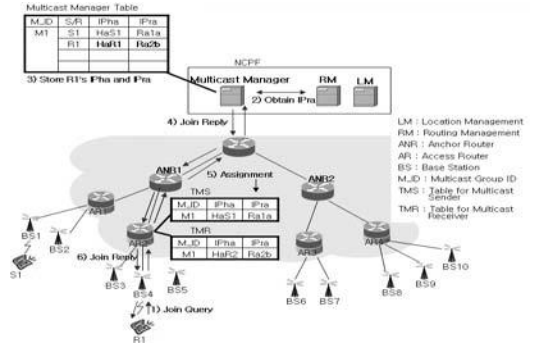


Fig. 5. Joining Multicast Session

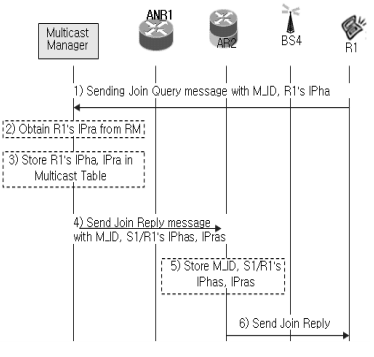


Fig. 6. Message Flow

and IPra of R1 (HaR1, Ra2b respectively) to Multicast-manager table having group ID, M1. 4) Multicast-manager sends M1, IPha and IPra of S1 and IPha and IPra of R1 to AR2, to which R1 belongs. 5) AR2 generates TMS and Table for Multicast Receiver (TMR) in its cache and stores the information about S1 and R1 (S1 / R1's group ID, IPha and IPra) to TMS and TMR respectively. 6) AR2 sends join reply message to R2 to notify it has been joined multicast group M1.

3.5 Mobility Support Procedure

Figure 7 and 8 show the procedure of updating the information of IPra of S1, which moved from BS1 to BS7 when sender of S1 and receivers of R1, R2 and R3 are within multicast group of M1. 1) S1 moves from BS1 to BS7. 2) After S1 receives advertisement message which is sent by BS7 periodically, S1 sends join query message, including multicast group ID of M1 and IP home address of S1, to Multicast-manager. 3) Multicast-manager obtains the new IP routing address of S1, Ra3b, from RM. 4) Multicast-manager updates IP routing address of S1 in Multicast-manager table having group ID of M1. 5) Multicast-manager sent the new IP routing address of S1 to AR1, to which S1 was belongs, and to ARs (AR1, AR2 and AR3) , to which members of group M belong. 6) ARs, to which members of group M belong, update IP routing address of TMS in their cache and delete TMS in their cache since AR1 does not has multicast member any longer. 7) After S1 moved, AR3, to which S1 is currently belong, sends join reply message to S1 and consequently S1 is notified that it has been joined multicast group M1 after its move. Figure 9 and 10 show message flow and the procedure of updating

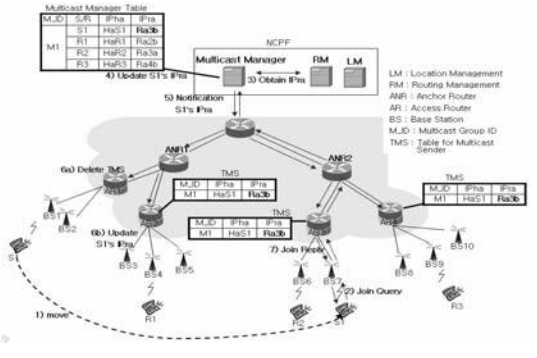


Fig. 7. Movement of Sender

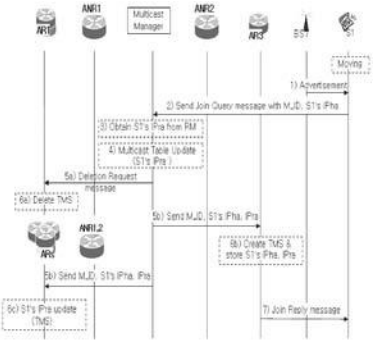


Fig. 8. Message Flow

the information of IPra of R1 in case that R1 moved into the area of BS8 on condition that sender S1 and receiver R1, R2 and R3 belong to multicast group M1. 1) R1 moves from BS4 to BS8. 2) R1 receives Advertisement message, which is periodically transmitted by BS8, and sends join query message, which includes multicast group ID M1 and IP home address of R1, to Multicast-manager. 3) Multicast-manager obtains Ra4a, the new IP routing address of R1, from RM.

4) Multicast-manager updates IP routing address of R1 in Multicast-manager table having group ID M1. 5) Multicast-manager sends old and new IP home address of R1 to AR2, to which mobile node R1 was previously belongs, and AR4, to which R1 is presently belongs, respectively. 6) AR4, to which R1 is presently belongs, stores IP home address and IP routing address of R1 to TMR in its cache and deletes TMR in its cache since AR2, to which R1 was belong, does not has multicast member any longer. 7) AR4, to which R1 is currently belong after its move, sends join reply message to R1. Therefore, R1 is notified that it has been joined multicast group M1 after its move.

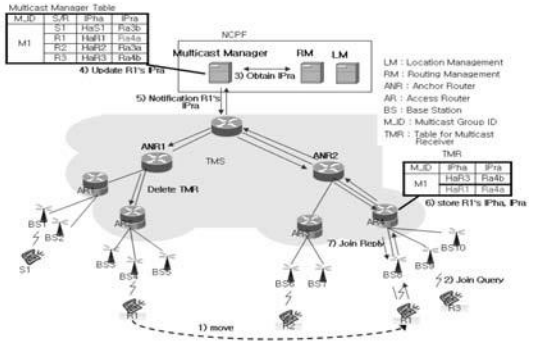


Fig. 9. Movement of Receiver

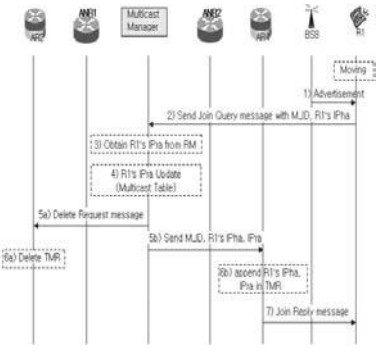


Fig. 10. Message Flow

3.6 Finishing Multicast Service

Figure 11 and 12 show the procedure of ending the service of multicast group M1 and message flow for finishing multicast service. 1) S1 sends termination message with multicast group ID M1 to Multicast-manager. 2) Multicast-manager deletes Multicast-manager table of group ID M1 from its date registry. 3) Multicast-manager sends termination message to AR1, AR2, AR3 and AR4, to which

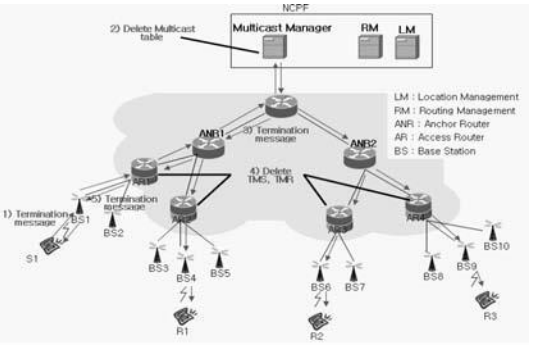


Fig. 11. Finishing Multicast Service

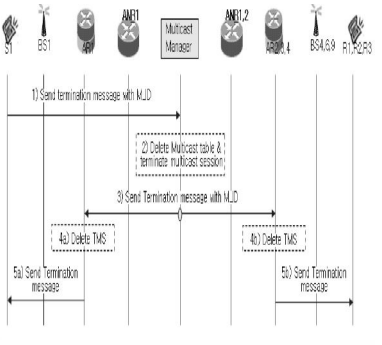


Fig. 12. Message Flow

members of M1 belong. 5) AR1, AR2, AR3 and AR4 sends Termination message to S1, R1, R2 and R3, which are members of AR1, AR2, AR3 and AR4 respectively. Therefore, all multicast group members is notified that multicast service has been finished.

4 Simulation

For the simulation for the proposed method of using Multicast-manager, topologies and simulation scenarios are organized using 2 anchor routers, 4 access routers and 10 base stations. Simulations of multicast tree reconfiguration are conducted by two parts: one part is a simulation after handoff of sender and the other is the one after handoff of receiver. For reliable test, performance analysis for each group member is conducted with 4 20 multicast group members on the assumption that mobile nodes are moving with constant velocity and a fixed direction.

4.1 Case of Sender's Mobility

Control traffic amount and delay of the network when mobile node S1 (sender) reconfigures multicast tree after handoff are measured. In figure 13 14, traffic and delay of Remote-subscription method is compared with those of the proposed method respectively. Table 1 shows the comparison result of delay and traffic, which is the number of signaling message in case of join query generation according to the number of groups. In addition, the relative performance

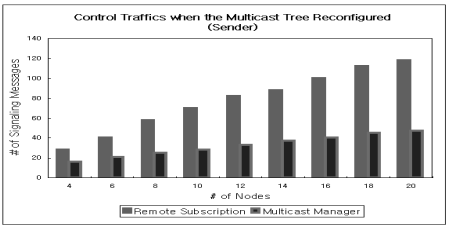


Fig. 13. Control Traffic (sender)

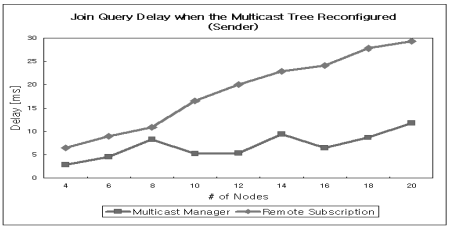


Fig. 14. Join Query Delay

Table 1. Network Traffic Increase and Delay (Sender) when Tree Reconfigured

Multicast Group members		4	6	8	10	12	14	16	18	20
Traffics	Remote-subscription	28	40	58	70	82	88	100	112	118
	Multicast-manager	16	21	25	28	33	37	40	45	47
	Improvement rate -MM	42.9	47.5	56.9	60	59.8	60	60	59.8	60.2
Delay (ms)	Remote-subscription	6.45	8.97	10.87	16.59	20.06	22.92	24.15	27.78	29.28
	Multicast-manager	2.87	4.53	8.26	5.21	5.34	9.42	6.43	8.70	11.76
	Improvement rate	55.5	49.5	24.1	68.6	73.3	58.9	73.4	68.7	59.8

improvement rate of the proposed method is also listed in the table. The proposed method showed average traffic reduction of 57.07

4.2 Case of Receiver's Mobility

Control traffic and delay of the network when mobile node R1 (receiver) reconfigures multicast tree after handoff are measured and shown in figure 15 16. Horizontal axis stands for the number of multicast group members and vertical axis for generated traffic in case of tree reconfiguration.

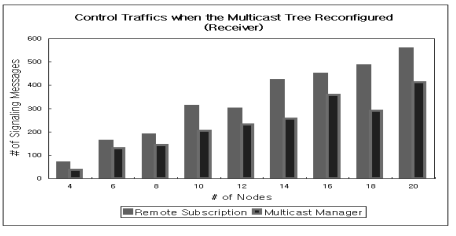


Fig. 15. Control Traffic (receiver)

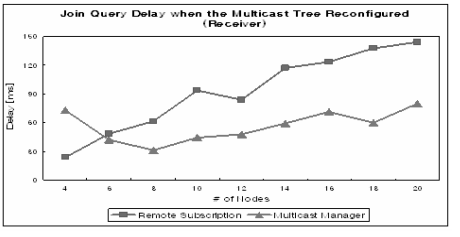


Fig. 16. Join Query Delay

Table 2. Network Traffic Increase and Delay (Reciever) when Tree Reconfigured

Multicast Group members		4	6	8	10	12	14	16	18	20
Traffics	Remote-subscription	68	162	188	310	298	422	448	484	556
	Multicast-manager	36	129	144	205	232	256	359	291	413
	Improvement rate -MM	47.1	20.4	23.4	33.9	22.1	39.3	19.9	39.9	25.7
Delay (ms)	Remote-subscription	23.75	48.37	61.32	93.79	83.99	117.0	123.4	138.0	144.3
	Multicast-manager	73.0	42.1	30.89	44.43	47.61	58.95	71.11	59.92	80.04
	Improvement rate	55.5	49.6	24.1	52.6	43.3	49.6	42.3	56.6	44.5

In figure 15, it is shown that, in case the number of multicast group member is 5, Sender and Receiver come closer each other and traffic of Remote-subscription method is smaller. As the number of multicast group member is increased, however, control traffic and delay of the proposed method, which uses Multicast-manager, is significantly reduced compared with Remote-subscription method. Table 2 shows that, in case of the proposed method, the average reduction of control traffic and delay are 30.2.

5 Conclusion

In this paper, a method for implementing multicast in IP-based IMT network by placing Multicast-manager in NCPF was proposed. In conventional mobile multicast method, there are problems that transmission route is not optimized

when sender moves and overall multicast tree has to be reconfigured whenever sender moves. With the proposed method, however, network load resulted from tree reconfiguration can be reduced by placing Multicast-manager in NCPF to manage MN's moves and consequently to reconfigure the route between sender and Multicast-manager only, not reconfiguring the overall multicast tree. Simulation results show that, when Sender moves, the average control traffic and delay of the proposed method are reduced by 57.07 For future works, a study on techniques to guarantee unique of multicast group ID, in case that sender's transmits new session message to Multicast-manager, has been underway.

Acknowledgments. This research was supported in parts by grant No. (IITA-2006-C1090-0603-0044) from the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment) and supported in parts by grant No, (R01-2006-000-10941-0) from the Basic Research Program of the Korea Science and Engineering Foundation. And This work was supported in parts by the 2006 faculty research program of Kookmin University in Korea.

References

1. ITU-R Draft Recommendation: Vision, framework and overall objectives of the future development of IMT-2000 and systems beyond IMT 2000.(2002)
2. H.Yumiba, et al.: IP-based IMT Network Platform. IEEE Personal Communication Magazine, Vol. 8. (2001) 18–23
3. K. Imai, M. Yabusaki, and T. Ihara: IP2 Architecture towards Mobile Net and Internet Convergence. WTC2002 (2002)
4. T. Okagawa, et al.: Proposed Mobility Management for IP-based IMT Network Platform. IEICE Trans Commun. Vol. E88-B. (2005) 2726-2734
5. C.Perkins: IP mobility support. RFC 2002 (1992)
6. C.Perkins: IP Mobility Support for IPv4. RFC 3220 (2002)
7. Y. Fang: Movement-based Mobility Management and Trade Off Analysis for Wireless Mobile networks. IEEE Trans. Comput. Vol. 52. (2003) 791-803
8. I. Akyildiz, and W. Wang: A Dynamic Location Management Scheme for Next-Generation Multitier PCS Systems. IEEE Trans. Wireless Communication. Vol. 1. (2002) 178-189
9. X. Zhang, J. Castellanos, and A. Capbell: P-MIP: Paging Extensions for Mobile IP. ACM/Kluwer Mobile Networks and Applications. Vol. 7. (2002) 127-141

Dual Priority Scheduling Based on Power Adjust Context Switching for Wireless Sensor Network

Taeo Hwang[1], Jung-Guk Kim[2], Kwang-Ho Won[1], Seong-Dong Kim[1], and Dong-Sun Kim[3,⋆]

[1] Ubiquitous Computing Center, Korea Electronics Technology Institute, 68 Yatap-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 463-816, Korea
{taeo, khwon, sdkim}@keti.re.kr
[2] Hankuk Univ. of Foreign Studies, Korea
jgkim@hufs.ac.kr
[3] DxB·Communication Convergence Research Center, Korea Electronics Technology Institute, 68 Yatap-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 463-816, Korea
dskim@keti.re.kr

Abstract. The wireless sensor network (WSN) node is required to operate for several months with the limited system resource such as memory and power. A general WSN node operates in the active state during less than 1% of the several-month lifetime and waits an event in the inactive state during 99% of the same lifetime. This paper suggests a power adjust dual priority scheduler (PA-DPS) for low-power, which has a structure to meet the requirements for the WSN by estimating power consumption in the WSN node. PA-DPS has been designed based on an event-driven approach and is based on the dual-priority scheduling structure, which has been conventionally suggested in the real-time system field. From experimental results, PA-DPS reduced the inactive mode current up to 40% under the 1% duty cycle.

1 Introduction

In order to achieve sensing operation and wireless network protocols, the wireless sensor node requires real-time system properties such as timeliness and predictability. In addition, for the applications such as voice transmission and peripherals of computer, low latency and guarantee of QoS are required in the wireless sensor network (WSN). The timeliness requires a different approach from the timeliness for on-time completion which meets deadlines required by a general real-time system. Because the sensor node is in the inactive state during the most of its lifetime, on-time execution of tasks, rather than on-time completion of tasks should be assured. TinyOS [1] is a representative OS for the WSN because of its small code size and the efficient component based programming model. For efficient structure, it has a simple scheduler which is based on an

⋆ Corresponding author.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 190–199, 2007.

event driven approach and uses a sleep and wakeup scheme for power management. When the system is in the idle state, the scheduler changes the system mode to the sleep mode which results in the lowest power consumption. It is also used by a preemptive multi-threading OS for the WSN like MANTIS [2]. If the WSN system operates under 1% duty cycle, the WSN node is in the idle state during more than 99% of its lifetime. More than 95% of the entire power is consumed in the idle state. If the sleep mode current is reduced to just a few , the battery operated system can have a significantly longer lifetime. If the power level provided in the MCU, the power of the peripherals and elements of the system can be controlled, the system can maintain in the relatively lower power state as long as on-time execution is assured. For this reason, this paper provides the following technical idea: In the general WSN node, peripherals such as RF transceiver and sensors steadily consume a small quantity of currents in spite of idle state, because the elements related to peripherals must be supplied power for the reliable execution of corresponding task. On the other hand, timers embedded in microcontroller significantly consume lower current than peripherals. The tasks for WSN can be organized to the predictable and periodic properties that are executed by the timer, so that the unused hardware blocks can be turned off completely in idle state. If necessary hardware blocks are enabled earlier enough to cover the time duration to initialize the hardware device, the desired tasks can be executed well and the system can maintain with a relatively lower power during the sleep mode. The rest of this paper is organized as follows. Section 2 describes the design of the power adjust context switching based dual-priority scheduling (PA-DPS). In developing the SoC for the transceiver of IEEE 802.15.4 [3], the function, which is required for the hardware, has already been designed and achieved. Section 3 describes the achieved chip set and hardware structure. Section 4 provides an analysis of the duty cycle in the WSN, and experiments and simulation to compare the scheduler suggested in this paper with an approach by an existing simple sleep and wakeup scheme for low-operation.

2 Scheduler Design

In order to achieve a low power real-time system for WSN, we propose the modified dual-priority scheduling (DPS) structure [4]. The modified DPS structure can determine the system operation speed by estimating the worst case execution time (WCET) based on deadlines. Because of the WCET, it is advantageous in the power management. The real-time tasks in the WSN node have the different long periods. The deadlines of the real-time tasks have no significance since its duty cycles are less than 1% in WSN. The WSN node should assure on-time execution of the periodic tasks. Thus, the modified DPS structure can be simplified in order to meet the requirements for the sensor node. A power_adjust() function is inserted into the scheduling time for low-power operation. The power_adjust() function is designated by each task. For each task, it defines the required hardware blocks and the power level provided by the MCU, so that it can reconstruct

the system prior to run a task. The modified DPS concept, which consists of the lower run queue (LRQ) and the upper run queue (URQ), is simplified to be the concept of the time-triggered message-triggered object (TMO) [5, 6] that was suggested for a real-time object programming model. Tasks of the sensor node are classified into the time triggered periodic tasks and the message or event triggered non-periodic tasks so that they are defined as different task control blocks. The periodic tasks are executed by a system timer in a periodic ready queue (PRQ), and the non-periodic tasks are executed by an event in a sporadic ready queue (SRQ). The periodic tasks of the PRQ have higher priorities than the non-periodic tasks of general SRQ. They are executed by the FIFO-based serialized (context-switching free) scheduling [7]. The events generated from the hardware interrupt handler are enqueued by the event handler. In case of system timer generated events, the executed sporadic tasks are preempted so that the tasks of PRQ are executed. The periodic tasks have guard time and time left fields in their control block. The time_left is information on time left and the guard_time relates to time information needed to enable the hardware device, which accesses the tasks. The periodic cycle which the scheduler is operated by the system timer is determined by subtracting the guard_time from the minimum time_left in the PRQ. The tasks of the PRQ can wake up as earlier as the guard_time so as to enable the needed hardware. After the PRQ tasks are completed, the contexts of the sporadic tasks are restored. If no sporadic tasks exist, the system is converted into the lowest power state, i.e., POWER_IDLE

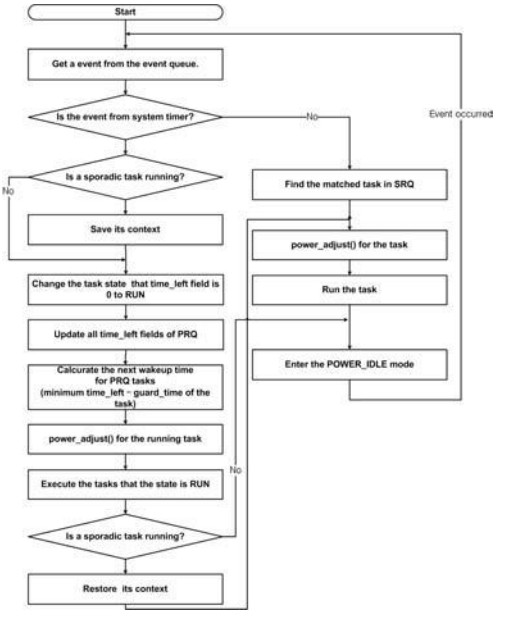


Fig. 1. Flowchart for the Power Adjust Based DPS

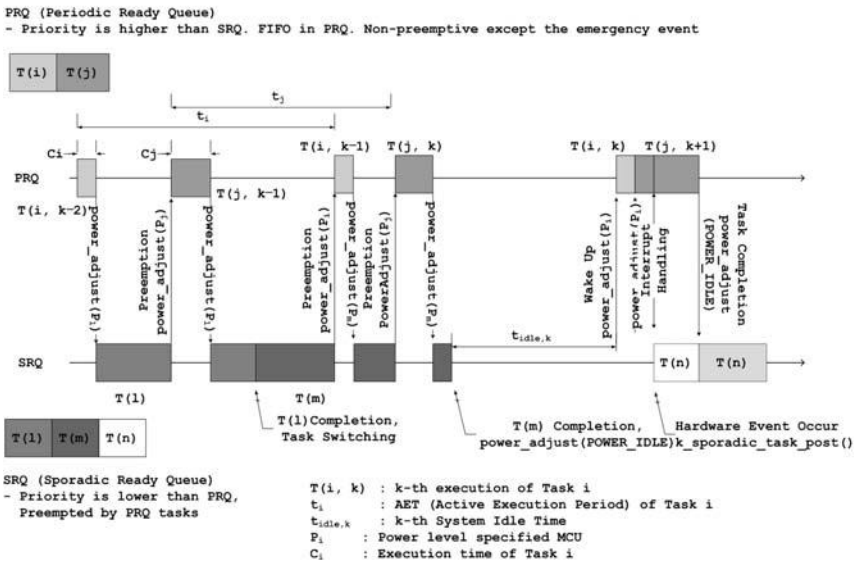


Fig. 2. The operation of PA-DPS

state. This scheduling process includes the power_adjust() process for each task and Fig. 1 shows a flowchart for the scheduler.

As shown in Fig. 2, tasks i and j are in the WAIT state to wait timer events in PRQ. The system timer is set by values obtained from subtraction of T(i) period from T(j) period when the tasks are generated. Since the system timer is achieved to have 32 bit scale per 1 μs, the timer can be set by large scaled values, which range from 16 μs to approximately 60 minutes. The value for the timer tick varies depending on the variable clock source and the 16 bit hardware counter scale. As a result, the timer interrupt for the periodic tasks can be minimized. When the system timer is expired, Task T(i, k-2) in Fg. 2 wakes up by the system interrupt and is executed during the Ci time period. When the execution is completed, T(1) of the SRQ is executed by events. T(1) can be executed at a low operation speed, and is preempted by the system timer interrupt during the execution so that T(j,k-1) is executed. T(j,k-1) is executed during Cj. When the execution is completed, T(1) continues to execute the tasks. After the tasks by T(1) is completed, task T(m) of the SRQ is executed by events. Likewise, the tasks of the PRQ and the SRQ are executed. Where there are no tasks to be executed in the SRQ and the PRQ, the system is converted into the POWER_IDLE mode through the power adjust process. In the POWER_IDLE mode, all power, except for the hardware block to maintain the system timer and the memory, is blocked. In the POWER_IDLE mode, the sensor node system wakes up by the hardware interrupt of the system timer, and then renews the system clock and the events to execute the corresponding tasks of the PRQ.

Fig. 3 shows a simple application program code for reading values from a temperature sensor per second. The application consists of one periodic task and one sporadic task, and has power adjusting functions. The sensing_trigger, which is a periodic task, enables waking up every second to read the values from the temperature sensor. Before the sensing_trigger is executed so as to enable and initialize the temperature and the hardware of analog-to-digital converter (ADC), the scheduler calls one of the pa-sensing-trigger, which is the power_adjust() function. The sporadic task is executed by the KE_ADC_COMPLETED which is a system event. When the AD conversion, which is initiated by the sensing_trigger, is completed, hardware interrupt occurs. An interrupt handler generates an event to deliver it to the scheduler. The scheduler receives the event to execute the pa_temp_detect, which is the power_adjust() for temp_detect, thereby powering off the sensor and the ADC hardware and executing the temp_detect.

```
POWER_ADJUST(pa_temp_detect) {
    h_temp_stop();
    h_adc_stop();
    k_power_mode(PM_POWER_SAVE);
}

SPORADIC_TASK(temp_detect) {
    uint32_t temperature;
    temperature = adcInfo.result;
  /* calculate temperature in hundredths of a degree.*/
    . . .
}

POWER_ADJUST(pa_sensing_trigger) {
    h_temp_init();
    h_adc_init();
    k_power_mode(PM_POWER_SAVE);
}

PERIODIC_TASK(sensing_trigger) {
    k_sporadic_task_post
    (
        temp_detect,        /* task func pointer */
        pa_sensing_trigger, /* power adjust function */
        KE_ADC_COMPLETED,   /* triggered event */
        FALSE           /* if TRUE,
                           repeat whenever the event */
    );
    h_adc_start();
}
```

3 Hardware Platform

We implement a baseband processor and MAC accelerator for sensor network. Baseband processor mainly consists of a digital modem, a 8-bit microcontroller, and MAC hardware. The digital modem basically consists of a modulator with 6-bit digital to analog converter (DAC) and a demodulator with 4-bit analog to digital converter (ADC) that transmits or receives the radio frequency signals. A general purpose of the microcontroller is to perform software MAC and sensor network protocols. The hardware MAC supports auto cyclic redundancy check (CRC), four programmable timers for checking the timing rules of MAC or network protocols such as acknowledgement response time. In addition, a MAC frame block generates a transmitted data frame and parses information from received frames. It also generates an acknowledgement frame and automatically transmits for carrier sense multiple access-collision avoidance (CSMA-CA) protocols. The hardware MAC also supports 128 bit advanced encryption standard (AES) encryption and decryption for reducing the processing time and power of embedded microcontroller in case of security operation. Fig. 3 relates a block-diagram of the baseband processor.

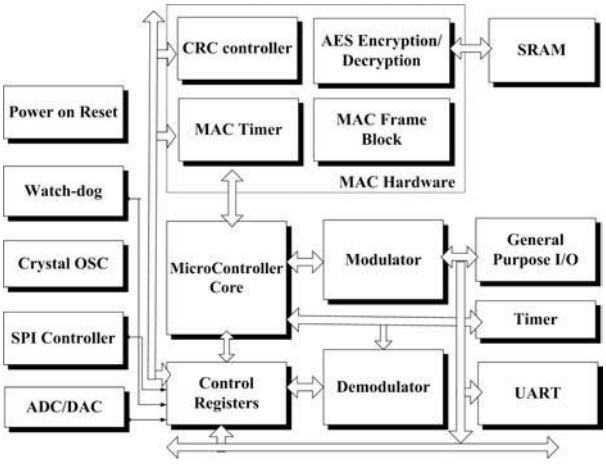


Fig. 3. Block Diagram of Baseband Processor

4 Experimental Results

System power consumption can be defined as a current consumption quantity (Q), which is calculated by multiplying needed current (I) by use time (t) [8]. The WSN node can be divided into an active time section and an inactive time section. The sum of these time sections refers to total consumption current capacity in the WSN node.

$$Q_{total} = Q_{active} + Q_{inactive} = t_{active} \cdot I_{active} + t_{inactive} \cdot I_{inactive} \quad (1)$$

Table 1. Power Consumption Parameters

POWER_FULL	7,000 μA	32 MHz clock
POWER_SAVE	3,800 μA	16 MHz clock
POWER_STANBY	296 μA	32.768 MHz clock
POWER_IDLE	1 μA	32.768 MHz clock
POWER_DOWN (Only Power on Reset)	1 μA	-
Radio_TX	17,700 μA	-
Radio_RX	20,000 μA	-
ADC	900 μA	-
Timer	10 μA/MHz	-
UART	12 μA	-
Temperature Sensor	1,000 μA	-
ETc elements	685 μA	-

Table 2. Power Estimation: Sense and forward with sleep()

Items	**Time (μs)**	**Current (μA)**	**Amount (μs μA)**
Sensing Temperature	220	2,628	578,138
Scheduler Processing	1,250	7,727	9,658,750
Tx/Rx Turn-Around	232	19,578	4,542,073
CSMA-CA	312	20,728	6,467,105
Tx Data Frame	732	18,878	13,818,623
Rx Ack Frame	672	20,716	13,921,085
MAC Processing Time	218	4,977	1,084,986
System Active	3,636	-	50,070,759
System Inactive	996,364	716	713,296,988
Total	1,000,000	-	763,367,747

The lifetime of the system can be estimated by use of the current consumption quantity as set forth below.

$$Lifetime = \frac{Q_{battery}}{Q_{total}} \tag{2}$$

In order to estimate the lifetime of the system suggested in this paper, the current, which is needed for each part of the hardware, is listed in Table 1. Table 1 describes the current for each power mode, which is provided by the micro-processor, the current consumed in the RF transceiver, and the current, which is needed for each major hardware block of the system. In the scheduler having the simple FIFO, the MCU performed managing power with sleep(), which is converted into the minimum clock when it is in the inactive state. We experimented the sense and forward that the temperature value was measured from the sensor per second, and the measured temperature, 4Bytes (8 symbols) was transmitted from the non-beacon mode of IEEE 802.15.4 PHY/MAC to the PAN coordinator [5]. In this case, the tasks executed in the WSN node are

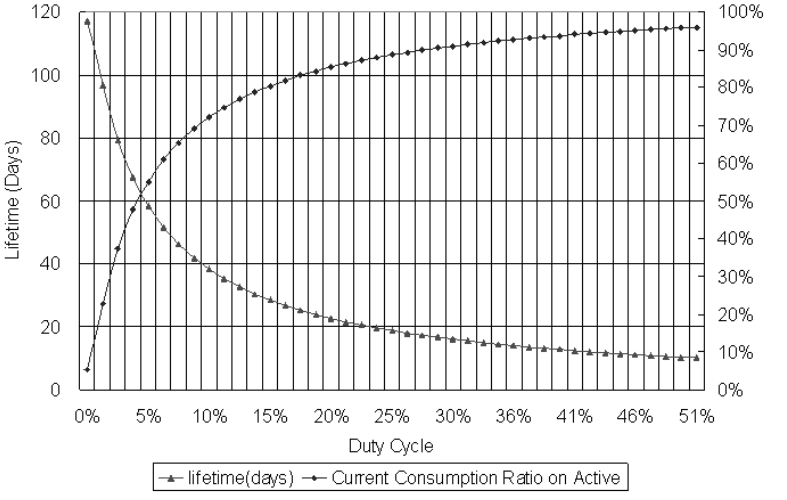


Fig. 4. Power Estimation Graph: Sense and forward with sleep()

divided into 7 types of tasks, which are listed together with the required time and current in Table 2. The tasks listed in Table 2 are classified by hardware blocks, and not tasks unit. The current in Table 2 refers to the sum of the current, which is required for the hardware blocks used in the corresponding tasks. The duty cycle is calculated when the system is in the active state for one second of the unit execution time and is approximately 0.25%. In other words, the system is in the active state during approximately 0.25% of its entire lifetime to execute the necessary tasks and is in the inactive state during 99.75% of its lifetime. When the system is in the inactive state, the lowest power of the system maintains. In that event, only 1011μA current is required. Assuming that two AA size batteries (3000mAhr) are used to maintain the system, the lifetime of the system would be approximately 117 days. Fig. 4 shows the lifetime of the system when the duty cycle increases up to 50% by increasing times for executing sense and forward per second.

An experiment was conducted on PA-DPS to achieve the sense and forward application in the same hardware. Table 3 provides the result of the experiment. In the experiment, the time and current were measured for each of the 7 tasks based on the hardware. When comparing Table 2 and 3, the PA-DPS has a more significant current consumption ratio than the sleep and wakeup in the active state. This result was attributable to applying the guard time for power_adjust() thereby increasing the total time in the active state. However, the current consumption was significantly reduced in the 7 tasks when the system is in the active state, because the PA-DPS converted all the inactive sections required for the tasks into the lowest power state through the power_adjust() process. The temperature sensor and peripherals such as the RF Tranceiver chip, etc.

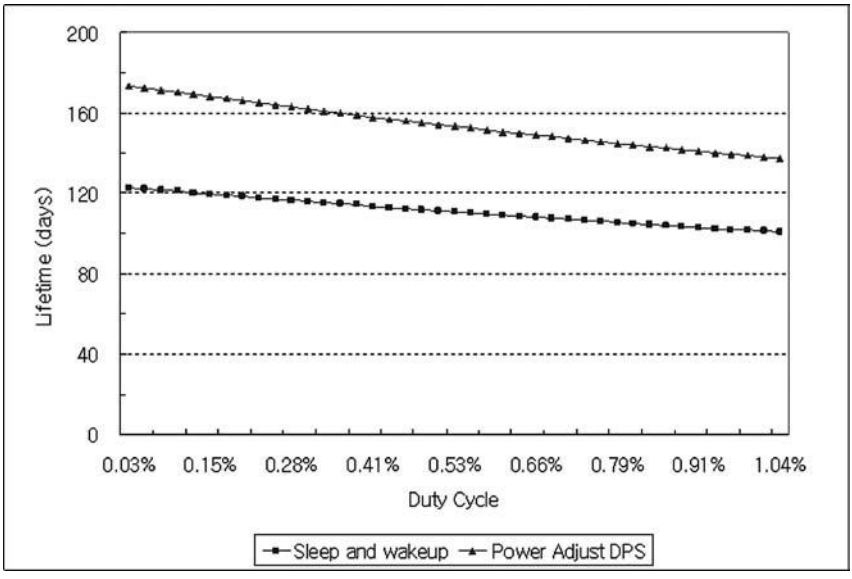


Fig. 5. Comparison: Lifetime under 1% duty cycle

Table 3. Power Estimation: Sense and forward with PA-DPS

Items	Time (μs)	Current (μA)	Amount (μs μA)
Sensing Temperature	200	10,569	2,113,800
Scheduler Processing	440	9,669	4,254,360
Tx/Rx Turn-Around	192	28,519	5,475,648
CSMA-CA	256	29,669	7,595,264
Tx Data Frame	672	27,369	18,391,968
Rx Ack Frame	672	29,669	19,937,568
MAC Processing Time	109	9,669	1,053,921
System Active	2,541	-	58,822,529
System Inactive	997,459	1,011	1,008,431,049
Total	1,000,000	-	1,067,253,578

were converted into the power-off state. In particular, the current consumption in the inactive state was reduced up to 30%, so that the entire current consumption was significantly reduced. The sleep and wakeup structure and the PA-DPS structure can be compared in terms of lifetime under 1% duty cycle as shown in Fig. 5.

5 Conclusion

In this paper, we proposed the PA-DPS (power adjust dual priority scheduling) structure. The PA-DPS has small size code, wherein the timer interrupt

have been minimized. This advantage is obtained from the existing event driven scheduling approach. In order to achieve low-power operation of the WSN node, it is most important to reduce the inactive mode current. When duty cycle is under 1% in WSN node, the PA-DPS can reduce the inactive mode current up to 30 40%. Further, the PA-DPS is simple in structure and assures the on-time execution.

Acknowledgments. This research was funded by MIC and IITA through IT Leading R&D Support Project. And, this research was supported by University ITRC Program funded by MIC of Korea and by DSRC Program funded by ADD Korea.

References

1. TinyOS, http://www.tinyos.net/
2. S. Bhatti, J. Carlson, H. Dai, J. Deng, J. Rose, A. Sheth, B. Shucker, C. Gruenwald, A. Torgerson, R. Han: MANTIS OS: An Embedded Multithreaded Operating System for Wireless Micro Sensor Platforms, ACM/Kluwer Mobile Networks & Applications (MONET). Special Issue on Wireless Sensor Networks Vol. 10. No. 4 (2005) 563-579.
3. IEEE Standard for Information Technology Telecommunications and information exchange between systems local and metropolitan area networks Specific requirements. Part 15.4: Wireless Medium Access Control(MAC) and Physical Layer(PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPAN)
4. Moncusi, M. A., Arenas, A., and Labarta, J.: Improving energy saving in hard real time systems via a modified dual priority scheduling. SIGARCH Comput. Archit. News 29 (2001) 19-24.
5. Kim K.H., Kopetz, H.: A Real-Time object Model RTO.k and an Experimental Investigation of Its Potentials. In Proc. Of the 18th IEEE Computer S/W & App. Conference (1994) 392-402
6. Kim, K.H.: Real-Time Object-Oriented Distributed Software Engineering and the TMO Scheme. Int'l Jour. of Software Engineering & Knowledge Engineering Vol. No.2, (April 1999) 251-276.
7. Kim, J. G., Kim, M. H., Shin Heu: Architectures and Functions of the TMO Kernels for ubiquitous & Embedded Real-time Distributed Computing. In Proc. of UIC (2006) 71-82
8. A. Mainwaring, J.Polastre, R. Szewczyk, D. Culler, J. Anderson: Wireless Sensor Networks for Habitat Monitorin. First ACM Workshop on Wireless Sensor Networks and applications (2002) 88-97.

WODEM: Wormhole Attack Defense Mechanism in Wireless Sensor Networks

Ji-Hoon Yun[1], Il-Hwan Kim[1], Jae-Han Lim[2], and Seung-Woo Seo[1]

[1] EECS, Seoul National University, Korea
{sjeus,ihkim}@ccs.snu.ac.kr, sseo@snu.ac.kr
[2] ETRI, Korea
ljhar@etri.re.kr

Abstract. The wormhole attack, which is accomplished by selectively relaying packets between two adversaries, can ruin routing and communication of the network without compromising any legitimate nodes. There have been a few countermeasures against the wormhole attack in generic ad hoc networks, but they are not appropriate for sensor networks since they require special devices (*e.g.* directional antenna) or put much overhead on each node. In this paper, we propose a new countermeasure against the wormhole attack for sensor networks, named WODEM. In WODEM, a few detector nodes equipped with location-aware devices and longer-lasting batteries detect wormholes, and normal sensor nodes are only required to forward control packets from the detector nodes. Therefore, WODEM is efficient in cost and energy. From the simulation results, we show that 10 detector nodes can detect a wormhole within the accuracy of 90% in a densely deployed sensor network.

Keywords: sensor network, security, wormhole attack.

1 Introduction

Recent advances of sensor and wireless communication technology have enabled us to apply wireless sensor networks to various applications such as military, commerce, environment and ubiquitous computing. In sensor networks, each node is normally limited on power, cost, size, processing capability and transmission capacity, thus a great deal of research efforts have mainly been focused on making the sensor networks feasible, practical and economical. Especially, because the limited energy is in close connection with the whole network lifetime, it is the most important challenge how to use the available energy efficiently. In addition to the power management issue, the sensor network has the similar inherent problem as other wireless networks: security. There are many applications that require secure communication and routing in sensor networks, such as military applications and confidential business operations. To make communications secure in sensor networks, many security protocols have been proposed to provide authenticity and confidentiality. For example, SPIN [1] provides a security architecture optimized for sensor networks with two building blocks: SNEP and TESLA. SNEP provides confidentiality, two-party data authentication, and

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 200–209, 2007.

freshness between nodes and a sink, while TESLA provides authenticated broadcast by introducing asymmetry through a delayed disclosure of symmetric keys. These security protocols make it difficult for an attacker to compromise legitimate nodes. However, a new type of attack called *wormhole attack* [2] can ruin routing and communication of the network without compromising any legitimate nodes.

1.1 Wormhole Attack

The wormhole attack is accomplished by two adversaries which simply relay incoming packets to the other adversary without decrypting or differentiating any packets. The two adversaries communicate with each other through a direct and dedicated channel by using a wired link or additional RF transceivers with a longer transmission range. The route via the wormhole looks like an attractive path to the legitimate sensor nodes because it generally provides less number of hops and less latency than normal routing paths. While relaying packets, the adversaries can arbitrarily drop the packets and thus data communications through the wormhole suffer from severe performance degradation. The wormhole attack can be a big threat to the security of sensor networks in two aspects: (1) it is difficult to detect the attack since the adversaries in the wormhole attack are normally invisible to the other legitimate nodes; (2) the wormhole attack can be accomplished easily because the adversaries of the attack do not need to compromise any legitimate nodes.

1.2 Previous Countermeasures Against Wormhole Attack

In order to cope with the wormhole attack, a few countermeasures have been proposed for mobile ad hoc networks. In packet leashes [2], the authors introduce geographical leashes and temporal leashes. The geographical leashes ensure that the recipient of the packet is within an allowed distance from sender. The temporal leashes restrict the lifetime of packets, which ensures that the packets have an upper bound of maximum travel distance. To implement these two schemes, the geographical leashes require that all nodes have a localization system like GPS, and the temporal leashes need accurate local clocks (*e.g.* cesium-beam clocks) and global time synchronization. In [3], Lingxuan and David propose a scheme using a directional antenna and ultrasonic signals. However, the schemes mentioned above require that each sensor node is equipped with special device(s) and thus they do not satisfy one of the most important design criteria of sensor networks: low cost. In LITEWORP [4], a node which resides simultaneously in the ranges of two neighboring nodes along a route is chosen as their guard. The guard monitors the traffic from both nodes and checks if one of them does not forward a packet from the other node. From monitoring, the guard can detect selective forwarding by the wormhole attack. LITEWORP operates without any special devices; however, nodes which are chosen as guards should monitor every data traffic received, which can be a big overhead on normal sensor nodes in processing and energy aspects. Moreover, LITEWORP cannot work with the

data aggregation in which data packets are aggregated and regenerated before being forwarded. Therefore, LITEWORP may not be appropriate for the protection against wormhole attacks in sensor networks either.

1.3 Contributions of This Paper

In this paper, we propose a new countermeasure against the wormhole attack, named WODEM (WOrmhole attack DEfense Mechanism), which is designed to satisfy the constraints of sensor networks. In WODEM, the node called detector is introduced and only this node is equipped with location-aware system (*e.g.* GPS) and longer-lasting battery. A pair of detectors exchange newly defined control packets and compare the traversed hop count with their distance. If the maximum possible distance for that hop count is shorter than the real distance, that means a wormhole exists along the route. After detecting the existence of a wormhole between the detectors, the detectors can find the exact position of the wormhole by adjusting their transmission range and TTL (time-to-live) of the control packets. In WODEM, all the processing required to detect the wormhole attack is performed by the detectors and sensor nodes only need to forward a few control packets from the detectors. Therefore, WODEM is an appropriate countermeasure for sensor networks in two aspects: (1) cost-efficient since only a few nodes require locaton-aware devices and longer-lasting batteries; (2) energy-efficient since normal sensor nodes do not require additional processing. Through ns-2 simulation, we show the wormhole attack detection probability of WODEM. For example, 10 detecting nodes are required to detect more than 90% of wormholes when sensor nodes with 150m transmission range are densely deployed over 1500m×300m rectangular space. We also provide the quantitative analysis on the effect of the wormhole attack on the network performance and demonstrate the efficiency of WODEM. One of the important findings through the simulation is that the wormhole attack cannot damage the network with local-repairing routing protocol if the wormhole drops traversing packets excessively.

The rest of this paper is organized as follows. In Section 2, we describe the system model considered in this paper. Section 3 explains our defense mechanism, WODEM, against the wormhole attack. We demonstrate the simulation results in Section 4. We conclude this paper in Section 5.

2 System Model

In this section, we describe the attack and network models to define the system considered in this paper.

2.1 Attack Model

The wormhole attack is launched by a pair of malicious nodes which are connected with each other via an out-of-band channel (*e.g.* high power transmission,

directional antenna or wired line). We assume that all the packets of the network are encrypted by a security protocol and thus the malicious nodes cannot compromise any legitimate sensor nodes. Instead, they just relay incoming packets from the sensor nodes selectively to each other without knowing the contents of the packets.

2.2 Network Model

We assume that the network is composed of many regular sensor nodes and some detectors, both of which are randomly distributed. Regular sensor nodes are MICA2 mote [5] class nodes with low-performance CPU and memory, spending less power. These nodes do not know their own locations. We assume that the transmission ranges of regular sensor nodes and malicious nodes are circular and the same with radius r. Each sensor node maintains its one-hop neighbor list. Sensor nodes send and receive packets only to and from their neighbors. Before detectors remove existing wormholes completely, there may be wrong neighbors in the list. Detectors know their location using location-aware systems. Detectors are assumed to have longer-lasting batteries than those of sensor nodes such that they are alive as long as the network is under utilization. We assume that the wireless channel is static and predictable by the log-distance path loss model [6], i.e., the relationship between the transmission power and its range. This assumption is reasonable because sensor nodes have no mobility basically. Under this assumption, each detector measures the path loss exponent of the path to its counterpart detector and it can control its transmission range from r to a sufficiently long distance by adjusting its transmission power. Each detector knows intermediate nodes along the route to its counterpart detector from the routing protocol which performs shortest hop routing. Each packet has a TTL field. The TTL value of each packet is decreased by one before the packet is forwarded to the next node. If the TTL value becomes zero, the node discards the corresponding packet without forwarding it.

3 WODEM

WODEM consists of three phases: detector scanning, wormhole detection, neighbor-list repair (shortly scanning, detection and repair, respectively). In the first phase, scanning phase, detectors scan their counterpart detectors to measure the path loss exponent of the wireless channel and to prepare for the detection phase. In the next detection phase, a pair of detectors detect the wormhole attack between them. If a wormhole is detected, the detectors start the repair phase where invalid neighbors in the neighbor lists will be removed. The detectors repeat the detection and repair phases until they cannot detect a wormhole anymore. We explain the operation of each phase in the following subsections.

3.1 Detector Scanning Phase

In the scanning phase, each detector scans its counterpart detector and measures the channel characteristics. To keep scanning secret, the scanning phase uses a separate channel different from the normal communication channel of the network so that control packets of the scanning phase do not traverse a wormhole. Therefore, all the detectors in the scanning phase are tuned to this channel. We define the detector which triggers the scanning phase as S-detector. The S-detector starts its scanning phase by broadcasting a *scanning packet* with TTL 1. The scanning packet contains the location L_S and transmission power level P_t of the S-detector. The S-detector repeats sending the scanning packet while increasing its transmission power by Δ_p in each transmission until it receives more than two replies from the other detectors. The reply packet contains P_t in the scanning packet and the location L_r of the replying detector. From the reply packets, the S-detector computes two characteristic parameters of its channel, i.e., path loss exponent n and constant k in the equation shown below:

$$P_t = k \times |L_S - L_r|^n \tag{1}$$

which is derived from the path loss model. The S-detector can obtain n and k since it has received more than two pairs of P_t and L_r. From now on, the S-detector can control its transmission range R_S by adjusting its transmission power to $k(R_S)^n$. If Δ_p is enough small, the granularity of P_t is also small and thus we get more accurate n and k, but should consume more energy and time to scan. Therefore, Δ_p should be determined by considering the measurement accuracy and efficiency simultaneously. After the computation, the S-detector chooses its counterpart detector, called *R-detector*, among the detectors which have replied to the scanning packets. Any detector can be either S-detector or R-detector.

3.2 Wormhole Detection Phase

In the detection phase, the S and R detectors check whether there is a wormhole between them. Let L_R be the location of the R-detector and H_{SR} be the hop count of the route from the S-detector to the R-detector. Then, the inequality below should always be true without any wormhole between two detectors:

$$H_{SR} \geq \min\{H_{SR}\} = \left\lceil \frac{|L_S - L_R|}{r} \right\rceil \tag{2}$$

where the right side is the minimum achievable number of hops between the S-R pair. If the above inequality is not true, it means that the hop count is reduced by a wormhole. Therefore, we can detect a wormhole between S-R pair by checking the above inequality. For this, the S-detector sends a *detecting packet* with a normal transmission range r. The detecting packet contains L_S and H_{SR}. Here, the TTL value of the detecting packet is set to be enough large for the

R-detector to receive the packet. When the R-detector receives the detecting packet, it checks the above inequality and knows whether there is a wormhole between the S-R pair. If a wormhole is detected, the R-detector notifies the S-detector of the existence of a wormhole on the route. Then, the S-detector receives this notification and enters the repair phase.

On the other hand, the detecting packet may traverse the wormhole and thus it can be dropped by the wormhole. Therefore, the R-detector needs to acknowledge the reception of the detecting packet regardless of the detection of the wormhole. If the S-detector does not receive an acknowledgement packet in a timeout time, it retransmits the detecting packet. This procedure is applied to all the exchanges of control packets between the S-R pair. The acknowledgement delay increases as the drop rate of the wormhole increases. However, the packet drop rate by the wormhole nodes will be deliberately controlled since, if the wormhole drops traversing packets excessively, packets will not traverse the wormhole anymore due to the route recomputation mechanism of the routing protocol.

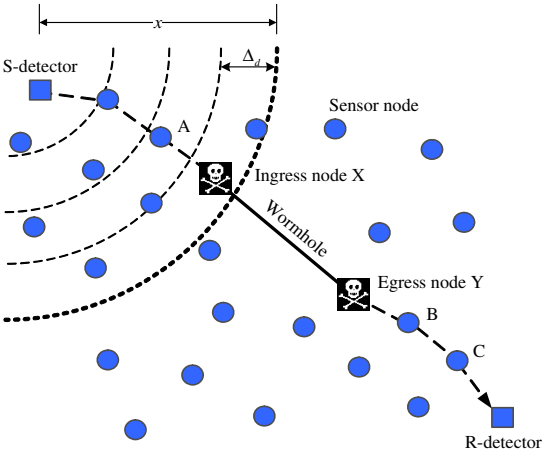


Fig. 1. The detectors can find node A and B when the S-detector sends a probing packet with the transmission range x and TTL 3

3.3 Neighbor-List Repair Phase

In the repair phase, the detectors find two sensor nodes in the route, between which the detected wormhole resides, and let them remove each other in their neighbor lists. For example, the goals of the repair phase in Fig. 1 are to find node A and B and let A remove B from its neighbor list, and vice versa.

The S-detector starts the repair phase by sending a *probing packet* to the R-detector with an initial transmission range r and TTL 1. The probing packet contains L_S, the transmission range of the corresponding transmission and the

initial TTL value. The S-detector repeats sending of probing packets increasing its transmission range by Δ_d in a stepwise manner until it receives a *probing reply packet* from the R-detector or the transmission range reaches the R-detector directly. If the S-detector still does not receive a probing reply packet from the R-detector after the above procedure, the S-detector resets the transmission range to r, increases the initial TTL value by 1 and repeats the above procedure.

For every probing packet received, the R-detector examines whether the packet is via the wormhole using the inequality below:

$$(\text{Initial TTL}) \geq \min\{H_{SR}\} \tag{3}$$

where the left side means the maximum number of hops which the probing packet can travel and the right side is the minimum valid number of hops of the route between the S-R pair. If the inequality is false, that means the packet traversed shorter hops than the minimum valid number of hops, which is possible only by traversing the wormhole. Here, $\min\{H_{SR}\}$ is obtained by:

$$\min\{H_{SR}\} = \left\lceil \frac{|L_S - L_R| - R}{r} \right\rceil + 1 \tag{4}$$

where R is the transmission range of the S-detector when it transmitted the corresponding probing packet. If the received probing packet does not satisfy the inequality, the R-detector sends a probing reply packet to the S-detector. The probing reply packet contains the initial TTL value of the probing packet which triggers the reply.

From the operations of the S and R detectors, the R-detector sends a probing reply packet to the S-detector for the first time when the transmission range of the S-detector becomes the distance between the S-detector and the ingress wormhole node and the initial TTL of the probing packet becomes the number of hops from the egress wormhole node to the R-detector. Here, the ingress wormhole node is defined as one adversary of the wormhole which is close to the S-detector along the route of the S-R pair, and the egress wormhole node as the other adversary. In Fig. 1, when the initial TTL of probing packets is 2, the R-detector cannot receive any probing packet if the transmission range R_S of the S-detector dose not cover the node C. If R_S begins to cover C, the R-detector receives probing packets; however, the R-detector does not reply to them since the inequality (3) is true for the received probing packets. When the initial TTL of probing packets become 3, the R-detector cannot receive any probing packet if R_S does not cover the ingress wormhole node X. Finally, if R_S begins to cover X, the R-detector receives the probing packet and the inequality (3) becomes false for the first time. Then, the R-detector sends a probing reply packet to the S-detector.

When the S-detector receives a probing reply packet for the first time, it can infer each sensor node neighboring the ingress and egress wormhole node, respectively, from the initial TTL value of the probing packet which triggered the probing reply packet, i.e., the (TTL-1)-hop previous node from the R-detector

in the route is the neighbor node of the egress wormhole node and the TTL-hop previous node is the neighbor node of the ingress wormhole node. Then, the S-detector sends *alarming packets* to both of the sensor nodes, which let them remove each other in their neighbor lists.

Although a pair of sensor nodes correct their neighbor lists in the repair phase, a new route from the S-detector to the R-detector may traverse the wormhole. Therefore, the S-R pair repeats the procedures from the detection phase until they cannot detect a wormhole anymore.

4 Simulation Results

In order to investigate the effect of the wormhole attack and the detection probability of WODEM, we conduct simulation using the ns-2. Our simulation is conducted over a 1500m×300m rectangular flat space with randomly distributed sensor nodes for 200 seconds. We choose a longer rectangular space in order to enforce the use of longer routes between nodes. A sensor node has a circular radio range of 150m radius and uses the IEEE 802.11 distributed coordination function (DCF) for the MAC layer protocol. Sensor nodes generate constant bit rate (CBR) traffic to the sink at the corner. The packet size is fixed as small as 64 bytes and the packet-sending rate of a source node is limited to one packet per second. For the wormhole attack, a pair of adversaries is deployed making a wormhole by a wired link, which randomly drops traversing packets according to the given drop rate. We use AODV [7] for the routing protocol since AODV is the basis of many routing protocols for sensor networks, e.g., ZigBee [8], which is a promising technology as the framework of sensor networks and WPANs (Wireless Personal Area Networks), also adopts AODV-like routing protocol.

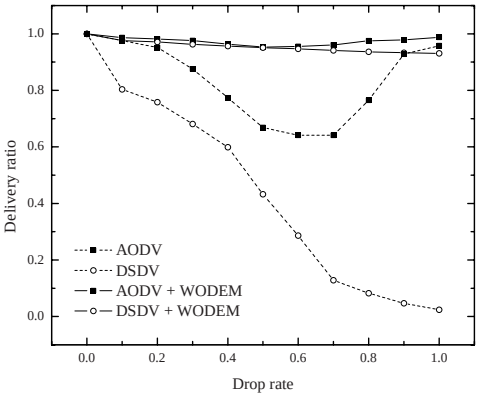


Fig. 2. Effect of wormhole attack as the drop rate varies

Fig. 2 shows the packet delivery ratios of the network with and without WODEM when 200 sensor nodes are deployed. Here, the packet delivery ratio is defined as the ratio of the number of packets received by the sink to the number of packets generated by the sensor nodes. In this simulation, we also consider DSDV [9] for the routing protocol in order to investigate the effect of routing protocol in the wormhole attack. Without WODEM, the network performance deteriorates more and more as the drop rate of the wormhole increases until 0.6 for both routing protocols. On the other hand, when AODV is used, the delivery ratio recovers as the drop rate is higher than 0.7 due to its local repair algorithm. As a result, it is noted that excessive drop rate may not damage the network if the routing protocol of the network has a local routing recovery mechanism. However, AODV cannot be a countermeasure against the wormhole attack. With the drop rate of 0.6, a single wormhole can degrade the delivery ratio of the network by as much as 30% even with AODV. With WODEM, the wormhole cannot damage the network at all for both routing protocols regardless of the drop rate.

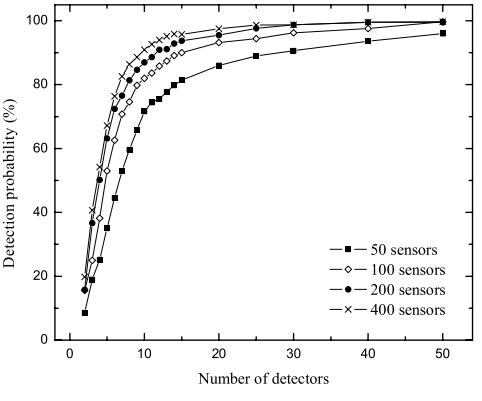


Fig. 3. Detection probability of WODEM vs. the number of detectors

Fig. 3 shows the detection probability as the number of detectors varies. The average detection probability increases rapidly with the increasing number of detectors. Note that with more than 400 sensor nodes, only 10 detectors are sufficient to detect the wormhole within the accuracy of 90%. When the density of sensor nodes is low, more detectors are needed because the probability that the number of hops along the illegal route through the wormhole is larger than the ideal minimum hop count becomes high.

Fig. 4 shows the simulation result of the relation between the detection probability and the distance L_w between two adversaries of a wormhole with 200 sensor nodes. From the figure, the longer L_w is, the higher the detection probability is. That is because it is more probable that the number of hops of the illegal route is smaller than the ideal minimum hop count.

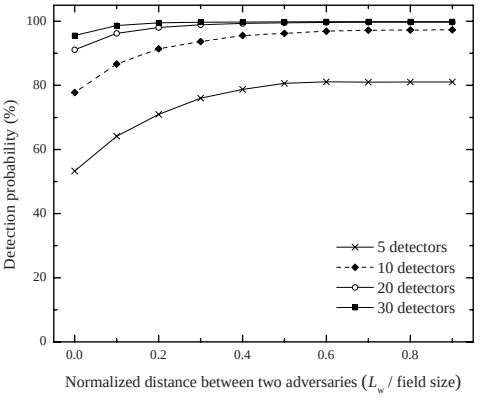


Fig. 4. Detection probability vs. wormhole length L_w (field size = 1500m)

5 Conclusions

In this paper, we showed that WODEM can defend the network efficiently against the wormhole attack. Furthermore, WODEM is shown to be suitable for sensor networks because it is cost effective by minimizing the use of location-aware equipments and tight synchronized clocks, and power efficient due to small processing overhead on normal sensor nodes. On the other hand, the effect of power management in sensor networks is not investigated in this paper, which needs further researches about that.

References

1. A. Perrig *et al.*: SPINS: Security protocols for sensor networks, in Proceedings of Mobile Networking and Computing 2001, 2001.
2. Y. C. Hu *et al.*: Packet leashes: a defense against wormhole attacks in wireless networks, IEEE INFOCOM 2003, Volume: 3, 2003.
3. Lingxuan Hu and D. Evans: Using Directional Antennas to Prevent Wormhole Attacks, The 11th Annual Network and Distributed System Security Symposium, San Diego, California, February 2004.
4. I. Khalil *et al.*: LITEWORP: A Lightweight Countermeasure for the Wormhole Attack in Multihop Wireless Networks, International Conference on Dependable Systems and Networks (DSN) 2005, June 2005.
5. MICA2 series, http://www.xbow.com.
6. Theodore Rappaport: Wireless Communications: Principles and Practice, Prentice Hall PTR, 2001.
7. Charles E. Perkins: Ad hoc On-Demand Distance Vector (AODV) Routing, RFC 3561, IETF, July 2003.
8. ZigBee Specification, Ver. 053474r10, ZigBee Alliance, July, 2006.
9. Charles E. Perkins: Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers, ACM SIGCOMM'94, 1994.

Layer-Based ID Assigning Method for Sensor Networks

Jung Hun Kang and Myong-Soon Park

Department of Computer Science and Engineering, Korea University
Seoul, 136-713, Korea
{kjhoun, myongsp}@ilab.korea.ac.kr

Abstract. A sensor network consists of a set of battery-powered nodes, which collaborate to perform sensing tasks in a given environment. Globally unique ID allocation is usually not applicable in a sensor network due to the massive production of cheap sensor nodes, the limited bandwidth, and the size of the payload. However, locally unique IDs are still necessary for nodes to implement communications to save energy consumption. Already several solutions have been proposed for supporting locally unique ID assignment in sensor networks. However, they bring much communication overhead, and they are complex to implement. We present a layer-based algorithm to solve the unique ID assignment problem. This algorithm can save energy consumption by reducing communication overhead while IDs are assigned.

1 Introduction

A sensor network consists of a large number of sensor nodes simultaneously engaged in environment monitoring and wireless communications. In traditional distributed systems, the name or address of a node is independent of its geographical location and is based on the network topology. However, in sensor networks, it has been widely proposed to use attributes external to the network topology and relevant to the application for low-level naming (1). Solutions gradually make some neighbor nodes into a group and simultaneously assigns the unique ID to each node. Unique IDs are assigned by a header node of each group. Globally unique IDs are useful in providing many network functions, e.g. configuration, monitoring of individual nodes, and various security mechanisms. ID conflict problem is a major issue of the ID assignment in sensor networks (2). In Figure 1, nodes $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ are connected to each other. Nodes $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same ID of $\boldsymbol{x}$, node $\boldsymbol{C}$ has a different ID of $\boldsymbol{y}$. If node $\boldsymbol{A}$ wants to send a packet to node $\boldsymbol{B}$, a traditional network layer protocol usually considers the packet destined for itself and will not deliver the packet to the underlying data link layer because the destination has the same address of the source. If node $\boldsymbol{C}$ wants to send a packet to either node $\boldsymbol{A}$ or node $\boldsymbol{B}$, because they both have the same address, both will receive the packet and process it, which will waste power. Thus, how uniquely and efficiently assigning ID in sensor network is the biggest issue.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 210–219, 2007.

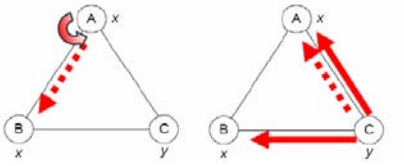


Fig. 1. ID Conflict Problems in Sensor Network

An obvious ID assignment strategy is to have each node randomly choose an ID such that the probability of any two nodes choosing the same ID is very low. However, for this probability to be low, we need the IDs to be very long, which is again costly in terms of energy (3). Any ID assignment solution should produce the shortest possible addresses because sensor networks are energy-constrained. The usage of the minimum number of bytes required in motivated by the need to limit the size of transmitted packets, in particular the header. In fact, communication is usually the main source of energy drain in sensor node (4).

This paper proposes a layer-based ID assignment, which is an efficient ID assignment in a sensor network. This paper is constructed as follows: Section 2 introduces work related to ID assignment in sensor networks. Layer-based ID assignment is distributed in Section 3. Then the efficiency of the scheme is supported by simulation results in Section 4. Finally, Section 5 concludes the paper and future works.

2 Related Work

In general, network-wide unique addresses are not needed to identify the destination node of a specific packet in sensor networks. In fact, attribute-based addressing fits better with the specificities of sensor networks (5). In this case, an attribute such as node location and sensor type is used to identify the final destination. However, different nodes can have the same attribute value, in particular in the same neighborhood. Thus, there is a need to uniquely identify the next hop node during packet routing. Several schemes have been proposed to assign locally unique addresses in sensor networks.

In (4), Schurgers, et al., developed a distributed allocation scheme where local ad-dresses are spatially reused to reduce the required number of bits. The preexisting MAC addresses are converted into locally unique addresses. Each locally unique ad-dress is combined with an attribute-based address to uniquely determine the final destination of a packet. This use of locally unique addresses instead of global ad-dresses does not affect the operations of the existing routing protocols. This solution assumes the pre-existence of globally unique addresses, which is not realistic in the case of sensor networks.

The scheme proposed in (6) utilized a proactive conflict detection method for a general sensor network, including a mobile sensor network, and a stationary

sensor network with new members joining. When a node boots up, it first chooses a random physical address and then announces it with periodic broadcasts of HELLO messages with the interval of 10 seconds. All the nodes record the source address of the HELLO message in a neighbor table, which is included in the subsequent HELLO messages. Therefore, every node will have 2-hop neighbor information, which is utilized to re-solve address conflicts among 2-hop neighbors. If a node finds that one of its neighbors chooses a duplicate address, it will notify this neighbor to change the ad-dress. Reactive ID Assignment (2) is introduced next. This algorithm defers ID conflict resolution until data communications are initiated. It leads to save communication overhead. However, every node can not choose a random ID in the beginning. Sensor network is getting enlarged; the number of communication is being increased extremely. In addition, many kinds of messages make this algorithm more complex.

As far the globally unique ID assigning scheme, Distributed ID Assignment is introduced in (7). In order to assign ID, Tree structure is used to compute the size of the network. Then Unique IDs are assigned using the minimum number of bytes. However, this scheme uses not only assigning temporary ID and final unique ID but also obtain-ing sub-tree size. In order to assign temporal ID and final unique ID, high communication cost is needed. In (8), Ali, et al., proposed an addressing scheme for cluster-based sensor networks (9). To prevent collisions, nodes within the same cluster are assigned different local addresses. Non-member one-hop and two-hop neighbors must also have different local addresses to avoid the hidden-terminal problem. The network is divided into hierarchical layers where the number of layers increases with the number of nodes in the network. Global IDs are obtained by putting the local address and the addresses of the head nodes of the different layers together. This solution suffers from the fact that the address size increases with the number of layers as 6 bits are added for each layer. However, this makes this solution less attractive due to the energy cost of using global IDs in the case of large sensor networks. In addition, this solution can be used only with cluster-based routing and does not extend to the case of multi-hop routing (10).

3 Layer-Based ID Assignment Algorithm

We proposed a layer-based ID assignment algorithm that assigns globally unique IDs to sensor nodes. In this section, the assumptions for our proposed ID assignment scheme are given first, and then the message types and proposed algorithms are de-scribed in detail.

3.1 Assumption

Initially, we define some assumptions as below

(1) The nodes in a sensor network are usually manufactured in batches.
(2) Neighbor node IDs must be stored in the memory of the sensor node during all its lifetime.

(3) ID assigned field is composed of 3 parts: Group ID, Section ID, and Node ID. (For example, assigned ID, 0123 means Group ID (01), Section ID (2), and Node ID (3)).

3.2 Message Types

Total 4 kinds of messages are used to assign IDs in each node.

(1) ***Layer1_SEARCH*** message: Within 1-hop, a Sink node or a Header node searches neighbor nodes that have no assigned ID. After collecting neighbor node Information, it assigns sequence ID to the searched neighbor nodes.
(2) ***Layer2_SEARCH*** message: Within 2-hops, the sink node lets the assigned neighbor nodes to search the other neighbor nodes that have no assigned ID. After collecting neighbor nodes' information, it assigns sequence ID to the searched neighbor nodes.
(3) ***Child_GROUPING*** message: The sink node can make extended other groups by unicasting this message to ID assigned border node in layer2. The border node broadcasts this message to neighbor nodes and chooses one node (by fastest responding time) to make it as a header node.
(4) ***Sink_REPORTING*** message: All header nodes can send the grouping information and ID assigning status to the sink node at the end of assigning ID task in each header.

3.3 Grouping and ID Assigning Methods

In order to assign globally unique IDs to each node, we divided the proposed ID assignment scheme into two parts: Parent grouping algorithm and Children grouping algorithm. They assign globally unique IDs to each node while they build groups. Firstly, the Parent grouping algorithm takes roles of building core group and as-signing IDs to neighbor nodes from the sink node. In order to expand children groups, these assigned IDs are working as a message forwarder. The Children grouping algorithm takes roles of building expanded groups and assigning ID globally. In each group, the sink node sets a header node as a sub-sink node to broadcast messages and collect information instead of the sink node.

In the Parent groping algorithm, the sink node builds the 1st layer area within 1-hop range by broadcasting ***Layer1_SEARCH*** message. This makes a 1-hop core area from the sink node. ID of sink node is set as 0000. In this core area, node members assign their ID as 0001, 0002 by according to the responding time. The second layer areas are expanded via 1st layer members when the sink node broadcasts ***Layer2_SEARCH*** message to those members. Up to the member nodes' ID, section IDs are decided in nodes of the second layer. After assigning IDs of member nodes in the second layer, the first layer members report their ID assigned status to the sink node. Via these organized group members, the sink node lets them forward message to header nodes in each group. This makes the

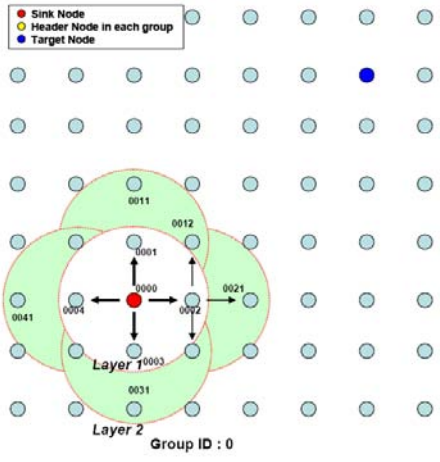


Fig. 2. ID Assignment in Parent Grouping Algorithm

Algorithm 1: **Parent Grouping Algorithm.**
Step 1. The sink node broadcasts ***Layer1_SEARCH*** message.
Step 2. The Sink node assigns the sequential IDs to the found neighbor nodes by the responding time.
Step 3. The Sink node transfers ***Layer2_SEARCH*** message to layer1 nodes
Step 4. Layer1 nodes broadcast received ***Layer2_SEARCH*** message to 1-hop neighbor nodes
Step 5. Layer1 node assigns sequential ID for layer2 nodes (upon the responding time)
Step 6. Layer1 nodes report ID-assigning results to the sink node by sending ***Sink_REPORTING*** message

sink node know how many members are assigned, and who can be a message forwarder by sending ***Sink_REPORTING*** message to the sink node. Steps of ID assigning in Parent grouping algorithm are illustrated in Figure 2 as well.

Algorithm 2: **Children Grouping Algorithm.**

Step 1.
The Sink node unicasts ***Child_GROUPING*** message to a specific border node in layer2.

Step 2.
The layer2 node chooses one node which has the fastest responding time among the neighbor nodes. The layer2 node broadcasts ***Child_GROUPING*** message to the node chosen.

Step 3.
Group ID and 00(header ID) are assigned to the chosen node as a header node.

Step 4.
The header node (Children-header node) broadcasts ***Layer1_SEARCH*** message to neighbors in a 1-hop distance.

Step 5.
The chosen layer1 nodes broadcast ***Layer2_SEARCH*** message to set layer2 nodes.

Step 6.
The header node reports ID-assigning result to Sink node by sending ***Sink_REPORTING*** message.

Children grouping algorithm starts with unicasting ***Child_GROUPING*** messages from the sink node to its second layer members. Member nodes, in the second layer, choose one node which has the fastest responding time among the neighbor nodes. And then, they broadcast the ***Child_GROUPING*** message to the selected node. After receiving this message, this selected node can be a header in a children group. With the given group ID, the header ID (00) is assigned to the chosen header node as a root of a children group. The given group ID is decided by the sink node. By up to the time of creating group, the sink node decides the ID sequentially. This header node broadcasts message to build the 1st layer area of the children group. And left processes are the same as the Parent grouping algorithm. When the other children group is extended, the intermediate header nodes record the IDs of new created header nodes in newly created children groups by the previous header node, and then report the assigned ID status to the sink node. Described steps of Children grouping algorithm are illustrated in Figure 3.

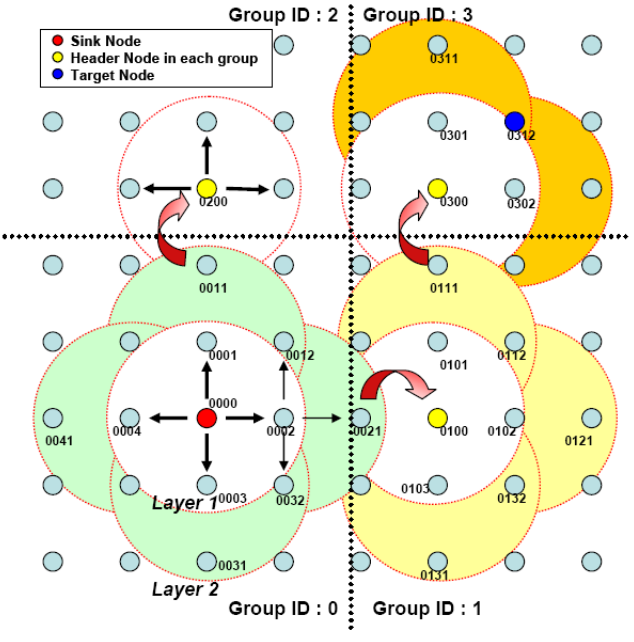


Fig. 3. ID Assignment in Children Grouping Algorithm

4 Simulation Results and Analysis

The simulations are implemented to compare their performance in the NS-2 network simulator (Version 2.27) with the modification of directed diffusion module. The sensor nodes are placed in a 13 x 13 grid for a stationary sensor network. The distance between two nodes is 100 meters so that a node in the middle of the network has 4 direct neighbors. The size for the address has 4 digits. The number of group IDs can be assigned up to 99. Also the number of section IDs can be assigned up to 9.

4.1 Communication Overhead

Figure 4 shows the sum of received message packets at all the nodes when assigning IDs globally. We set the node density as 100, 169, 225, and 324. In the case of proactive case, it broadcasts periodic *HELLO* messages including its neighbor table. In every transmission, each node broadcast the periodic *HELLO* message to each other to assign ID. This causes extremely high communication overhead. Reactive scheme broadcasts the *HELLO* message in the end of the simulation to build the neighbor table for analysis. This broadcasting message cause much lower communication over-head than the proactive scheme. However, in order to avoid the ID conflict problem, this scheme keeps going on sending *CHANGE*

messages to each other. This overhead is quite high. In the layer-based ID assignment scheme, communication overhead is much lower than the other two schemes because in each group, when ID is assigned to members of groups, it locally communicates the assigning message to member nodes. Also in the header node, they target only a few member nodes to assign ID in each group. Indeed, only reporting messages and control messages such as ***Child_GROUPING*** message are transferred to the sink node. The simulation shows the comparison among those three schemes in Figure 5. According to the simulation result, the proposed scheme causes much lower communication overhead than the other two schemes; proactive and reactive ID assignments.

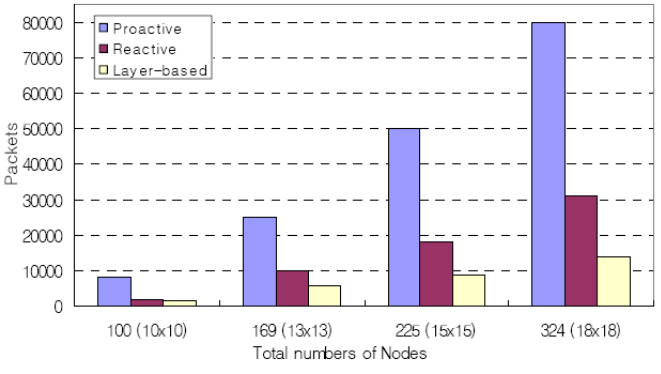


Fig. 4. Communication Overhead

4.2 Energy Consumption

We compare the energy consumption when ID is assigned in each node. The energy consumption is measured by the sum of energy consumed by all the sensor nodes on the data transmission when IDs are assigned. In order to evaluate the energy consumption, we set parameter values as Table 1.

In Figure 5, it shows the comparison result of average values in energy consumption. Proactive scheme causes a lot of message transmission because of the *HELLO* message transmission. In the reactive ID assignment scheme, in order to avoid the ID conflict problem, nodes communicate messages many times. This causes high energy consumption in transferring messages. In contrast, this layer-based ID assignment algorithm consumes lower energy than the reactive scheme because of the globally unique ID. When the sink node or the header node assigns IDs to other nodes, it communicates messages with only a few nodes. And in each children group, each header controls to assign IDs to neighbor nodes locally. This causes the total energy consumption lower. Thus, layer-based ID assignment scheme saves both bandwidth and power more than 25%. Furthermore, in each group, the header node takes ID assigning task instead of the sink node which leads to a shorter length of finishing time and a less power consumption to assign IDs to all nodes.

Table 1. Parameter Settings in Simulation

Parameter Name	**Value**
Radio Bandwidth	20Kbps
Radio Transmission Range	100m
Packet Length	10bytes
Transmit Power	8.2mA
Receiver Power	4.6mA

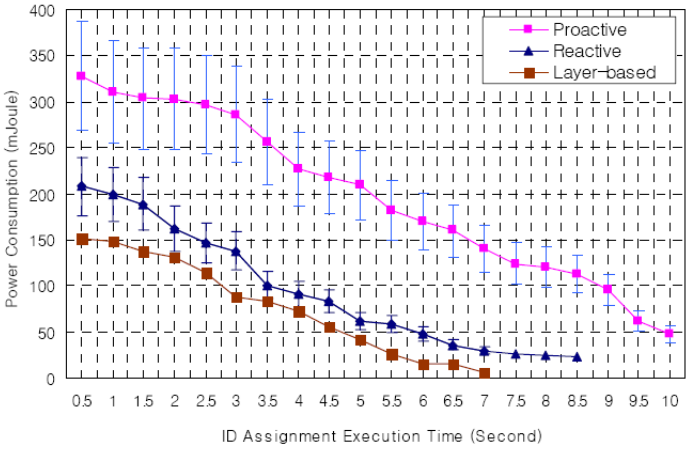


Fig. 5. Energy Consumption

4.3 Analysis

The total amount of energy consumption of the layer-based ID assignment scheme can be described as Equation 1. When each parameter is defined as below: x is the number of the sink node, the header node i, the number of created second layer groups can be j, total number of all created group is $\hat{t}_x$, the total amount of nodes in group is t, $P_{x,i,i}(t)$ is the total amount of communication cost in the first layer, and $P_{x,i,j}(t)$ is the total amount of communication cost in the second layer.

$$E(x, \hat{t}_x) = \sum_{t=\hat{t}_1}^{\hat{t}_x} \left[\Delta_i(t) P_{x,i,i}(t) + \sum_{j \in K(x)} \Delta_j(t) P_{x,i,j}(t) \right] \ldots (1)$$

5 Conclusion

In this paper, we presented a solution to the globally unique ID assignment problem in sensor networks. Our proposed algorithm aims at assigning globally unique IDs to each node by using two grouping algorithms. Through these two

grouping algorithms, it forms two-layer groups. In each group, headers take roles of the sink and they assign neighbors' IDs instead of sink node. The sink node can not only easily assign IDs to all other nodes via header nodes, but also saves the energy consumption up to approximately 25% according to the simulation. The challenge for our future work is to establish an efficient routing architecture based on our scheme in sensor networks. In addition, we will study the global group-ing management and the grouping resilience in the sensor networks.

Acknowledgements. This work was supported by the Second Brain Korea 21 Project. Dr. Myong-Soon Park is the corresponding author.

References

[1] J. Heidemann, F. Silva, C. Intanagonwiwat, R. Govindan, D. Estrin, and D. Ganesan. "Building Efficient Wireless Sensor Networks with Low-Level Naming," In Proceedings of the Symposium on Operating Systems Principles, pp. 146-159. Chateau Lake Louise, Banff, Alberta, Canada, ACM. October, 2001.

[2] H. Zhou, M. W. Mutka, and L. M. Ni, "Reactive ID Assignment for Sensor Networks," In Proceedings of IEEE International Conference on Mobile Ad-Hoc and Sensor Systems, Washington DC, November 2005.

[3] J. R. Smith, "Distributing identity," IEEE Robotics and Automation Magazine, Vol.6, No.1, March 1999.

[4] C. Schurgers, G. Kulkarni, and M. B. Srivastava, "Distributed On-demand Address As-signment in Wireless Sensor Networks," In Proceedings of IEEE Transactions on Parallel and Distributed Systems Vol.13, October 2002.

[5] D. Estrin, J. Heidemann, and S. Kumar, "Next century challenges: Scalable coordination in sensor networks," in Proceedings of MOBICOM,pp. 263-270, 1999.

[6] C. Schurgers, G. Kulkarni, and M. B. Srivastava, "Distributed Assignment of Encoded MAC Address Assignment in Wireless Sensor Networks," In Proceedings of MobiHOC 2001, Long Beach, CA, October 2001.

[7] E. Ould-Ahmed-Vall, D. M. Blough, B. S. Heck and G. F. Riley, "Distributed Unique Global ID Assignment for Sensor Networks," In Proceedings of IEEE International Con-ference on Mobile Ad-Hoc and Sensor Systems, Washington DC, November 2005

[8] M. Ali and Z. A. Uzmi, "An energy efficient node address naming scheme for wireless sensor networks," in Proceedings of the International Networking and Communications Conference (INCC), 2004.

[9] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless micro sensor networks," IEEE Transactions on Wireless Communications, Vol.1, No.4, October 2002.

[10] W. B. Heinzelman, J. W. Kulik, and H. Balakrishnan, "Adaptive protocols for informa-tion dissemination in wireless sensor networks," in Proceedings of MOBICOM, 1999.

A Smart Sensor Overlay Network for Ubiquitous Computing*,**

Eui-Hyun Jung[1], Yong-Pyo Kim[2], Yong-Jin Park[2], and Su-Young Han[3]

[1] Dept. of Digital Media, Anyang University, 708-113, Anyang 5-dong, Manan-Gu, Anyang City, Kyounggi-do, 430-714, Korea
jung@anyang.ac.kr
[2] Dept. of Electronic and Computer Engineering, Hanyang University, Haengdang-dong, Sungdong-gu, Seoul, 133-791, Korea
{ypkim, park}@hyuee.hanyang.ac.kr
[3] Dept. of Computer Science, Anyang University, Samseong-ri, Buleun-myeon, Ganghwa-gun, Incheon, 417-833, Korea
syhan@anyang.ac.kr

Abstract. Sensor networks have been considered as a base technology for constructing ubiquitous computing environment. However, most researches about sensor networks have not focused on ubiquitous computing but mainly focused on efficiency of sensor networks itself. Ubiquitous computing requires sensor networks to be accessed in a transparent and abstract manner from the outer world. However the requirements cannot be easily achieved because of physical restrictions of sensor networks and centralized data processing scheme. In this paper, we have designed a sensor overlay network that enables ubiquitous computing applications to consider sensor networks as an ordinary component of ubiquitous computing environment. Virtual sensors and actuators on the overlay network cooperate with each other and provide intelligent decision that was impossible to be implemented with current physical sensor networks. The proposed overlay network also provides dynamic service composition and device integration essential to ubiquitous computing. To evaluate the usefulness of the designed system, the overlay network is implemented and simulated on J-Sim simulator.

1 Introduction

Due to the rapid development of computing and network technologies, ubiquitous computing that enables users to enjoy computing and communication at anytime and anywhere has been emerged [1][2]. The objective of ubiquitous computing is to have devices sense changes in their environment and automatically provide computing services based on the preference of users [3][4][5]. These researches

* This work was supported by Korea Research Foundation Grant funded by Korea Government (KRF-2006-331-D00365).
** A part of authors who work for this research are supported by BK21's research funding.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 220–231, 2007.

have a common goal to recognize user context and provide personalized service in a transparent way.

To provide ubiquitous computing, several technological issues should be resolved such as Heterogeneity, Invisibility, Expandability, Autonomy and Context-Aware/management [2][3]. One of the most fundamental technologies to fulfill these issues is collecting environmental data about users' current state and surroundings [4]. Sensor networks have been considered as a prominent technology for this purpose. Sensor networks made up of a lot of wired or wireless sensors can provide an effective method to collect data about users and their surroundings [6]. However, the structure and functions of sensor networks to support ubiquitous computing environment are not clearly defined and not deeply researched by now [6]. Researches on sensor networks have focused on energy-efficient routing protocol and OS for sensor nodes. There are also researches on middleware such as SensorWare [7] and Cougar [8]. However, the purpose of these middleware researches is not for ubiquitous computing, but to provide unified execution environment for heterogeneous sensors or reprogramming for dynamic update.

Seamlessly to be integrated to ubiquitous computing environment, sensors on sensor networks should be able to be accessed via transparent and abstract manner. However, this requirement can not be achieved with current sensor networks schemes because it is very difficult and not desirable to communicate directly to an individual sensor on sensor networks. To make the matter worse, current sensor networks assume a centralized information processing architecture for their data processing; sensors collect data, sensed data are delivered to a central unit, the central unit performs processing and issues commands back to sensors and actuators. Ubiquitous computing needs cooperation and rapid data fusion among sensors for dynamic service composition, whereas, in the centralized architecture, sensor networks are considered as a set of dummy data collectors, so these can not be integrated components accessed independently from various ubiquitous computing applications.

Another issue is the extensibility when a new kind of devices is dynamically added in ubiquitous computing environment. Introduction of a new kind of devices will cause modification of runtime processing code of the central unit, but conventional hard-wired coding cannot provide the flexible extension and integrations with existing modules in the central unit. For these reasons, it is desirable that basic processing for ubiquitous computing is performed on sensor networks autonomously, but physical restrictions on sensor nodes prohibits achieving this goal.

This paper proposes a smart sensor overlay network that supports cooperation among sensors and data fusion by locating virtual counterparts of physical sensors and actuators in the sink node. The proposed overlay network can provide ubiquitous computing service by applying virtual counterparts in the sink node without any centralized computing support. Moreover, when a new kind of device is applied, the proposed structure enables integration only by loading virtual counterpart of that device into the sink node. To verify the proposed structure, we implemented the overlay network on the J-SIM simulator. The

simulation results showed that the proposed system operates properly according to changes in environment and addition of a new kind of devices.

The remainder of this paper consists of four subsections. Section 2 explains existing researches on sensor networks and several issues when sensor networks are applied to ubiquitous computing. Section 3 describes the concept of virtual counterpart and the structure of implemented system. Section 4 evaluates proposed structure with J-Sim [9] simulator. Finally, section 5 concludes the paper.

2 Issues of Ubiquitous Computing Based on Sensor Networks

2.1 Current Researches

Researching fields on sensor networks by now can be categorized into routing and operating system. Researches on routing have focused on network lifetime enhancement by energy-efficient routing schemes [10]. Operating systems for sensor networks have been researched to load lightweight and energy-efficient functions to sensor nodes [11]. Since sensors have limits in energy and computing power, energy-efficiency has been a main issue in both researching fields. Compared to these researches, there are few researches on the information system of sensor networks. Since most researches regard sensor networks as device networks simply collecting data rather than an information system, researches by now mainly focused on energy-efficient transmission of collected data from sensors to the central system. To use sensor networks as an information infrastructure, several researches on the middleware are ongoing recently in the view of dynamic code update by reprogramming in sensor nodes and data centric routing protocol. Researches such as SensorWare and Cougar are representatives of researches on the middleware for sensor networks. In the SensorWare, a middleware executing Tcl script is implemented on each sensor node for supporting dynamic code update. Sensor application programs written in Tcl script are dynamically loaded to the middleware. On the other hand, Cougar focuses on data centric processing. Sensor networks are assumed as a large distributed database in the Cougar. Query Proxy Layer analyzing SQTL sentence is equipped as a middleware in each sensor node and various INTERESTs can be processed using SQTL in real-time. In the view of applying the middleware in sensor networks, these researches are meaningful. However, researches on the sensor middleware by now have been focused on providing identical execution environment or reprogramming, so they are far from the middleware that can provide ubiquitous computing services based on sensor cooperation and data fusion.

2.2 Technological Issues

To utilize sensor networks in ubiquitous computing, it is important not to deal with sensor networks as dummy data gathering networks. Usually, ubiquitous computing requires rapid data fusion and dynamic service composition. To fulfill these requirements, each sensor node should be able to easily communicate with

other sensor nodes and issue commands directly to actuators in the sensor field. However, most current researches on sensor networks assume sensor networks as a set of dummy devices. This assumption inevitably requires a centralized information processing architecture in which a central unit collects sensing data from the sensing field(s) and makes a decision with collected data. This kind of structure is not adequate for ubiquitous computing where sensors and sensors' data should be easily accessed through a transparent and abstract way from ad-hoc ubiquitous computing services. Another problem is the extensibility of a central unit's code. A central unit tends to be hard-wired code programmed in the development phase. In this development cycle, when a new kind of devices or service is introduced to running sensor networks, corresponding code must be added and integrated with existing code in the source level. For these reasons, this centralized processing architecture is not adequate for the ubiquitous computing adopting sensor networks, none the less; most researches inevitably adopted the centralized processing architecture due to sensors' physical restrictions (i.e. limited computing resources and unstable wireless medium).

3 Structure of the Proposed System

3.1 Introduction of Virtual Counterpart

Concept of virtual counterpart is a structure in which physical objects in the real circumstance are mapped to virtual objects on the cyber space [12]. This structure is widely used in simulations or online games but not applied in sensor networks yet. In this paper, we designed a structure that provides ubiquitous computing service by locating virtual counterpart corresponding to each real sensor in the sink node (i.e. base station) as shown in Fig. 1. Virtual counterparts maintain collected data from corresponding real sensors and provide ubiquitous computing service by communication and cooperation among virtual counterparts.

The special position of the sink node in sensor networks makes this structure possible. Generally, the sink node plays a role as a gateway connecting the sensor networks and the outer networks. Therefore, all the collected data from the sensing field is transmitted to the sink node and, on the other hand, requests from the central unit are disseminated through the sink node. Since the sink node has computing power and energy resources superior to other ordinary sensor nodes, it is able to maintain a virtual counterpart correspondent to each sensor node. Cooperation among sensors and data fusion essential to ubiquitous computing are not easily achieved due to physical limits of sensor networks, but the virtual counterpart architecture can fulfill these requirements easily with existing sensor networks.

3.2 Detailed Architecture

To run virtual counterparts in the sink node, a sensor overlay network is designed to hold instances of virtual counterparts and provide communication channel among virtual counterparts. Designed overlay network consists of Virtual

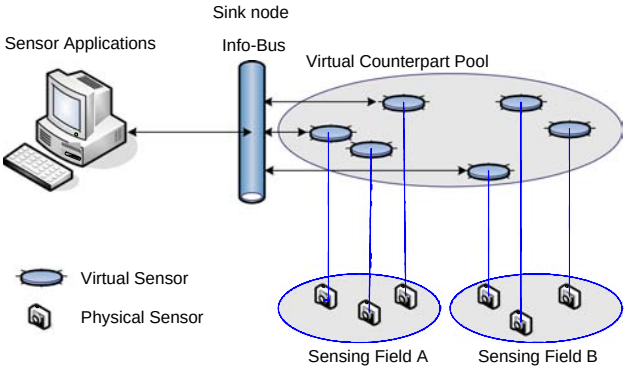


Fig. 1. Structure of the overlay network

Counterpart classes, Virtual Counterpart Pool and Info-bus. Virtual Counterpart classes represent virtual sensors and actuators. Virtual Counterpart Pool maintains instances of virtual counterparts and Info-Bus is a channel for these instances. The structure of the overlay network is shown in Fig.2.

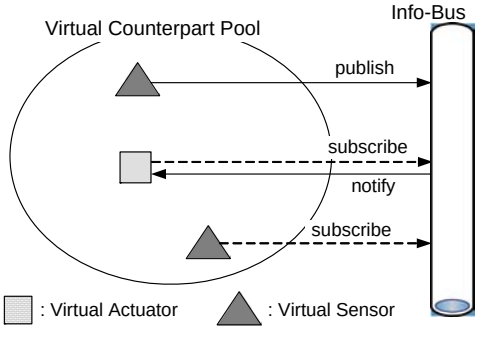


Fig. 2. Structure of the overlay network

3.2.1 VirtualCounterpart Class

Since a virtual counterpart corresponds to a sensor or an actuator, data of real sensors should be maintained as attributes in the virtual counterpart. In this paper, a VirtualCounterpart abstract class is designed to represent a virtual counterpart. Attributes of the VirtualCounterpart class can be any value, but essential data such as location, power, device type and sensing data must be included. The VirtualCounterpart class also has basic functions to spawn itself in the Virtual Counterpart Pool and communicate with other virtual counterparts. Attributes and functions of the VirtualCounterpart class are summarized in Table 1.

Table 1. Attributes and functions of the VirtualCounterpart class

Member variables : Attributes		
Name	Data type	Description
locatioin	Attr Class	Subclass of Attr Class, having (x,y) coordinates
battery	Attr Class	Subclass of Attr Class, indicates a range of 0∼100
deviceType	int	Defined by applications
sensingData	Attr Class	Subclass of Attr Class, processed by each sensor and actuator class
Methods : Functions		
Name	Description	
register()	Spawn virtual sensors or actuators into Virtual Counterpart Pool	
subscribe()	Subscribe interested sensor or data to Info-Bus	
publish()	Publish sensed data to Info-Bus	
notify()	Called from the Info-Bus to deliver interested messages	

3.2.2 Attributes Representation in Virtual Sensors and Actuators

Since the VirtualCounterpart class cannot represent whole sensors and actuators, each sensor or actuator should use its own virtual sensor or actuator class derived from the VirtualCounterpart class. These derived classes also have attributes and methods that represent their real counterparts, but the classes should follow the common interface of the VirtualCounterpart class to be uniformly accessed from the overlay network. In spite of enforcement of the common interface, virtual sensors and actuators classes can have their own functions by method overriding, but they cannot have different types of attributes from the VirtualCounterpart class. To resolve this problem, we design attributes using the Strategy pattern [13]. The pattern enables different implementations to be accessed through a unified interface. In this paper, an attribute class, Attr, is defined for this purpose and each virtual sensor or actuator represents its own attribute classes derived from the attribute class. Fig. 3 shows the class hierarchy of attribute classes.

The Attr class is an abstract class that provides isMatch() method. This method is used whether its member attribute matches to the given attribute. The decision for the matching depends on the characteristic of each virtual sensor or virtual actuator. By using this approach, a virtual counterpart can figure out if other virtual counterparts have target data in a unified way. For example, in the case that a virtual thermal sensor has TempData (i.e. a subclass of SensingData) as a sensingData member variable and a virtual humidity sensor has HumiData as its member variable, a caller can query temperature to whole virtual counterparts in the Virtual Counterpart Pool regardless of sensor types. When this query is arrived, the virtual humidity sensor will neglect this query because the query is not targeted for data of humidity type. On the other hand, the thermal virtual sensor will compare its attribute with the query. Code snippet of isMatch() in TempData is as follows. As you can see in the code, before

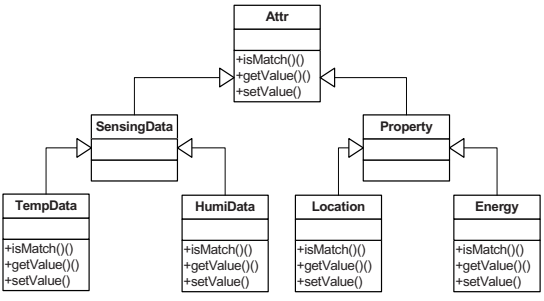


Fig. 3. Class hierarchy of attribute classes

executing isMatch() code, there is a part that checks whether the requested attribute is the same type of its member attribute. This approach enables the unified query handling for heterogeneous attributes.

```
bool isMatch(Attr queryAttr) {
    if(queryAttr instanceOf TempData) {
        int temp = queryAttr.getValue();
        int myTemp = sensingData.getValue();
        // member sensingData's type is TempData
        if(temp-2 <= myTemp  &&  myTemp <= temp+2) {
            return true;
        } else {
            return false;
        }
    } else {
        return false;
    }
}
```

3.2.3 Virtual Counterpart Pool and Info-bus

To actually operate virtual counterparts corresponding to sensors and actuators, each virtual counterpart have to be spawned and registered to Virtual Counterpart Pool using register() method. Virtual counterparts registered to Virtual Counterpart Pool have to communicate with each other for ubiquitous computing service. However, synchronous and direct communication is not suitable because virtual counterparts just mimic sensors and actuators in the sensing fields. Sensor networks usually do not adopt direct communications because of several problems related to host identification and transport layer support. For this reason, asynchronous and connectionless communication can be a better communication infrastructure for virtual counterparts. In this paper, we implemented a communication infrastructure, namely Info-Bus, adopting Message Oriented Middleware (MOM) [14] structure. Each Virtual counterpart

individually subscribes for interesting sensor data to the Info-Bus using subscribe() method. When some data is reported from a real sensor, its corresponding virtual sensor broadcasts the data to the Info-Bus using publish() method. In the Info-Bus, the data is directed to the virtual counterparts that subscribed their interest for the sensor. The example in Fig. 4 shows that a virtual air conditioner subscribes interest for thermal sensors. When data is reported from a virtual thermal sensor, it is propagated through the Info-Bus and directed to the virtual air conditioner using notify() method.

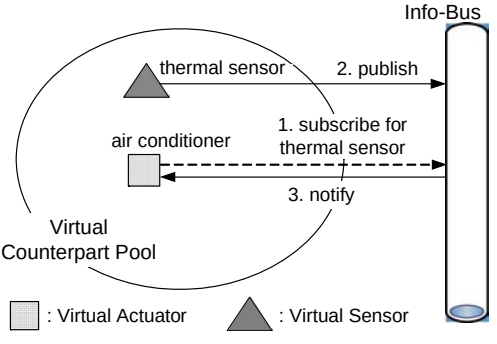


Fig. 4. Communication flows in the Info-Bus

Code snippet of notify() method in the virtual air conditioner is as follows. In the code, when data is delivered from the Info-Bus, the virtual air conditioner checks if the corresponding sensor is located at the same place. For the case of an air conditioner, the temperature in the same place will be important.

```
void notify(VirtualCounterpart sensor) {
   Attr loc = sensor.getLocation();
   if( this.location.isMatch(loc) ){
   // Sensor and Actuator exist in same area ?
      Attr seningData = sensor.getSensingData();
      int temp = sensingData.getValue();
      if(temp >= 28) {    // temperature is higher than 28C
         // turn on power
      } else if (temp < 20){ // temperature is lower than 20C
         // turn off power
      }
   }
}
```

Individually configured for its purpose, a virtual sensor or a virtual actuator can have intelligence in a distributed manner without centralized management. In the example above, thermal sensor's data also can be delivered to a heater

or a fire alarm without any modification of the existing virtual thermal sensors and air conditioners. This structure makes complex action and sensor cooperation possible in a distributed way. For another example, an asthmatic patient monitoring sensor subscribes for humidity data. When a change of humidity is detected, the patient monitoring sensor figures out the patient's location and cooperates with the humidity sensor to control a humidity controller actuator. By supporting asynchronous communication among virtual counterparts using the Info-Bus, an individual sensor or an actuator can configure its own customized action without affecting other elements. This enables intelligent cooperation and easy service modification that cannot be supported in the physical sensor networks.

4 Implementation and Evaluation

4.1 System Implementation

Proposed overlay network is implemented on J-Sim [9] simulator that is a component-based, compositional simulation environment. For sensor networks, the simulator provides BaseStation, SensorNode and TargetNode components and it supports sensor network simulation. To emulate physical sensors and actuators, given SensorNode and TargetNode classes are used in the evaluation. We designed VirtualCounterpart classes, Virtual Counterpart Pool and Info-Bus. Virtual Counterpart Pool and Info-Bus are implemented by modifying the sink node (i.e. BaseStation) class. Virtual thermal sensor class is newly designed to act as a proxy for a real thermal sensor and communicate with other virtual counterparts through the Info-Bus. For the evaluation, VirtualFireAlarm and VirtualAircon classes are used to control an external physical fire alarm and an air conditioner. Info-Bus is designed to support MOM structure for supporting asynchronous communication among virtual counterparts. Relation among designed classes is shown in Fig. 5.

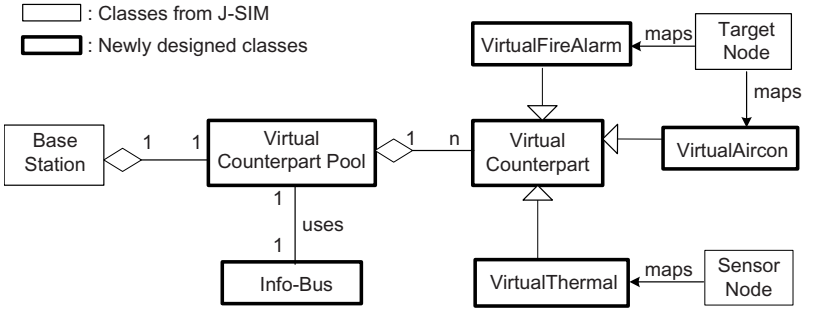


Fig. 5. Relation among classes

4.2 Operation Test

To test proper operations assumed in the design phase, a test environment is setup on the J-Sim as shown in Fig. 6. In the test environment, air conditioners and thermal sensors are located in separated room A and B. Virtual counterparts for every sensor and every actuator in each room exist in the sink node. Each virtual air conditioner actuator is configured to turn on the physical air conditioner if temperature of the room is higher than 28 °C. It will turn off the air conditioner if temperature is lower than 20 °C. In the first test, thermal sensors read a script containing temperature changes and report the data to the corresponding virtual sensors. After the first test, a fire alarm is additionally located in room B. When room temperature is higher than 80 °C, the virtual fire alarm will turn on the real fire alarm. Initial temperature is set as 25 °C.

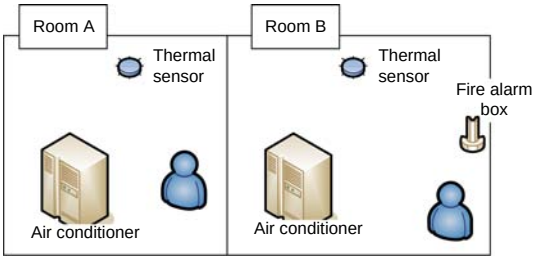


Fig. 6. Test environment

Table 2. Attributes and functions of the VirtualCounterpart class

			Room A	Room B
		Initial temperature	25 °C	25 °C
Test-1	Script input	Temperature change	5 °C up	5 °C down
	Expected Operation	Thermal sensor	Sensed 30 °C	Sensed 20 °C
		Actuator	Activate air conditioner	Deactivate air conditioner
Test-2	Script input	Temperature change	15 °C up	55 °C down
	Expected Operation	Thermal sensor	Sensed 40 °C	Sensed 80 °C
		Actuator		Activate Fire alarm

Changed amount of temperature included in each script is shown in table 2. When the test script is read, the message output is shown in Fig. 7.

This message output result shows orders of actions in the system and data flow. As shown in the output message during Test-1 phase (09:01:03∼09:05:32), temperature changes of sensors affect the virtual air-conditioner's action. The result shows that sensor cooperation is possible without direct communication among sensors and actuators in the physical sensing field. In the Test-2 phase (10:12:43∼10:15:34), a fire alarm actuator is added dynamically in the sensing field. When temperature is higher than 80 °C, the virtual fire alarm decides activation. The result shows that dynamic addition of a new kind of devices does not affect other existing devices or services.

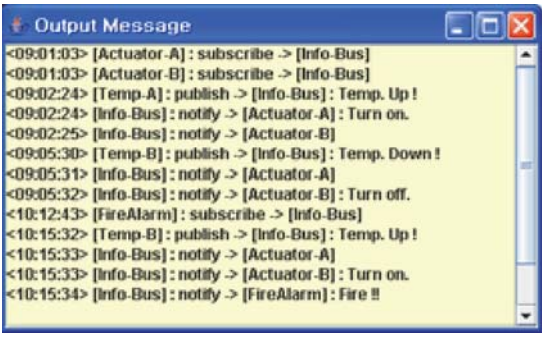


Fig. 7. Output message during simulation

5 Conclusions

Sensor networks are expected to be widely used as an infrastructure for collecting data in ubiquitous computing environment. However, physical limitation and centralized information scheme of sensor networks prohibit seamless integration of sensor networks to ubiquitous computing environment. To achieve transparent and abstract access of sensor and sensor's data essential to ubiquitous computing, direct communication and cooperation among sensors are required, but it is very difficult to fulfill these requirements on current sensor networks. In this paper, a virtual counterpart concept is adopted to resolve these issues. Virtual counterparts corresponding to physical sensors are located in the sink node and they cooperate with each other for ubiquitous computing services. This structure enables highly structured intelligent actions that are unable to be taken in the physical sensor networks. It also shows that dynamic service composition and device addition are possible without affecting other runtime components. The proposed overlay network was simulated using J-Sim simulator to verify the usefulness. The simulation result showed that the proposed structure performs requested functions properly. Further research is working on porting of proposed structure to the Tiny-OS.

References

1. Weiser, M. : "The Computer for the 21st Century". - Scientific American 265(30), Page(s). 94–104, 1991.
2. Debashis Saha, and et al : "Pervasive Computing: A Paradigm for the 21st Century". - IEEE Computer Society, March 2003.
3. IRFAN A. ESSA : "Ubiquitous Sensing for Smart and Aware Environments". - IEEE Personal Communications, October 2000.
4. B. Brumitt, and et al : "EasyLiving: Technologies for Intelligent Environments". - Proc. Handheld and Ubiquitous Computing 2nd Int'l Symp. HUC 2000. Springer-Verlag, NewYork, 2000. Page(s) 12–29.
5. M. H. Coen, and et al : "Meeting the Computational Needs of Intelligent Environments: The Metaglue System". - Proc. 1st Int'l Workshop Managing Interactions in Smart Environments, 1999.
6. Ian F. and et al : "A Survey on Sensor Networks". - IEEE Communications magazine, August 2002.
7. Athanassios Boulis, and et al : "A Framework for Efficient and Programmable Sensor Networks" - OPENARCH, July 2002.
8. Yong Yao, and et al : "The Cougar Approach to In-Network Query Processing in Sensor Networks". - In SIGMOD, 2002.
9. J-Sim is a component-based, compositional simulation environment. AKA(Autonomous Component Architecture) : (http://www.J-Sim.org/)
10. Heinzelman W.R, and et al : "Energy-efficient communication protocol for wireless micro-sensor networks". - System Sciences, 2000. Proceedings of the 33rd Annual Hawaii Int'l Conference on Jan 4–7 2000. Page(s) 10. vol.2.
11. TinyOS Community Forum —— An open-source OS for the networked sensor regime. : (http://www.tinyos.net/)
12. Kay Romer, and et al : "Smart Identification Framework for Ubiquitous Computing Applications". - Proceedings of the First IEEE International Conference on Pervasive Computing and Communications. (PerCom'03).
13. Erich, G, and et.al., "Design Patterns: Elements of Reusable Object-Oriented Software," Addison-Wesley, 1995.
14. G. Banavar, and et al : "A Case for Message Oriented Middleware". - In 13th International Symposium on Distributed Computing. (DISC 99), September 1999. Page(s) 1-18

An Efficient Sensor Network Architecture Using Open Platform in Vehicle Environment

Hong-bin Yim, Pyung-sun Park, Hee-seok Moon,
and Jae-il Jung

Department of Electrical and Computer Engineering Hanyang University,
17 Haengdang-dong, Sungdong-gu, Seoul, 133-791, Korea
{hbyim,xclass}@mnlab.hanyang.ac.kr, hsmoon@katech.re.kr,
jijung@hanyang.ac.kr

Abstract. Vehicles have been developed with an objective of safety. A large number of sensors will be required in Advanced Safety Vehicles that provide intelligent and automatic services in future ITS(Intelligent Transport Systems) circumstances. Because current in-vehicle networks must be changed to add new sensors, the number of sensors that can be added is restricted in current in-vehicle networks. To manage the sensors more efficiently and to provide extensibility, we propose a SCSN (Smart Car Sensor Network), which is an in-vehicle architecture based on AMI-C and OSGi standards. In this architecture, Vehicle Interface (VI), defined in the AMI-C standard, performs as a gateway in an AMI-C network. An integrated VI structure has been developed to provide a Vehicle Service (VS) on a standard platform. An interworking structure with a CAN(Controller Area Network) interface is implemented to provide an efficient VI. In current telematics architecture, time delay occurs between the CAN network start-up time and the platform booting time. Message loss occurs during this time delay. In this paper, we propose an efficient gateway architecture to minimize message loss due to this time delay. The efficiency of this platform has been verified using CANoe, which is a vehicle-network simulation tool.

Keywords: SCSN(Smart Car Sensor Network), Telematics, ITS, Sensor Network, Sensor Network Gateway, Sensor Clustering Node.

1 Introduction

At present, the stability of in-vehicle networks is suboptimal due to increases in wire length, difficulty in diagnosing sensor failures, and fault tolerance issues. Because control and sensing data are transmitted in a single network, it is difficult for in-vehicle portable devices to collect sensor data to check a vehicle status. Furthermore, current in-vehicle networks have problems adding new sensors. To solve this problem, we propose an additional in-vehicle sensor network, namely SCSN (Smart Car Sensor Network). The proposed in-vehicle sensor network collects sensor data from distributed smart sensors, using a sensor clustering node,

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 232–241, 2007.

and sends these data to sensor network gateway. The role of the sensor network gateway is to maintain and manage the overall network, as well as to process sensor data. The processing module of a sensor network gateway creates new information using sensor data fusion techniques.

New information created by sensor network gateway is provided to in-vehicle devices, multimedia terminals, and additional control boxes. For this, vehicle middleware is necessary to provide a vehicle status and travel information to in-vehicle devices.

The SCSN platform is based on OSGi[10] and AMI-C[11], which are open standards for telematics. These standards provide extensibility and interoperability for next-generation in-vehicle software and devices.

The rest of this paper is organized, as follows. Section 2 investigates the current in-vehicle network technologies and problems. Section 3 outlines the proposed SCSN platform architecture and components and illustrates the implementation challenges of the gateway in our SCSN. Section 4 explains the simulation environments and evaluates the performance of the SCSN network, compared with current in-vehicle networks.

2 Related Works

2.1 Architecture of Current In-Vehicle Networks and Its Problems

A vehicle consists of about 11,136 electrical devices, 60 electrical control units (ECU) and 3 Controller Area Network (CAN)[1] buses, with 2,500 signals and 250 CAN messages.

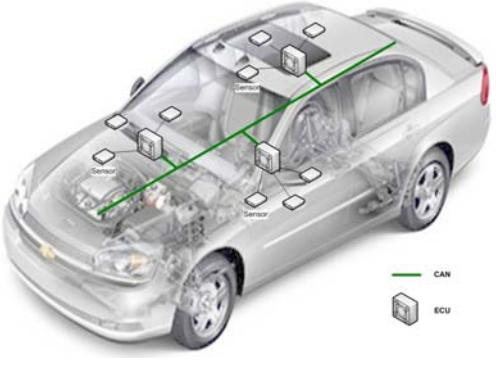


Fig. 1. Current in-vehicle network

Each ECU has multiple sensor nodes and communicates using multi-master communication technology. Sensing data to be collected by ECU are broadcast on the in-vehicle network. Sensing data created by distributed sensors are sent

to the actuator in order to perform a job. However, it is difficult to distinguish needed information from all sensing data. Additional overhead is required to acquire proper data from CAN messages collected by ECU. Only proper CAN data needs to be retained from the complete CAN message. Fig 1 shows an example of a current in-vehicle network. Current in-vehicle networks do not support a telematics terminal for travel conditions and location information [2][5].

2.2 Current In-Vehicle Sensor Allocation and Its Problems

A vehicle has many kinds of sensors, such as for safety and airbags. These sensors are connected to each other over a single in-vehicle network. Fig. 2 shows an example of current vehicle sensor allocation. These sensors consist of existing sensors, advanced airbag system sensors, hybrid vehicle sensors, and safety sensors.

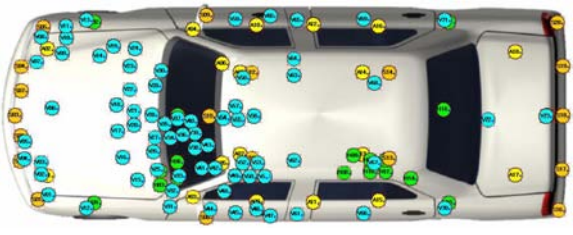


Fig. 2. Vehicle sensor allocation

Current in-vehicle networks are becoming very complex due to increasing numbers of sensors and wire lengths. This network needs to change the existing in-vehicle network to add new sensors. Because all information collected by ECU must be examined to distinguish needed sensing data, technical overhead exists to obtain only the sensing data needed by the ECU.

2.3 Current Telematics Architecture and Its Problems

The AMI-C standard defines a logical vehicle information architecture, namely Vehicle Service (VS) [6][7]. This architecture provides a vehicle information and control service to a platform application by cooperating with the in-vehicle network. The telematics platform based on OSGi operates on the Java Virtual Machine (JVM) [8]. This platform has a problem in that it cannot support VS with the AMI-C standard.

As we see in Fig. 3, this platform uses a vehicle network driver using Java Native Interface (JNI) [8] to provide vehicle information to in-vehicle devices.

JNI on JVM is a method to communicate with other communication technology except TCP/IP [9]. Because a single bundle handles all real-time vehicle messages, using only JNI in this architecture, this communication method is not efficient. A message created by the in-vehicle network may be lost due to platform start-up time delay when the vehicle is started up or when the bundle is restarted. This is further explained in section 3.2.

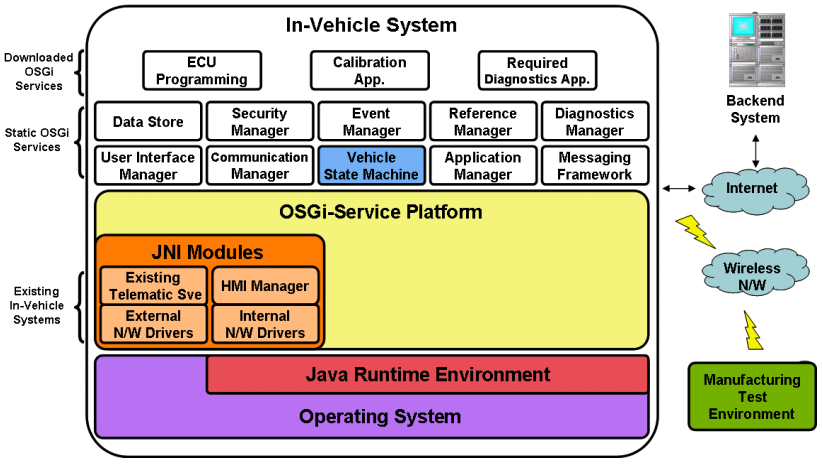


Fig. 3. Current telematics platform architecture[9]

Architecture based on AMI-C provides vehicle travel and status information to in-vehicle devices through middleware [11]. This architecture efficiently provides sensing data management and a processing function for multiple sensors. In this paper, we propose an additional in-vehicle sensor network, using the AMI-C standard, namely a smart car sensor network. The role of this network is to collect and manage sensor data. This network consists of several sensor clustering nodes to collect sensing data from sensors and one sensor network gateway to provide this data to a telematics terminal, multimedia devices, and an additional control box.

3 Smart Car Sensor Network (SCSN)

3.1 Smart Car Sensor Network Architecture

Fig. 4 shows the proposed SCSN architecture. This network architecture provides flexibility to add new sensors and efficient sensor data management using a sensor network gateway and sensor clustering nodes. As we see in Fig. 4, the SCSN consists of one sensor network gateway, several sensor clustering nodes, the current in-vehicle network, in-vehicle devices, and an additional control box.

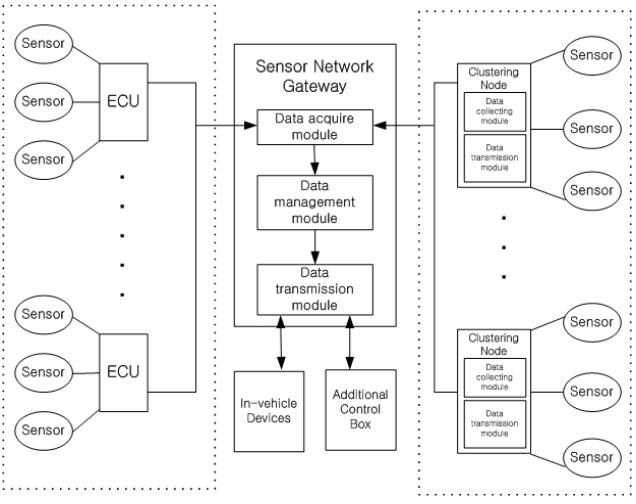


Fig. 4. Smart car sensor network architecture

The sensor network gateway networks among the current in-vehicle network, the sensor clustering node, in-vehicle devices, and an additional control box. The sensor network gateway collects sensing data from the existing in-vehicle network and sensor clustering nodes. This gateway sends this data to a telematics terminal, multimedia devices, and an additional control box. Also, this gateway performs functions, such as data maintenance, processing, and management. The sensor clustering node collects sensing data from distributed sensors. This node transmits sensing data to the sensor network gateway and manages many kinds of sensors connected to this node.

In this architecture, the existing in-vehicle network does not need to be changed to add new sensors so new sensors may be installed on the right side of Fig. 4. A new sensor is connected to a sensor clustering node and is managed by it. This characteristic of SCSN provides flexibility and extensibility for new sensor installation.

3.2 Implementation of the Sensor Network Gateway

Fig. 5 shows the internal architecture of an in-vehicle sensor network gateway. As we mentioned in Section 2.3, message loss occurs during the platform start-up time delay or platform initiation time in current in-vehicle network communication technology. Current in-vehicle communication operates on a JNI driver. This communication technology creates a time delay between the CAN network start-up time and the platform boot time. In current in-vehicle architecture, message loss occurs during this time delay. In our SCSN architecture, the sensor network gateway has components to reduce message loss by decreasing this time delay. These components are as follows.

– CAN gateway
– Vehicle Service Interface (VSI) gateway bundle to provide AMI-C standard vehicle service based on the OSGi platform
– Vehicle Service

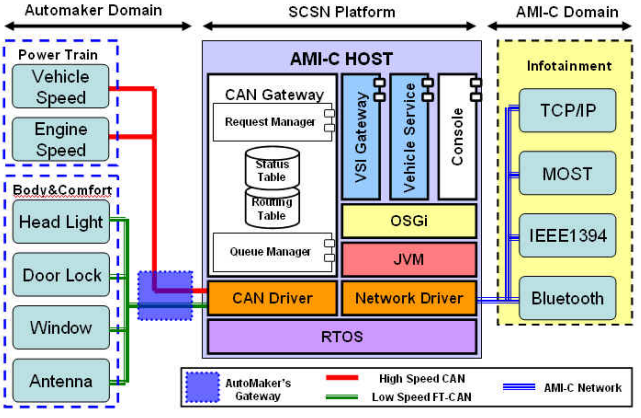


Fig. 5. Internal architecture of sensor network gateway

The CAN gateway cooperates with the VSI gateway using a TCP/IP protocol that is supported by JVM. This communication technology uses mutual communication to request service from the VSI gateway bundle to the CAN gateway. For this communication, we propose a new protocol, namely the GCP (Gateway Communication Protocol), which has the following functions:

1. Receiving/Transmitting the CAN message request
2. ID-based CAN message Registration/Release
3. Inquiry/Modification of routing and status table in CAN gateway
4. Confirm/Reset of processed messages

Mainly, function 1 performs the CAN message receiving/transmitting request and sends the stored message in a status table to the proper application bundle. Function 2 performs ID start/stop for a specific message by message subscription and message transmission request by setting a period. Function 3 and function 4 manage the routing and status table. In other words, the CAN gateway performs processing of vehicle status and sensing data by a message subscription method to control message transmission. This gateway sends the needed message to the VSI gateway so that it performs the role to control CAN and the AMI-C message conversion rate.

3.3 Sequence of Gateway Internal Cooperation

After start-up of the CAN gateway, the VSI on the platform receives stored vehicle data by connecting with the CAN gateway. After this procedure, the VSI

on the platform cooperates with the in-vehicle network. This VSI converts each message transmission requirement to a suitable GCP and controls transmission rates by checking the arrival of messages.

Fig. 6 shows sequence diagram of VSI cooperation with the CAN gateway.

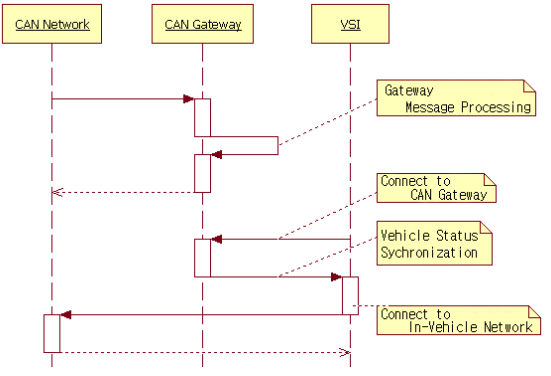


Fig. 6. Sequence diagram of VSI cooperating with the CAN gateway

4 Simulation

4.1 Simulation Environment

Fig. 7 shows a simulation environment using the CANoe vehicle network simulation tool. The simulation environment consists of a power train network with 500 kbps and a body network with 125 kbps. The power train network consists of the engine, ABS, gear box, and body network, which consists of a door control module, dashboard, and radio channel control console.

These components connect with the sensor network gateway and synchronize the vehicle velocity with the vehicle status information. Window and headlight information for the vehicle are controlled by the VS of the simulation environment.

Fig. 7. Simulation environments

4.2 Simulation Results

Vehicle start-up time delay is defined by the OS, CAN driver, VM, and OSGi message processing time, which depends on the bundle initiation time of a standard telematics platform. After this time delay, the sensor network gateway can process CAN messages.

This simulation measures time delay, message loss, and processing efficiency according to two different message processing methods.

1. Message processing method using a CAN module based on JNI
2. Message processing method using a CAN gateway in cooperation with VSI.

The measured items are as follows:

1. Time and message loss until CAN driver initiation
2. Message loss depending on the time from module initiation to message reception completion
3. The number of messages created by the platform application, which is controlled by message processing methods

Fig. 8 shows the overall results for time delay, message loss, and the number of processed messages.

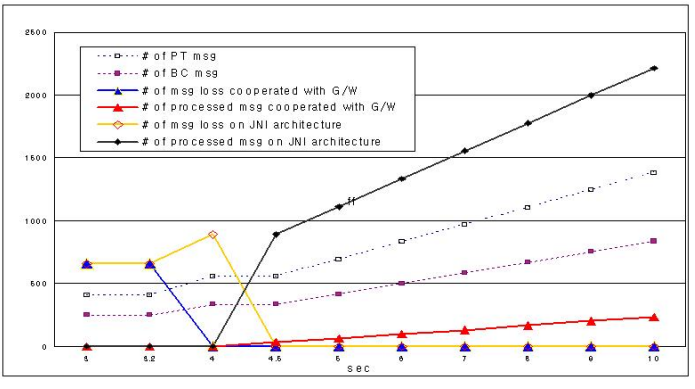


Fig. 8. Overall results for time delay, message loss, and the number of processed messages

The number of message loss cooperated with gateway is less than the number of message loss on JNI architecture, because the start-up time delay of proposed architecture is less than current telematics architecture.

Tables 1 and 2 show the averages of measurement results. PT stands for powertrain and BC stands for body control.

These 2 tables show that message loss is decreased by 27% in the case of powertrain and 25% in the case of body control. These results related to the average start-up time delay. In the case of sensor network gateway, average start-up time delay is decreased by 1.3 seconds in comparison with current telematics architecture.

Table 1. Message loss according to driver start-up delay time

Average of start-up delay(sec)	Average of vehicle message loss			
	Number of PT message	Frame/sec	Number of BC message	Frame/sec
2.711	273	132	168	83

Table 2. Start-up delay time and message loss according to architecture

Architecture	Average of start-up delay(sec)	Average of vehicle message loss	
		Number of PT message	Number of BC message
Currnet telematics	4.5	558	335
Sensor network gateway	3.2	404.6	251

5 Conclusion

In this paper, we have proposed a SCSN platform to solve the problem presented by an increasing numbers of sensors. The SCSN consists of a sensor network gateway and a sensor clustering node, which transmits sensing data to multimedia devices, an additional control box, and a telematics terminal. This network architecture improves sensing data sharing and management efficiency.

According to the simulation results, the SCSN platform is more efficient than current in-vehicle network architecture in terms of message loss and message processing rates. The SCSN platform provides extensibility, using an in-vehicle network standard, and supports easy installation of additional devices or sensors.

The merits of the SCSN platform are as follows:

1. Provides extensibility for an in-vehicle sensor network
2. Shares sensing data using sensor data management
3. Increases interoperability using a vehicle network standard
4. Improves in-vehicle network management by separating control and sensor data
5. Guarantees a predictive start-up time and efficient message processing

In the future, we will refine our proposed network architecture through improvement of the sensor network gateway function. A refined scheme should include an efficient data fusion algorithm and transmission algorithm from the sensor clustering node to the sensor network gateway. We will also study fault-tolerant architecture in an in-vehicle network.

Acknowledgements

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center)

support program supervised by the IITA(Institute of Information Technology Assessment) (IITA-2005-(C1090-0502-0020)).

References

1. Peng Shuai, Taeyeon Lee , Ealgoo Kim , Jaehong Park, "A Study on an algorithm of a Network Node for CAN-Based Wiring System Using Polling Structure", IPC-13 1105 - 1108, 2005
2. Choi Chang-hee, "Trends of Telematics Based or Network for Vehicle", Auto Journal, Vol 27, No 6, pp 9-16, 2005
3. http://www.can.bosch.com
4. AMI-C 1003 "AMI-C Release 2 Architectural Overview v1.00"
5. Syed Masud Mahmud, Sheran Alles, "In-Vehicle Network Architecture for the Next-Generation Vehicles", SP-1918, 99-108, 2005 SAE
6. AMI-C 1003, "AMI-C Release 2 Architectural Overview v1.00", 2003
7. Peter Abowd, Gary Rushton, "Extensible and Upgradable Vehicle Electrical, Electronic, and Software Architectures", SAE TRANSACTIONS, VOL 111, pp 495-498, 2002 SAE
8. http://java.sun.com
9. Vivek Kapadia, Veena Bai and G Parthasarathy, "Telematics System Architecture", WHITE PAPER, wipro Ltd., 2005
10. http://www.osgi.org
11. http://www.ami-c.org
12. AMI-C 4002, "AMI-C Requirements and specifications for Human Machine Interfaces v1.00", 2003
13. AMI-C 3001, "AMI-C Requirements and guidelines for software host platforms v1.00", 2003
14. AMI-C 2002, "AMI-C Common Message Set v1.01", 2003
15. AMI-C 2001, "AMI-C Network protocol requirements for vehicle interface access v1.00", 2003

Improved Reinforcement Computing to Implement AntNet-Based Routing Using General NPs for Ubiquitous Environments

Hyuntae Park[1], Byung In Moon[2], and Sungho Kang[1]

[1] Department of Electrical and Electronic Engineering, Yonsei University, 134 Shinchon-Dong, Seodaemoon-Gu, Seoul, 120-749, Korea
radiob@yonsei.ac.kr, shkang@yonsei.ac.kr
[2] School of Electrical Engineering & Computer Science, Kyungpook National University, 1370 Sankyuk-dong, Buk-gu, Daegu, 702-701, Korea
bihmoon@knu.ac.kr

Abstract. In the ubiquitous convergence era, the traffic managements and quality of services will be made much of a role. Because traditional routing mechanisms are lacking scalability and adaptability, a kind of adaptive routing algorithm called AntNet has attracted the attention. AntNet is an adaptive agent-based routing algorithm that imitates the activities of the social insect. In AntNet, there are implementation constraints due to complex arithmetic calculations for determining a reinforcement value. Besides, a housekeeping core in network processors will be overwhelmed by increasing routing workload for a processing of agents. In this paper, we propose a new reinforcement computing algorithm to overcome these problems. This can be implemented efficiently on packet forwarding engines of conventional network processors. The simulation results show that the proposed AntNet is more adaptive and effective in the performance of the implementation than the original AntNet.

1 Introduction

According to the advent of the ubiquitous convergence era, a unitary autonomous system will have more and more nodes and paths. The controls of an enormous network whether it is wired or wireless in the near future will become more difficult obviously. Besides, in order to support new various ubiquitous applications for a real-time multimedia, the traffic managements and quality of services are made much of a role. Accordingly, new paradigm for routing will be required in next generation networks. A kind of adaptive routing algorithm called *AntNet* has attracted the attention of researchers, among many different kinds of routing algorithms [1]. It is inspired by the adaptive and distributive behaviors of real ants to locate the shortest route from the nest (source node) to the food source (destination node) by deposing a chemical substance called *pheromone* on the trail[1].

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 242–251, 2007.

In previous works, the efficiency of AntNet has been verified by comparison with state-of-art routing algorithms such as RIP and OSPF in dynamic traffic environments [2]. However, up until now, there have been few attempts to implement it in a real network, and it is not used yet for real network routing systems. If AntNet is implemented by software on a housekeeping core as a traditional approach, an overload of a housekeeping core for processing incoming agents with a line-speed become further burdensome. Therefore, our prospective view is that, in the future, the processing of an agent-based routing protocol is performed on packet forwarding engines instead of the housekeeping core in order to improve the processing of agents. In other words, we expect that the routing in ubiquitous environments is a part of fast-path processing, not slow-path. A reinforcement computing is the most complex among the tasks of AntNet-based routing because it requires complex calculations. In this paper, the improving reinforcement computing on general packet forwarding engines is proposed to implement AntNet-based routing in real network environments.

Firstly, reinforcement in the original AntNet is briefly introduced in section 2. In the next section, we analyze constraints to implementing the original AntNet. The optimized AntNet algorithm developed to solve these indicated problems is presented in section 4. The performances of each algorithm are compared and evaluated in section 5. Finally, the conclusions of this paper are given in section 6.

2 The Reinforcement in the Original AntNet

AntNet algorithm, an approach to the adaptive learning of routing tables in communication networks, was first introduced in 1998 [1]. The concept of AntNet was derived from the adaptive behaviors of social insects such as ants, which behaviors are called *stigmergy*. Then, the proposed AntNet is introduced in [3]. However, because an improvement of reinforcement computing is insufficient in [3], we refer to the original reinforcement computing method in [1]. And, a detailed description of the operation of AntNet is not given here. According to the original AntNet, the equation of the reinforcement computing is as follows:

$$r = c_1 \left(\frac{W_{best}}{T} \right) + c_2 \left[\frac{I_{sup} - I_{inf}}{(I_{sup} - I_{inf}) + (T - I_{inf})} \right] \tag{1}$$

In the above Eq.(1), W_{best} is the shortest trip time experience by the ants traveling to the destination node d, which is observed within the scope of the window size. This window size reflects the number of considered samples before resetting the W_{best}, and it is assigned as a base of η, which is weighted by the number of samples effectively giving a contribution to the value of μ estimate.

$$\begin{gathered} P_{fd'} \leftarrow P_{fd'} + r(1 - P_{fd'}) \\ P_{nd'} \leftarrow P_{nd'} - r \cdot P_{nd'} \\ n, f \in N, n \neq f \end{gathered} \tag{2}$$

Eq.(1) is used to update a probabilistic entry in the routing table at the node k coming from the node f. $P_{fd'}$ represents probabilistic value of taking the node f as the next neighbor node towards the destination; whereas r represents a positive reinforcement. Eq.(2) is used to update probabilistic entries in the routing table at the node k for taking a neighbor node other than the node f as the next node towards the destination. According to Eq.(1) and (2), $P_{fd'}$ is increased but $P_{nd'}$ of other neighbor nodes is decreased at the point that the sum of $P_{fd'}$ and $P_{nd'}$'s is always 1.

3 Constraints of Implementation Using NP

We can consider two ways for the implementation of the AntNet-based routing. One is to build AntNet by adding the extra hardware for AntNet routing. The other is to build the AntNet using existing network devices and systems.

In the former case, existing network systems should be changed to become suitable for conventional state-of-the-art routing algorithms. It is a heavy burden with much cost and time because a current communication network is already occupied massively. Moreover, we never know whether it can be used widely and can be available on any network environment. This is a problem of compatibility. According to [1], AntNet is not available on all network environments in spite of the goodness of AntNet in a dynamical network environment. Therefore, we have to implement the AntNet without the extra hardware.

The latter case also has a problem. In order to implement AntNet-based routing with compatibility and efficiency, we premise that AntNet is implemented on network processors. Furthermore, in order to reduce a burden of a housekeeping core in network processors, agent processing for AntNet is performed on packet forwarding engines instead of the housekeeping core.

Reinforcement computing as briefly described in the previous section is not very complex mathematically. However, general packet forwarding engines have a concise arithmetic unit for simple processing called parsing and comparison. Accordingly, if the reinforcement computing based on the original AntNet by itself is performed on packet forwarding engines instead of the housekeeping core, the performance of computing is worse. Nowadays, network devices for routing are mainly made of network processors which are the products of Intel, Agere Systems, Lextra, Sitera, Clearwater Networks and so on. Considering the architecture of each packet forwarding engine in these processors, packet forwarding engines that consist only of ALU and shifter are sufficient to perform packet forwarding functions [4][5][6][7]. Moreover, these only support an integer type processing even though AntNet needs floating point type processing for handling probabilistic values. Therefore, we proposed the improved the reinforcement equation to implement AntNet on existing packet forwarding engines as it is.

4 The Proposed Algorithm

The excellent merit of the proposed algorithm is that the reinforcement is computed with simple arithmetic units. The proposed algorithm is induced as follows:

$$r = normal\left[\frac{bCost}{curCost}\right] \tag{3}$$

While several factors are induced to get a reinforcement value in the original AntNet, we only use the round trip time of the ants as Eq.(3). In Eq.(3), *bCost* is the same as W_{best} in Eq.(1) and *curCost* is the same as T in Eq.(1). In Eq.(1), the first and most important term weighs the goodness of the current trip time compared to the best trip time. The second term is used to complement the first term [2]. Even though the second term is less influential than the first term in the performance, it takes more time to calculate than the first term. Therefore, the second term in Eq.(1) is neglected unavoidably. Although it is an approximation of the original AntNet, we can simplify the process to improve the efficiency of the implementation without a serious loss of an adaptability.

In Eq.(3), a division mechanism is required for an arbitrary value. According to programming reference manuals of conventional network processors, a multiplication and a division are operated ineffectively by recursive additions or subtractions. Therefore, we propose Eq.(4) and Eq.(5) as an advanced revision of Eq.(3). These formulas are proposed heuristically based on the following rules:

$$r' = 255 + (bCost - curCost) \tag{4}$$

$$r = normal\left[\frac{r' \cdot bCost}{C_{bCost} \cdot C_{curCost}}\right] \tag{5}$$

Rule 1. *A relative term transforms into an absolute term.*
In Eq.(3), the division presents the relation between a denominator and a numerator. Because a denominator is an arbitrary value, the division procedure consumes a lot of time for the conventional network processors which perform the division by recursive subtractions. Thus, we try to transform a relative term as the division into an absolute term as the subtraction. In order to find the effect of *bCost*, if *curCost* value is fixed, *bCost* is positively in proportion to r'. Therefore, *bCost* must be the subtracted number in the subtractive term as the second term of Eq.(4) is presented. Because *curCost* is always equal to or larger than *bCost*, it is necessary to add any particular constant so that r' has a positive value. This particular constant is 255 as determined by a heuristic method. Finally, we can draw Eq.(4).

Rule 2. *Weighed values complement the transformed absolute term, r'.*
The transformed absolute term, r' in Eq.(4) is inferior in adaptability to the original reinforcement r. Considering the mutual relation between r' and the reinforcement r, if *bCost* and *curCost* are large values, the reinforcement r

should have a large value regardless of r'. On the other hand, if *bCost* and *curCost* are small values, the reinforcement r should have a small value. That is, even if r' is the same as in the above two cases, we should weight r' according to *bCost* and *curCost*. In order to complement this, we propose that weighted values, C_{bCost} and $C_{curCost}$, are assigned by *bCost* and *curCost* for division. In addition, in order to simplify the processing and save the calculation time, we propose that weighed values are multiples of 2 specially. The relation to determine the weighted values are shown in Table 1.

Table 1. Assign C_{bCost} and $C_{curCost}$ according to *bCost* and *curCost*

bCost and *curCost*	C_{bCost} and $C_{curCost}$
4 ∼ 7	2
8 ∼ 15	4
16 ∼ 31	6
32 ∼ 63	8
64 ∼ 127	10
128 ∼ 255	12
256 ∼ 511	14
512 ∼ 1024	16

Rule 3. *The reinforcement r is directly proportional to bCost sensitively.*
In rule 1, we confirm that the mutual relation between r' and *bCost* is proportional, when we assume that *curCost* value is fixed. We examine the mutual relation between the reinforcement r and *bCost*. Through various experiments, we learned that this relation is directly proportional. In addition, we verify that this relation is very sensitive. That is, the reinforcement r is nearly directly proportional to *bCost*. For this reason, r is multiplied by *bCost* itself as the role of the weighted value. There is no need to worry more complex by this. Because the multiplication is simpler than the division about calculation onto network processors, we can expect processing time to increase a little by adding this procedure. Therefore, it is not a great burden.

The proposed procedure for calculating the reinforcement value is presented as the following pseudo code at Fig. 1. Firstly, it stores the weighted values that are the multiples of 2 from 2 to 16. Then, it checks the validity of *bCost*. Next, if *curCost* is less than *bCost*, *curCost* becomes the new *bCost*. At this time, the reinforcement r is assigned the defined maximum value. Else, C_{bCost} and $C_{curCost}$ are assigned by the relation of Table 1. Finally, the reinforcement r is calculated by Eq.(4) and Eq.(5). If this calculated reinforcement exceeds the limit of maximum or minimum, it is assigned the available maximum or minimum values.

```
reF : Reinforcement value r
bCost : Best cost value within the available time
curCost : Current cost value reported by backward agent
C_bcost : Weighted value according to bCost
C_curCost : Weighted value according to curCost
Comptable : Assigned weighted value by interval

int comptable [] = {2, 4, 6, 8, 10, 12, 14, 16};
void setReinforcement ( int curCost ) {
      if ( ! checking the available time of bCost ) {
        bCost reset;
      }
      if ( curCost < bCost ) {
        bCost = curCost ;
        reF = max [ r ] ;
      }
      else {                    // assign by 2n interval
        for ( i=1; i=amount of interval ; i++) {
            if ( 2i+2 < bCost < 2i+3 ) C_bcost = comptable [i];
            if ( 2i+2 < curCost < 2i+3 ) C_curCost = comptable [i];
        }
        reF = normal {(255 + (bCost - curCost))* bCost / C_bcost / C_curCost } ;
      }
      if ( reF > 25 ) reF = 25;          // limit max . & min. reF
        else if ( reF < 1 ) reF = 1;
```

Fig. 1. Pseudo Code for the Calculation of the Proposed Reinforcement r

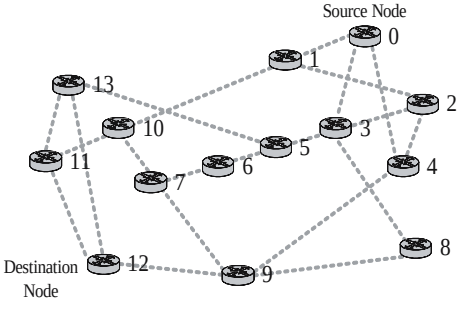


Fig. 2. Network Topology for the Simulation

5 Simulation Results

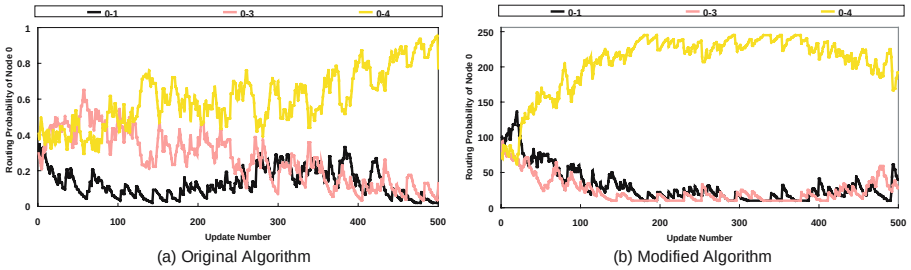
The proposed algorithm is evaluated by the comparison with the original AntNet in the topology presented in Fig. 2. This topology is like NSFnet. The node "0" is the source node and the node "12" is the destination node. The probability for routing of the source node "0" is evaluated and the variation of the best path is observed in the dynamic topology. The update number, which is a value of x axis, equals the frequency of changing probabilities of routing tables on the source node. If a particular path has the most routing probability from the source node to neighborhood nodes "1", "3", and "4", we can confirm that this way is the best path. The routing probability on the source is presented in y axis. We also define the four states of the network traffic patterns as follows:

- `Light:` All traffics are equally distributed. It means the routing time between nodes is nearly the same.
- `Biasing:` The routing time of a specific path is shorter than the others. It means the cost of links by a trip time are different from those by hopping.
- `Heavy:` All traffic are uniformly distributed as the light pattern. However, the amount of traffic is 10 times heavier than that of light state.
- `Dynamic:` The network traffic pattern changes periodically at each 500 update. The traffic pattern is rotated sequentially the light, the biasing and the heavy traffic state.

We assume that node "0" is the source node and node "12" is the destination node. There are many paths from node "0" to node "12". The time delay on each link sets as cost according to each of the above traffic pattern from the light, to the biasing, and to the heavy traffic state.

In the light state, the path via 0-4-9-12 has the smallest number of links to reach the destination node. Therefore, since all traffic is equally distributed between nodes, this path is the best path and the probability of choosing the 0-4 link is larger than others. It is shown as Fig. 3.

In the biasing state, we assume that the path via 0-3-5-6-7-9-12 would have the lowest cost even though this has a larger number of links to the destination



(a) Original Algorithm (b) Modified Algorithm

Fig. 3. Simulation Results: Light State

node. As observed in Fig. 4, the probability of choosing this best path is the highest. This indicates that the modified algorithm correctly finds the best path in the same way as the original algorithm.

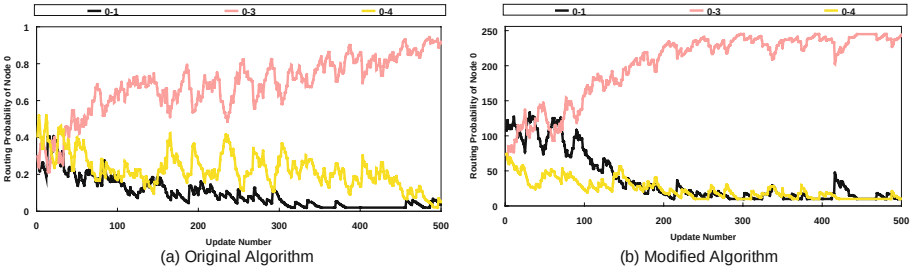


Fig. 4. Simulation Results: Biasing State

In the heavy state, because distributed traffic is the same as the light state despite the 10 times larger routing time between nodes needs, we expect the same results as for the light state. We observe this clearly in Fig. 5. Based on this result, we confirm both algorithms can select the best path by relatively distributed costs.

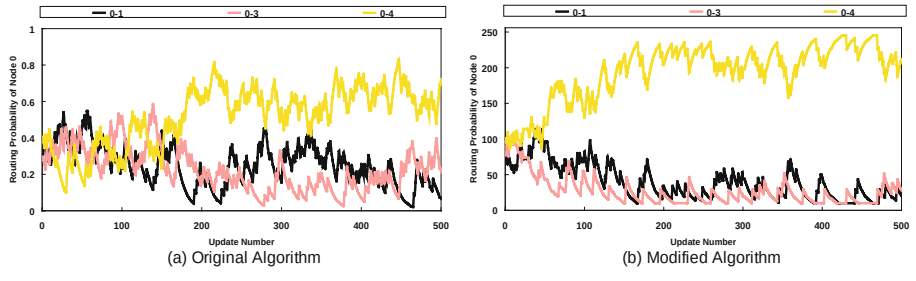


Fig. 5. Simulation Results: Heavy State

In the dynamic state, when updating the routing probability on the source node become each 500 time, the network traffic state changes to the next state. The order is light, biasing and heavy states. The results are shown in Fig. 6. As the original AntNet are affected by the second term of Eq.(1), it needs to have a lot of time to adapt to new environments. So, the original AntNet delays slightly while adapting to dynamic network topology. However, the proposed algorithm quickly reacts to the traffic variation due to its simpler processing. Therefore, the shapes in Fig. 6(b) are clearer than those in Fig. 6(a). Although the two graphs differ in detail, both algorithms selected the best path.

The simulation results confirm that the performance of the proposed algorithm which is implemented efficiently on existing network devices is similar to

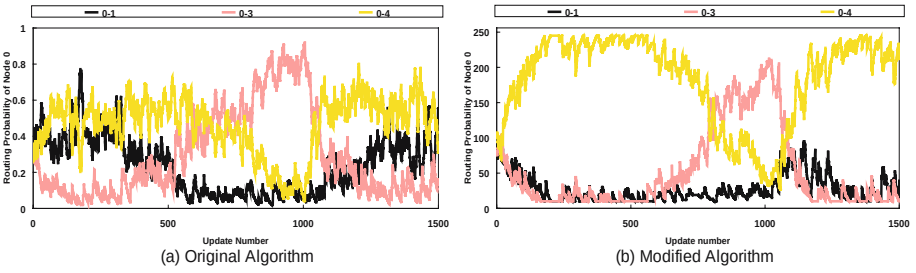


Fig. 6. Simulation Results: Dynamic State

that of the original algorithm. Moreover, we can implement AntNet-based routing without the extra hardware or replacement of existing network devices. In addition, to find the best path, the proposed algorithm is clearly more efficient than the original algorithm in a dynamical traffic environment. Selecting the best path of the proposed algorithm is also faster than that of the original algorithm. It is certain that the best path from the source to the destination is found by the proposed algorithm. As it shortens the processing time on hardware for routing, it improves the performance speed as well.

6 Conclusions

If AntNet is implemented on a housekeeping core in network processors around Ubiquitous Environments, a housekeeping core become overwhelmed by increasing routing workload for a processing of agents. There are constraints to implement the original AntNet on conventional packet forwarding engines with simple arithmetic units as it is. In this paper, the reinforcement computing method with simple arithmetic units is proposed. It can be implemented efficiently onto packet forwarding engines of conventional network processors. The results of the simulation show that the proposed AntNet is more adaptive and effective in the performance of the implementation than the original AntNet.

Acknowledgment

This research was supported by University IT Research Center Project of Korea.

References

1. Dorigo, M., Di Caro, G.: AntNet : Distributed Stigmergetic Control for Communication Networks. Journal of Artificial Intelligence Research, Number 9. (1999) 317–365
2. Sim, K. M., Sun, W. H.: Ant Colony Optimization for Routing and Load-Balancing: Survey and New Directions. IEEE Transactions on Systems, Man and Cybernetics, Part A, Vol. 33, Issue 5. (2003) 560–572

3. Baran, B., Sosa, R.: A New Approach for AntNet routing. Proc. of Ninth International Conference on Computer Communications and Networks. (2000) 303–308
4. Intel Corp.: Intel IXP1200 Network Processor Family Hardware Reference Manual. Intel Press. (2001)
5. Halfhill, T. R.: Lexra's NetVortex Does Networking. Microprocessor Report, Vol.13, No.12. (1999) 15–19
6. Halfhill, T. R.: Sitera Samples Its First NPU. Microprocessor Report, Vol.13, No.12. (1999) 7–10
7. Clearwater Networks.: Introducing the CNP810 Family of Network Services Processors. Clearwater Networks. (2001)
8. Kurose, J. F., Ross, K. W.: Computer Networking. Addison-Wesley. (2002)

Design of a Cooperative Distributed Intrusion Detection System for AODV

Trang Cao Minh and Hyung-Yun Kong

Department of Electrical Engineering
University of Ulsan 680-749, Ulsan, Korea
trangcm@gmail.com
http://wcomm.ulsan.ac.kr

Abstract. The Ad hoc On Demand Distance Vector (AODV) protocol is a routing protocol designed for mobile ad hoc networks. A mobile ad hoc network can be defined as a network of mobile nodes that communicate over the wireless radio communication channel. To transmit data over such a network, the AODV protocol enables dynamic, self-starting, multihop routing between mobile nodes. However AODV is vulnerable to various kinds of attacks. In this paper we have developed a Cooperative Distributed Intrusion Detection System to protect against some of attacks in AODV such as denial of service and sequence number modification in RREQ (Route Request) packets. We have modified original AODV in network simulator (ns2) to evaluate our solution. The simulation results show that the performance of AODV routing protocol when using our approach is improved significantly under attacks and even better than the original AODV in some cases.

1 Introduction

During the last few years, wireless and mobile communication networks have quickly developed and are widely used. A mobile ad hoc network (MANET) is a multi-hop wireless network formed by a collection of mobile nodes without the intervention of fixed infrastructure. Typical application areas of mobile ad hoc network include battle fields, emergency search, rescue sites and data acquisition in remote areas. A mobile ad hoc network is also useful in classrooms and conventions where participants share information dynamically through their mobile computing devices. The lack of fixed infrastructure and dedicated nodes that provide network management operations like the traditional routers in the fixed network poses many new challenges. The first challenge is how to maintain connectivity in network. The routing problem in MANET has been solved by some protocols such as AODV (Ad hoc On-Demand Distance Vector), DSR (Dynamic Source Routing). The second challenge is security. Due to the characteristic such as dynamic changing topology, heterogeneous and decentralized control, limited resources and unfriendly environment, ad hoc networks are vulnerable to various attacks including spoofing, modification of packets and distributed denial of service (DDoS), etc. To secure ad hoc networks, the prevention based approach is

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 252–262, 2007.

not enough (it cannot protect against the attacks from internal nodes), it is necessary to develop intrusion detection mechanisms. Intrusion detection involves the runtime gathering of data from system operation, and the subsequent analysis of the data; the data can be audit logs generated by an operating system or packets "sniffed" from a network.

In this paper, we proposed a simple method to detect the malicious nodes in MANET. We complemented an IDS agent on every node in network. This agent stores the records of incoming route request packets and detects any intrusions based on the predefined constraints.

In the remainder of this paper we start by summarizing related works in section 2 and briefly presenting the overview of AODV protocol and some attacks for it in section 3. In section 4 we describe the Intrusion Detection System (IDS) and a distributed architecture of IDS. Section 5 presents in detail our proposed solution for AODV-based networks. In section 6 we evaluate our solution that has been implemented by using ns-2. Finally, we draw the conclusion and give some directions for future works in Section 7.

2 Related Work

Most protocols for ad hoc networks have been developed without any consideration for security, assuming that every node in the environment is cooperative and trustworthy, i.e. AODV [1], DSR [9]. The few protocols for ad hoc networks developed with consideration for security can rather be seen as patches to the existing protocols, e.g. ARIADNE (for DSR) [10] and SAODV (for AODV) [11].

ARIADNE is a secure on-demand routing protocol proposed for ad hoc networks and based on DSR protocol to discover route path through ad hoc networks and TESLA [12] to peer-to-peer or broadcast authentication. It requires clock synchronization, which can be considered to be hard to achieve in public ad hoc networks.

In SAODV [11] each node has certified public keys of all network nodes. SAODV requires that routing messages must be signed by a private key; this prevents nodes from sending most false messages. In addition, when a node replies to a route request it also adds a signature from the destination to prove that it has a route. Finally, hash chains are used in order to protect the part of a message that intermediate nodes will change on each hop. However, even if SAODV can protect against several attacks it still has problems. The use of public-key cryptography imposes a high processing overhead on the intermediate nodes and can be considered unrealistic for a wide range of network instances. Furthermore, it is possible for intermediate nodes to corrupt the route discovery by pretending that the destination is their immediate neighbor, advertising arbitrarily high sequence numbers and altering the actual route length. Additional vulnerabilities stem from the fact that the IP portion of the SAODV traffic can be trivially compromised, since it is not protected, unless additional hop-by-hop cryptography and accumulation of signatures is used [13].

Although, these protocols give a considerable improvement to security as compared to the original protocols their implementation is difficult and there are still unresolved security issues. Therefore, these prevention techniques should be complemented by intrusion detection techniques. Zhang and Lee [2] present an intrusion detection technique for wireless ad hoc networks that uses cooperative statistical anomaly detection techniques. Each intrusion detection agent runs independently and detects intrusions from local traces. Only one-hop information is maintained at each node for each route. If local evidence is inconclusive, then neighboring IDS agents cooperate to perform global intrusion detection. The authors avoid the reliance on known attack patterns by using an anomaly detection model. However, all such IDS models suffer from performance penalties and high false alarm rates. Furthermore, the authors do not present any performance or detection accuracy analysis of their proposed architecture.

Venkatraman [14] extends the Zhang and Lee model by modifying the protocol so that two-hop information is maintained at each node for each route. The detection scheme uses threshold levels to identify falsified route requests and replies, packet dropping, and route hijacking attacks. However, this scheme requires a modified protocol in addition to requiring an intrusion detection system on each node.

Stamouli proposes an architecture for Real-Time Intrusion Detection for Ad hoc Networks (RIDAN) [6]. The detection process relies on a state-based misuse detection system. In this case, every node needs to run the IDS agent. There is no mention of a distributed architecture to detect attacks that require more than one-hop information.

3 AODV and Attacks in AODV

3.1 Overview of AODV

AODV is an on-demand routing protocol. The route discovery process is only initiated when a node needs to communicate with other nodes. First, the source node broadcasts a route request (RREQ) packet to its neighbors. When receiving RREQ, a node will only unicast route reply packet (RREP) if it has a fresher route to source node or it is the destination of RREQ. Otherwise it will broadcast the RREQ to other nodes until the RREQ reaches the destination. Once the first RREP is received by source node, it can begin data transmission. Route request, route reply, route error are control messages that are sent during the route discovery process for updating the route table of the source, destination, and all the other intermediate nodes. In addition, AODV uses sequence number and hop count to select newer or better routes. Figure 1 illustrates the process of route discovery in AODV. Node S wants to send data to node D. First, node S broadcasts Route Request (RREQ) to its neighbors, node A and node C. Intermediate nodes (A, B, C) set up the reverse path and forward the RREQ packet to other nodes until it reaches to node D. After receiving the RREQ, the destination D unicasts a Route Reply (RREP) packet back along the reverse

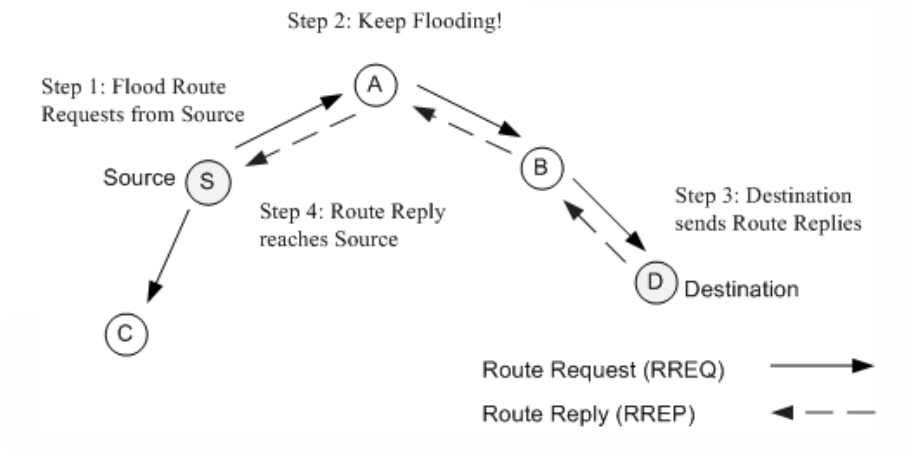


Fig. 1. How AODV works

path to the source. When the source S receives the RREP, it can start sending data to node D.

3.2 Some Attacks in AODV

Currently, AODV does not specify any special security measures. Route protocols, however, are prime targets for impersonation attacks. These attacks include transmitting RREQs with false routing information and denial of service attacks through the repeated broadcast of RREQ messages. Additionally, wireless transmission is inherently insecure. Packets are received by anyone within transmission range, and, if they are not encrypted, they can also be read by anyone.

Denial of Service (DoS): Preventing DoS attacks in an ad hoc wireless environment is extremely challenging. The attacker can create DoS attacks easily by broadcasting too many RREQ messages. Without any assumptions about the mobility of nodes in the network, other nodes simply cannot decide whether such a large number of RREQ messages are because of a DoS attack or broken links due to high mobility, and therefore continue to respond to these fake RREQ messages. Because DoS attacks consume a lot of resources of network, they are especially dangerous for wireless ad hoc networks which have limited resources. AODV itself prevented the sending of same RREQ by checking broadcast id. But the attacker can pass this checking process easily by generating broadcast id increasingly.

Modification of Sequence Number: Since AODV uses sequence number to determine the freshness of routing information and guarantee loop-free routes, the malicious node can send an RREQ or RREP packet with a forged larger sequence number than that of normal nodes and the route will be changed to the malicious node. Once the malicious device has been able to insert itself between

the communicating nodes, it can do anything with packets passing between them such as dropping packets to perform denial of service attack or using its place to perform other attacks (e.g. man in the middle attack). For example, in figure 2, if M sends a RREQ to B with sequence number (200) larger than sequence number (100) of RREQ from A to B, M will take precedence over A.

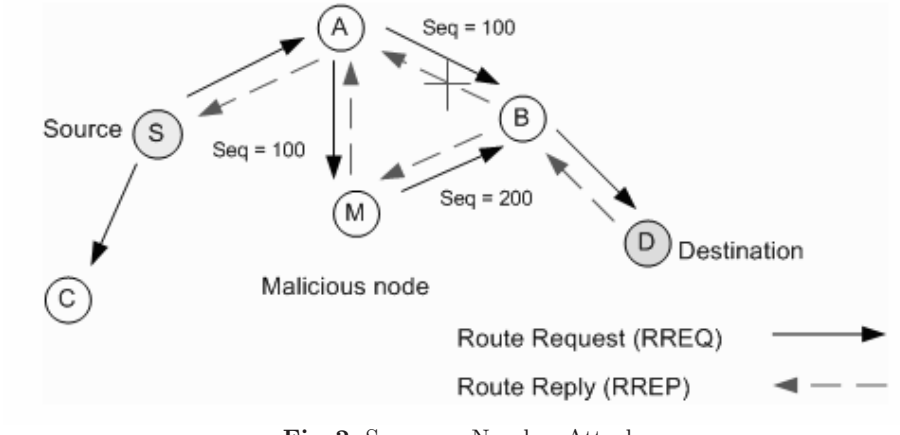


Fig. 2. Sequence Number Attack

4 Intrusion Detection System

4.1 Overview of IDS

Intrusion detection can be defined as the process of identifying malicious activities that may compromise system security. An intrusion-detection system (IDS) is a defense system to help identify, assess, and report unauthorized or unapproved network activities. An IDS does not actually detect intrusions - it detects activity in traffic that may or may not be an intrusion by continuously monitoring the network for unusual activity. Due to the characteristics of wireless ad hoc network such as the lack of fixed infrastructure and concentration points where IDS can collect audit data for entire network like wired network, the high mobility of nodes that makes disconnection, and limited resource, the IDS in MANET must run continually, minimize overhead and the algorithms it uses must be distributed in nature, and should take into account the fact that a node can only see a portion of the network traffic.

4.2 Distributed and Cooperative IDS

Yonguang Zhang and Wenke Lee [2] proposed a distributed wireless intrusion detection and response system architecture, shown in Fig. 3. The individual IDS agents are placed on each and every node. Each IDS agent runs independently and monitors local activities (user, system, communication activities within the radio range). It detects intrusion from local traces and initiates response. If

anomaly is detected in the local data, or if the evidence is inconclusive and a broader search is warranted, neighboring IDS agents can cooperatively participate in global intrusion detection actions.

The methods for collecting data, local detection, and local response are independent of each other and the other IDS agents in the network. A secure communications channel must be standardized among IDS agents to communicate and perform cooperative detection and coordinate global response strategies. This means there must be agreement, or knowledge, of IDS type efficiencies, the meaning of confidence levels if shared, and the costs of attacks on the shared network resources [8].

The intrusion detection system discussed in [2] was anomaly based and provided experimental results exploring the performance of anomaly based detection using different ad hoc routing protocols. It is useful to note that the architecture is not limited to using anomaly or signature based intrusion detection systems, or a hybrid of both. In addition, the focus of [2] was primarily on detection and did not explore the response methodology in significant detail.

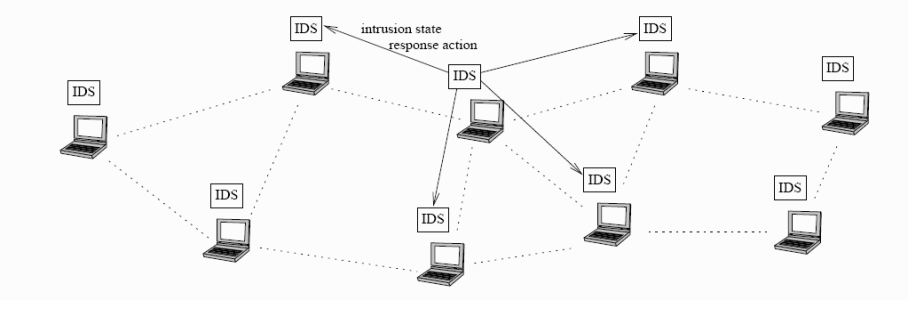


Fig. 3. Distributed Architecture IDS

5 Proposed Solution

We designed and implemented an intrusion detection system for AODV based on the model presented by Yonguang Zhang and Wenke Lee with assumption that the malicious node does not have IDS agent. Our IDS architecture consists of six components (local data collection, local detection, local response, global-cooperative detection, global response and secure communication).

Local data collection: This module is responsible for capturing incoming RREQ packets and updating data log table of each node in network. The data log table stores the newest sequence number of RREQ, number of RREQ packets and their source address correspondingly.

Local detection: After a specific time period, the IDS agent will automatically read data log table. If it detects the number of RREQ packets of some nodes is higher than the defined threshold it will add those nodes to black list. Otherwise the number of RREQ will be reset.

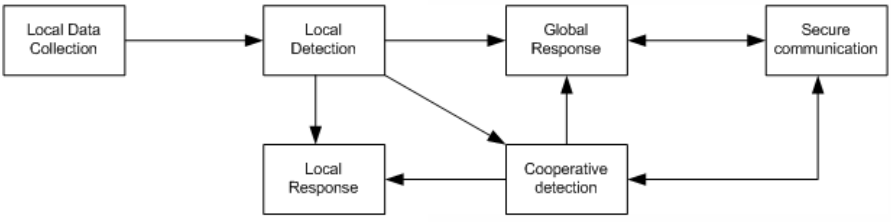


Fig. 4. A model for an IDS agent

Cooperative detection: As shown in figure 2, it is difficult to detect the modification of sequence number in RREQ packet from M by using cryptographic methods if M is an inside node because M can use its pair of keys to generate a legal RREQ packet. To detect any modification of sequence number in RREQ, we utilized a characteristic of AODV protocol, the flooding of RREQ. In AODV, when an intermediate node receives a RREQ, if it has already received a RREQ with the same broadcast id and source address, it drops the redundant RREQ and does not rebroadcast it. Our IDS will check whether the sequence number of rebroadcast RREQ is equal to the sequence number of the same RREQ that stored in current node before AODV drops redundant RREQ packets. If two sequence numbers are different, we can consider it as an anomaly activity. Now the problem is that we don't know exactly which node is malicious: the node comes later or the node has stored sequence number. In order to solve this problem we create a new validate control message, VREQ (Validate Request). Upon detecting an anomaly activity, IDS agent will send VREQ packet with pre-alarm state (state = 0) to node that has higher sequence number because this node can change routing information. This message can be only replied by the IDS agent in each node. Due to our assumption, the malicious node cannot reply this message. If a reply message of VREQ does not arrive within the given period, we can consider that node as a malicious node and add to black list node. For example in Figure 5, when receiving rebroadcasted RREQ message from node M, node A compares the sequence number in this RREQ with the sequence number that has the same broadcast id and source address which it is storing and detects that they are different. The IDS agent in node A will send a VREQ message (alarm state = 0) to node M to validate whether node M is normal. If M can reply a confirm message (VREP - Validate Reply) to A in an interval t, A can know M is normal. If the timer expires without receiving VREP, A can conclude that M is malicious and trigger Global Response module.

Local Response: This module is responsible for dropping any RREQ packets from nodes in black list.

Global Response: Upon being trigged, it will update its routing table to remove wrong information related to malicious node and broadcast a beacon message (alarm state = 1) to its neighbors to notify the presence of sequence number attack. When receiving this beacon, other nodes will update their routing table and rebroadcast it. (See Fig 5)

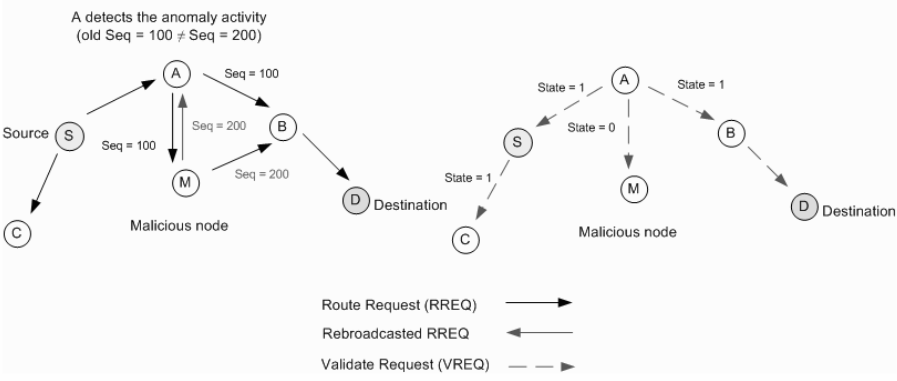


Fig. 5. Cooperative Detection

6 Simulation Results

6.1 Simulation Setup

We have used ns-2 network simulator [5], with CMU Monarch Project wireless and mobile ns-2 extensions, for our implementation. Mobility scenarios are generated by using a random waypoint model with 50 nodes moving in an area of 670m by 670m. Each node independently repeats this behavior and the average degree of mobility is varied by making each node remain stationary for a period called pause time every time before it moves to the next position. We used IEEE 802.11 as the MAC Layer and CBR (Constant -Bit - Rate) as the traffic source. The simulation parameters are summarized in Table 1.

The original AODV protocol in ns-2 is modified to include some attack modules and our proposed IDS. The denial of service attack can be performed by broadcasting continuously RREQ packets with increasing broadcast id into network. In order to carry out the sequence number attack, the malicious node generates a random number between 5 and 200 and adds it to the sequence number that is included in RREQ. The reason that a randomly generated sequence number is used instead of a static one is to make the attack more realistic. In our simulation, the malicious node only drops data packets that it receives.

We used the following metrics to evaluate the performance of our scheme.

Packet Delivery Ratio: is defined as ratio of the total over all nodes of the number of data packets received, divided by the total number of data packets sent from the sources.

Routing Overhead: the number of routing packets transmitted per data packet delivered at the destination.

6.2 Simulation Results

Flooding Attack (denial of service). Figure 6(a) shows that successful packet delivery ratio decreased under flooding attack. However when the IDS

Table 1. Simulation Parameters

Parameters	Values
Nodes	50
Mac Layer	IEEE 802.11
Traffic Model	CBR
Packet size	512
Area	670 x 670
Packet rate	4 m/s
Number of malicious node	1
Simulation time	100 s (Flooding attack)
	300 s (Sequence number attack)

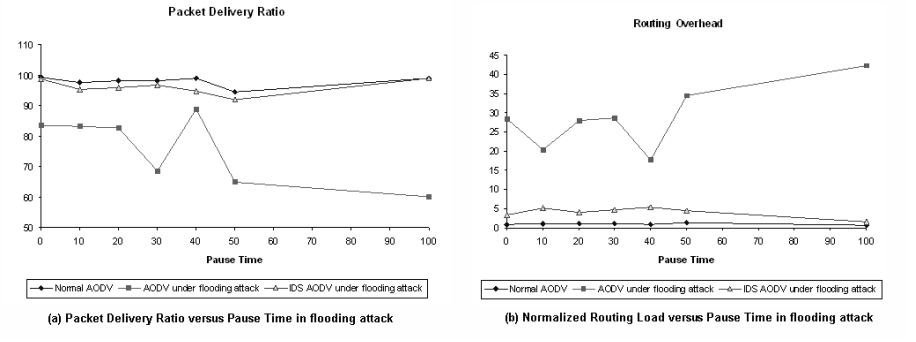


Fig. 6. Simulation results in case of flooding attack

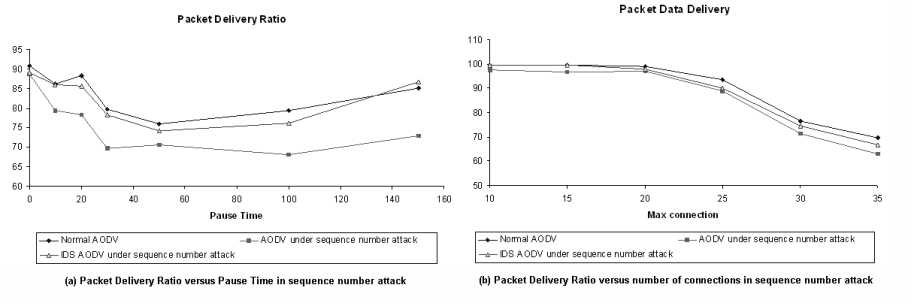


Fig. 7. Simulation results in case of sequence number attack

is enabled the delivery ratio stays within acceptable performance limits almost only 5% lower than it should normally be.

The second evaluation metric for this attack is the normalized routing load. In AODV under attack, the routing overhead is much larger than the normal AODV because the malicious node broadcasts many RREQ. Our approach had a significant improvement. The routing overhead in IDS AODV is only larger than the normal AODV in limit of 5%.

Sequence Number Attack. In order to evaluate the effect of IDS in sequence number attack, we used packet delivery ratio with the variable max connections and pause time. Figure 7(a) shows that the packet delivery ratio is even higher than normal AODV when the pause time increases. As shown in Figure 7(b), when the max connection is low, the ratio of delivery packet of our solution is approximately equal to that in normal AODV and larger than that under attack.

All intrusion detection systems suffer from false alarms that occur whenever the system incorrectly concludes in an alarm but there is actually no malicious behavior in the network. In our IDS, the reason of false alarms is that the reply of VREQ message can be lost.

7 Conclusion and Future Work

We have proposed an intrusion system tool for preventing some internal attacks in AODV. The results of our implementation show that the performance of AODV routing protocol is improved significantly under attacks. The overhead in our solution is minimized because we don't use any cryptographic mechanisms. But our solution has some disadvantages. First, we can't detect attacks which involve the impersonation. Second, the accuracy is decreased when the mobility of network is high. We are working to identify solutions for other attacks and incorporate those with our Intrusion Detection System as well to overcome above shortcomings.

Acknowledgments. This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment). This work was supported by the BK21 Research Fund of Korea.

References

1. C.E Perkins, E.M. Royer, and S. Das: Ad hoc On-demand Distance Vector (AODV). RFC 3561, July 2003.
2. Yongguang Zhang and Wenke Lee: Intrusion detection in wireless ad-hoc networks. Proceedings of the 6th annual international conference on Mobile computing and networking, p.275-283, August 06-11, 2000.
3. Jiejun Kong, Haiyun Luo, Kaixin Xu, Daniel Lihui Gu, Mario Gerla, and Songwu Lu. Adaptive security for multi-layer ad hoc networks. Wireless Communications and Mobile Computing, 2002.
4. Chin-Yang Tseng, Poornima Balasubramanyam, Calvin Ko, Rattapon Limprasittiporn, Jeff Rowe and Karl Levitt: A specification-based intrusion detection system for AODV. Proceedings of the 2003 ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN-2003) Fairfax, Virginia, USA, October 31, 2003.
5. Kevin Fall and Kannan Varadhan: The ns Manual, 2006.

6. Ioanna Stamouli, Patroklos G. Argyroudis, and Hitesh Tewari: Real-Time Intrusion Detection for Ad Hoc Networks. Proceedings of the Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM'05), p.374 - 380, 2005.
7. Paul Brutch and Calvin Ko: Challenges in Intrusion Detection for Wireless Ad Hoc Networks. Proceedings of the Workshop on Security and Assurance in Ad-hoc Networks in Orlando, January 2003.
8. Timothy R. Schmoyer, Yu Xi Lim and Henry L. Owen: Wireless Intrusion Detection and Response - A case study using the classic man-in-the-middle attack. Proceedings of the 2004 IEEE Wireless Communications and Networks Conference, Atlanta, Georgia, USA, March 2004.
9. D. Johnson, D. Maltz, Y. Hu, and J. Jetcheva: The dynamic source routing protocol for mobile ad hoc networks. Internet Draft, Internet Engineering Task Force, March 2001.
10. Y.-C. Hu, A. Perrig, and D. B. Johnson: Ariadne: A secure on-demand routing protocol for ad hoc networks. MobiCom 2002, Atlanta, Georgia, USA, September 2002.
11. Manel Guerrero Zapata: Secure Ad hoc On-Demand Distance Vector (SAODV) Routing. Mobile Ad Hoc Networking Working Group Internet draft. October 2002.
12. A. Perrig, R. Canetti, J.D. Tygar, D. Song: The TESLA broadcast authentication protocol. RSA CryptoBytes, 5(Summer), 2002.
13. Panagiotis Papadimitratos and Zygmunt J. Hass: Secure Routing for Mobile Ad Hoc Networks. Simulation Conference (CNDS 2002), San Antonio, TX, January 27-31, 2002.
14. L. Venkatraman: Securing Routing Protocols for Ad Hoc Networks. Master's thesis, University of Cincinnati, November 2000.

Towards a Security Policy for Ubiquitous Healthcare Systems (Position Paper)

Joonwoong Kim, Alastair R. Beresford, and Frank Stajano

University of Cambridge Computer Laboratory
15 JJ Thomson Avenue, Cambridge CB3 0FD, United Kingdom
{joonwoong.kim,alastair.beresford,frank.stajano}@cl.cam.ac.uk

Abstract. U-Healthcare promises increases in efficiency, accuracy and availability of medical treatment; however it also introduces the potential for serious abuses including major privacy violations, staff discrimination and even life-threatening attacks.

In this position paper we highlight some potential threats and open the discussion about the security requirements of this new scenario. We take a few initial steps towards a U-Healthcare security policy and propose a system architecture designed to help enforce the policy's goals.

1 Introduction

Granny Alice is so pleased with the special "U-Health Shirt" she received this weekend from her son Bob: it monitors her vital signs and sends them wirelessly to a medical centre. Thanks to this ongoing monitoring, she will be able to continue to live in her own flat instead of having to move into one of those horrible, crowded nursing homes. She feels safe, independent and empowered.

On Monday morning, on his way to his office, Bob checks his schedule on his PDA and is pleased but surprised to see that all his meetings have been cancelled. In the office, he finds Carol sitting at his desk: something feels wrong. "Didn't you get any messages?", she asks with a hint of embarrassment. He checks email and discovers that he has been transferred to the post room and that Carol, not him, is now leading the department. There is also another message, from healthcare services, booking him in for a detailed medical check-up: his body sensors show high levels of nitric oxide, suggesting the possibility of cancer.

Bob feels dizzy. He knows that, even if the check-up later reveals no cancer, he has now lost his prestigious position in the company. There are too many candidates for that post: his own turn may only come back in several years. While he drives back home, he nervously removes all of his body sensors, only regretting he can't easily get rid of the implanted ones. Soon his mobile phone rings: a voice mail tells him he should replace and reconnect his sensors or he will lose his insurance discount, and that he should contact customer service if this is a sensor failure.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 263–272, 2007.

Despite the obvious exaggerations, with the rapid evolution of sensor technologies most of the pieces of the above scenario are rather close to feasibility. Healthcare projects using body sensors as remote monitoring devices are already under way[1]. Body-sensor-based 24-7 monitoring will enable remote diagnosis without the patient having to visit a hospital, thus providing cheaper healthcare services. At the same time, more detailed body sensor data, combined with data from infrastructure sensors, will provide a "life log" or "activity diary" of the patient. If such sensitive personal data is shared among interested parties—employers, insurance companies, drug companies and the government, to name a few—the possibility of abuse is great.

Most readers (except perhaps those who were recently downsized) will point out that real companies can't afford to be as ruthless and nosey as Bob's if they wish to retain talent; but Bob might have given his explicit consent to his doctor about sharing his body sensor data with his employer, to prove to the employer that it was safe to promote him to division head because he was a healthy employee who would be able to work hard and handle severe stress levels. And he could have also granted access to his health insurance company for a discount. Or his employer might have done it for him, assuming Bob lives in a country where the employer customarily pays for the employee's health insurance.

The fact that the inappropriate disclosure of private medical information can harm the patient has been clear since at least the 4th Century BC, as the Hippocratic Oath indicates. But what exactly are the security requirements in this age of increasing computerization? Ten years ago, Anderson's BMA Security Policy [1] described the protection goals of clinical information systems: its motivation was that storing patient medical records on a nationwide distributed computer system endangered the principle of patient consent and increased the possibility of data aggregation. U-Healthcare, in turn, brings new threats and vulnerabilities, as illustrated by Bob's story above, which are not all adequately covered by the BMA policy.

The first contribution of this position paper is to point out such new threats and to open up the debate about security for U-Healthcare. Secondly, in order to clarify the protection goals, we propose and discuss some possible principles for a U-Healthcare security policy. Thirdly, we suggest a system architecture consistent with the proposed policy.

1.1 Terminology

Electronic Healthcare Systems, or *(Electronic) Clinical Information Systems* are the existing healthcare information systems that use networked computing systems for recording and accessing medical records. *Ubiquitous Healthcare Systems*, instead, adopt ubiquitous computing as an enabling technology, with sensors

[1] See for example the Codeblue paper [14] and the web sites of the PIPS, MyHeart and Proactive Health projects (`http://www.pips.eu.org/`, `http://www.hitech-projects.com/euprojects/myheart/` and `http://www.intel.research/prohealth/` respectively.

monitoring the patient continuously, and include Wellness Systems, Disease Care Systems, and Independent Living Systems [11].

We prefer to say *Patient* rather than *User* because a *Clinician*, too, is a user of the U-Healthcare system. There are several *Healthcare (Service) Providers*, including but not limited to *Clinicians* and GPs: for all of them we may also use the term *Caregiver*. The more general terms are preferable if we consider that the clinics can be replaced by other healthcare services such as gyms and healthcare web services.

As for sensor devices, there are body sensors and infrastructure sensors. Examples of the latter include scales and sensors of ambient temperature, light or movement. These sensor devices transfer data to base stations such as PDAs, Smartphones and PCs. The union of these sensors and base stations forms a *Personal Healthcare System* which is used and controlled by an individual patient, or by some trustee on behalf of the patient. The sensor data is then transferred to a *Clinical System* for further analysis, if needed.

Lastly, for economy of expression, we will use the same gender convention as the BMA Policy [1]: the clinician is female, and the patient male.

2 Threats and Vulnerabilities

Currently, most patient medical records are accessed through a standard desktop workstation which requires the caregiver to be in a particular place at a specific time. Therefore the environment and architecture of the hospital or surgery provides some additional social control to prevent unauthorised access to medical data. The use of PDAs and laptops in ubiquitous healthcare to access patient records on the move, or from a remote location, is likely to improve the timeliness of patient care, but may represent a greater temptation for an underpaid caregiver who is offered a bribe by a pharmaceutical company or a private investigator.

Many monitoring systems in hospitals today use physical access control to provide privacy. For example, a heart rate monitor is typically situated beside the patient, and the device only provides data to a caregiver who is standing near the machine. In addition, data might only be available in real-time—any historical data is lost unless a caregiver explicitly records it separately. Ubiquitous healthcare extends the computerisation of medical records to the domain of monitoring and diagnosis—monitored data will be recorded and the historical record used in subsequent analysis.

In a ubiquitous healthcare system, remote access to patient data by a caregiver may become normal. Because sensors will be cheap and portable, a personal healthcare system is likely to be used to record sensitive medical data continuously during everyday life, not just whilst the patient is in a hospital. This record of data will be of great interest to third parties, such as insurers, medical researchers and employers and therefore, without adequate control, the ability to data mine this resource becomes compelling. The recorded data is also likely to contain many personal facts (such as dates, times and durations of the patient's

sexual intercourses) which, whilst irrelevant to any specific medical diagnosis, are hard to remove from the dataset without reducing the quality of the sensor data itself.

It is also likely that a caregiver, or even a computer program, will be able to remotely administer drugs through a body area network. This scenario requires integrity of sensor data and rules used to decide when to administer drugs.

In current out-patient practice, a caregiver will typically engage in a short consultation with the patient and ask a series of questions about his health. The questions must necessarily be on a level that the patient understands. In this scenario the patient is able to query the relevance of any question with the caregiver, ask what the consequences of failing to answering the question might be and, if the patient feels it is necessary, provide a false answer. In contrast, ubiquitous monitoring of physiological signs will generate a large amount of data which the patient cannot interpret without help: the dataset will be too large for manual analysis, and it is likely to require a good deal of technical skill to understand.

The BMA Policy [1] was concerned that the aggregation of many patient records may lead to abuse. In ubiquitous monitoring, a combination of sensor readings of a *single* patient may also be problematic. For example, the symptoms of depression may be inferred from changes in body weight and sleeping patterns, even if this conclusion was not the original intention of monitoring. Disclosure of such medical histories might be unwelcome or used against the patient. For example, the medical history of a US politician who had suffered from depression was disclosed just before an election [18].

A patient may also configure a body sensor network to record data for other purposes. For example, Bell's MyLifeBits project [7] records a wide variety of audio, visual and location data which can be used to aid memory recall of specific events. The patient will probably not want to give unconditional access to these data, yet a caregiver may be able to give a better diagnosis with access to some information contained within the dataset.

The additional problems presented by a ubiquitous healthcare system can be summarised into four broad areas:

Ubiquitous access: easy remote access to data amplifies the vulnerability of medical records to unauthorised access;

Ubiquitous monitoring: monitoring and diagnosis will be computerised and sensors will travel with the patient wherever he goes, potentially providing the caregivers with the ability to record, search and archive sensor data remotely;

Ubiquitous care: patients will receive tele-prescription and tele-infusion of drugs and receive professional advice remotely;

Ubiquitous sensor data: The collection and recording of medical sensor data will be useful to researchers but may contain many personal facts.

For the extensive security analysis, we need to consider confidentiality, integrity and availability. In a sense, confidentiality is more related with privacy, and the latter two with safety and dependability. However, in the remainder of

this paper we focus on addressing the privacy issues of ubiquitous monitoring as a starting point. Because we believe this will become the most prominent part of ubiquitous healthcare systems in the near future, and is something which is missing from the existing discussions on security in healthcare systems, such as those found in the BMA Policy [1]. Studying the availability and integrity of U-healthcare systems will be part of the future work.

3 Towards a Security Policy

3.1 Monitoring

Traditionally, health status was measured either directly by the caregiver or the patient; more recently such measurement may have received some form of technological assistance. Such collected data is usually analysed in real-time and is summarised and discussed before being recorded. In contrast, when a ubiquitous monitoring environment is used, computing devices may create a permanent record in much greater detail. To protect the privacy of the patient we propose:

> **Principle of self care:** Data collected in a ubiquitous monitoring environment must be processed and stored on a personal healthcare system under the sole control of the patient. No sensor data shall leave the personal healthcare system without the patient's consent.

This principle reflects our current notion of healthcare: a patient will contact a caregiver only after a medical problem is discovered and caregivers only receive medical facts from the patient or perform an examination with the informed consent of the patient.

In some cases we will want the ubiquitous monitoring environment to analyse, report and automatically execute actions based on the sensor data. For example, a diabetes patient may use a body sensor network to keep him informed of his current glucose level and perhaps even automatically trigger the delivery of insulin. In this case, the principle of self care means that glucose level readings and insulin delivery must operate within the personal healthcare system and run independently of all clinical systems under the control of the caregiver. It is worth noting that this type of autonomous operation may be sensible from a safety and reliability perspective too.

3.2 Consultation

There will be times when a patient will seek the advice of a caregiver. This might happen at pre-defined intervals, whenever the personal healthcare system reports a potentially life-threatening reading, or during an emergency. In these cases, the patient (or, in the case of the young or seriously ill, their next-of-kin) will require some help interpreting the data recorded by the personal healthcare system. Since the patient cannot know what facts the sensor data contains, he

cannot give his *informed* consent to the release of all sensor data directly into his medical record. Therefore, to protect the privacy of the patient we propose:

> **Principle of non-disclosure:** The patient may transfer sensor data from his personal healthcare system into a temporary repository which is also accessible by a caregiver. Only data useful in assessing the state of health of the patient is transferred. By default, data may not be transferred out of the repository, which shall exist for a limited time.

In practice it is impossible to delete all traces of the analysis since the caregiver and patient may mentally recall some of the information. Nevertheless, this principle means that at the end of any consultation between a patient and the caregiver, there should be no electronic record of either the raw sensor data or any derived data.

Some forms of analysis may require several caregivers to collaborate and this might make it difficult to arrange for all the specialists and the patient to meet at once. In this case, the principle of non-disclosure means that as the data is analysed, the patient is kept informed of what data is collected from his personal healthcare system. It is important to limit both the amount of time data can be kept, and the number of caregivers who may access the repository. If this is not the case, the lifetime of the repository may last as long as the lifetime of the patient, and it becomes a medical record in all but name. The length of time data can be held in a repository will depend on the medical condition under analysis; for complex situations this is something which needs to be reviewed by caregiver and patient at regular intervals.

3.3 Permanent Records

The principle of non-disclosure means that, whilst caregivers can analyse data from a personal healthcare system, they cannot maintain a summary of the results of the analysis. Such a record may be needed to provide a prescription, charge a fee or provide continuity of care. We believe it is important that the patient controls and understands the meaning of any data which is written to a permanent medical record as the result of analysis in a temporary repository.

> **Principle of limitation and necessity:** Any results from the analysis of sensor data stored in a temporary repository may only be transferred into the patient's permanent medical record if the patient's *informed* consent for the transfer is obtained and the long-term storage of such data is required to protect the patient's future well-being.

Or in other words: record the outcome of the analysis (if it is relevant and useful) rather than the raw sensor data itself. In some sense this principle is nothing new: caregivers have previously summarised information written into a medical record rather than transcribing the entire conversation. The aim of this principle is to prevent the raw sensor data from being written into a permanent

record; this is important since raw data may contain lots of hidden personal facts the user did not consent to releasing, but may be obtained later by data mining.

In many cases, data may be summarised on the personal healthcare system itself. For example, a diabetic may provide the caregiver with a summary of the highs and lows of their glucose level. In other cases, the caregiver will need to see the raw data: an electrocardiogram (ECG) trace provides much more information than simply the heart rate—the data requires an expert to interpret it.

4 Architecture

In the last section we derived a security policy which provides the patient with a method to control access to any sensor data recorded by a personal healthcare system. We believe, from a computer science perspective at least, that it is practical to build a system which conforms to this security policy. In order to support a temporary repository, we envisage a software mediator which logically sits between a personal healthcare system used by the patient and any clinical system used by the caregiver. The concept of an intermediate component exists already in many diverse research areas of computer science, and includes proxies, agents, guardians, Trusted Computing Bases, etc.

The mediator (Figure 1) should provide an interactive environment in which a patient and a caregiver can explore the data recorded by the body sensor network, extract the relevant medical facts from the collected data and, with the patient's informed consent, append those facts to the medical record. In order to meet the criteria set in the security policy, it is important that the patient be in control and be able to limit: (1) the raw sensor data sent to the mediator and (2) the derived facts transferred from the mediator to the medical record. Obviously it is paramount that all data stored by the mediator be deleted at the end of any period of consultation.

5 Related Work

The BMA Security Policy [1] was developed by Anderson for the British Medical Association to protects patient records in clinical informations systems. It is based on nine principles, including access control, consents, audit, information flow and data aggregation. A few updates [2,3,4] also appeared.

In the 1990s, threats to privacy in Electronic Patient Record (EPR) were widely recognized in the U.S. As a result, a few reports [8,13,17] about security in EPR were released. Besides these works, most security research in healthcare systems [16,19] have been based on variants of the Role Based Access Control (RBAC) model. Gostin [9] discussed healthcare information from an ethical perspective, while Health Privacy Project [15] provided a small collection of privacy-breaching incidents in U.S. medical systems.

Many researchers have worked on privacy in ubiquitous computing environments, including at least [12,10]. Langheinrich [12] proposed the infospace concept as the trust boundary, and the privacy tag. Jiang et al. [10] discusses how

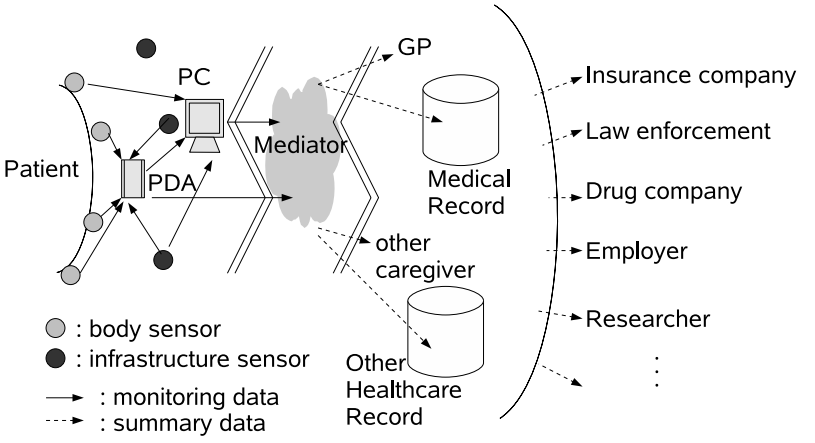


Fig. 1. Ubiquitous Healthcare Systems Architecture Option

the user is notified about data collection by sensors and how a policy can be negotiated. However these two privacy frameworks refer to a general ubiquitous computing or context-aware computing context and are not directly applicable to healthcare information systems. More relevantly, Bohn [6] analysed the dependability issues in U-Healthcare, and Beckwith [5] discussed the perception of privacy based on the case study of a sensor-rich, eldercare facility.

To the extent of our knowledge there has not yet been any proposal of a formal security policy to regulate ubiquitous healthcare systems, along the lines of what the cited BMA policy did for in clinical information systems. Hence our work.

6 Conclusion

U-Healthcare introduces great convenience, but at the same time equally great risk. The shift to 24/7 patient monitoring via body sensors is not just an incremental improvement over the existing practice: it is a qualitative step change. So is the shift to remotely-activated drug dispensers implanted in the patient's body. The main message of this paper is that such major paradigm shifts demand a rethinking of the security and privacy aspects: solutions that were appropriate for yesterday's situation are insufficient for tomorrow's. We pointed out some of the new threats.

We believe it is still too early to propose a complete technology solution: what is most needed at this stage is instead an informed debate. We wish to engage all parties, including clinicians and patients, and understand what is acceptable and desirable before the coming generation of U-Healthcare systems is deployed. This is why we presented our principles in natural language rather than using equations or formal security terminology. There will certainly be tension between security and usability, between patient privacy and clinician convenience, and

we don't presume to have got the balance exactly right at the first attempt; we solicit opinions and corrections, particularly from practicing clinicians, but we all need to understand the issues at stake.

In this context, a security policy is first of all an instrument of communication. By writing down, at least as a working draft, the protection goals of future U-Healthcare systems, we allow the community of stakeholders to think, agree, disagree and debate. We hope that the outcome of this process will be a strong specification upon which to build U-Healthcare systems that, like Isaac Asimov's brilliantly imagined robots, can never be misused to cause harm to their patients.

References

1. Ross Anderson. *Security in Clinical Information Systems*. BMA Report. British Medical Association, Jan 1996. ISBN 0727910485. `http://www.cl.cam.ac.uk/~rja14/Papers/policy11.pdf`.
2. Ross Anderson. "A security policy model for clinical information systems". In "IEEE Symposium on Security and Privacy", 1996. `http://www.cl.cam.ac.uk/~rja14/Papers/oakpolicy.pdf`.
3. Ross Anderson. "An Update on the BMA Security Policy". In "Cambridge workshop on Personal Information — Security, Engineering and Ethics", 1996. `http://www.cl.cam.ac.uk/~rja14/Papers/bmaupdate.pdf`.
4. Ross Anderson. "Healthcare Protection Profile — Comments", 1998. `http://www.cl.cam.ac.uk/~rja14/Papers/healthpp.pdf`.
5. Richard Beckwith. "Designing for Ubiquity: The Perception of Privacy". *IEEE Pervasive Computing*, **2**(2):40–46, 2003.
6. Jürgen Bohn, Felix Gärtner and Harald Vogt. "Dependability Issues of Pervasive Computing in a Healthcare Environment". In "Security in Pervasive Computing 2003", vol. 2802 of *Lecture Notes in Computer Science*. 2004. `http://www.vs.inf.ethz.ch/res/papers/bohn_pervasivehospital_spc$_$2003_final.pdf`.
7. Steven Cherry. "Total Recall". *IEEE Spectrum*, **42**(11), Nov 2005. `http://www.spectrum.ieee.org/nov05/2153`.
8. Paul D. Clayton (ed.). *For the Record: Protecting Electronic Health Information*. National Academy Press, 1997.
9. Lawrence Gostin. "Health Care Information and the Protection of Personal Privacy: Ethical and Legal Considerations". *Annals of Internal Medicine*, **127**(5), Oct 1997. `http://www.annals.org/cgi/content/full/127/5_Part_2/683`.
10. Xiaodong Jiang and James A. Landay. "Modeling privacy control in context-aware systems". *IEEE Pervasive Computing*, **1**(3), 2002. `http://guir.cs.berkeley.edu/projects/ubicomp-privacy/pubs/infospace.pdf`.
11. Ilkka Korhonen, Juhan Pärkkä and Mark Van Gils. "Health Monitoring in the Home of the Future". *IEEE Engineering in Medicine and Biology Magazine*, **22**(3):66–73, May 2003.
12. Marc Langheinrich. "Privacy by Design — Principles of Privacy-Aware Ubiquitous Systems". In "Ubicomp 2001", 2001. `http://www.vs.inf.ethz.ch/publ/papers/privacy-principles.pdf`.
13. William W. Lowrance. *Privacy and health research a report to the U.S. Secretary of Health and Human Services*. U.S. Department of Health and Human Services, 1997.

14. David Malan, Thaddeus Fulford-Jones and Matt Welsh. "CodeBlue: An Ad Hoc Sensor Network Infrastructure for Emergency Medical Care". In "International Workshop on Wearable and Implantable Body Sensor Networks", Apr 2004. http://www.eecs.harvard.edu/~mdw/papers/codeblue-bsn04.pdf.
15. Health Privacy Project. "Meidcal Privacy Stories", Nov 2003. http://www.patientprivacyrights.org/site/PageServer?pagename=True_Stories#True_Stories.
16. Jason Reid, Ian Cheong, Matthew Henricksen and Jason Smith. "A Novel Use of RBAC to Protect Privacy in Distributed Health Care Information Systems". In "Eighth Australasian Conference on Information Security and Privacy (ACISP 2003)", 2003.
17. Thomas C. Rindfleisch. "Privacy, information technology, and health care". *Communications of the ACM*, **40**(8), Aug 1997.
18. A. Rubin. "Records No Longer for Doctors' Eye Only". *Los Angeles Times*, 1 Sep 1998.
19. Longhua Zhang, Gail-Joon Ahn and Bei-Tseng Chu. "A role-based delegation framework for healthcare information systems". In "The Seventh ACM Symposium on Access Control Models and Technologies (SACMAT'02)", 2002.

Architecture of an LBS Platform to Support Privacy Control for Tracking Moving Objects in a Ubiquitous Environments

JungHee Jo, KyoungWook Min, and YongJoon Lee

Telematics-USN Research Division,
Electronics and Telecommunications Research Institute,
161 Gajeong-dong, Yuseong-gu, Daejeon, Korea

Abstract. Due to the rapid growth of the mobile communication technology, obtaining in-detail location information by combining GPS and networks is now available. In such an environment, legal steps are needed to ensure the protection of personal location information on the one hand, and also to regulate the use of personal location information for public purposes on the other hand. In Korea, based on such considerations, the Ministry of Information and Communication (MIC) established and announced a new LBS law. As the new law is proposed, it is essential to make it applicable to the existing LBS technology. In this paper, we propose our solution for applying such a new law to the LBS platform. To do this, we propose architecture for an LBS platform conforming to the new LBS law and the specific scenario of privacy control for tracking moving objects. Especially, we suggest new privacy control mechanisms using privacy profile, such as time clocking, time blurring, area cloaking, and area blurring.

1 Introduction

Many studies have shown that LBS subscribers seriously care about their privacy and are wary of intrusions. For this reason, mobile telecommunication operators and marketers are sensitive about the way they handle the localization of others [1]. As privacy issues emerge as potential inhibitors to the commercial LBS, a privacy control mechanism for location information became a critical factor for mobile telecommunication operators to protect their subscriber's privacy. The privacy control mechanism should verify each subscriber's privacy profile and determine whether or not to allow dissemination of their information, and how much of their information is allowed to be shared. According to the research reports from Frost & Sullivan [2], it is mandatory to include such privacy control mechanisms on the LBS platform.

However, concrete privacy regulations to ensure competition in the LBS market aren't yet established so some mobile telecommunication operators are still experienced a leakage of their subscriber's real-time location information. In order to maintain a consistent privacy regime, wireless location information collection must be regulated across the diversity of spectrum and technologies. In

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 273–282, 2007.

other words, legal steps of privacy protection are needed to prevent abusing personal location information. By considering such situations, the Ministry of Information and Communication (MIC) of Korea established and announced the new LBS law. The main purposes of this law are: to enhance the level of personal location information protection, to support the activation of LBS, to create boundaries for the use personal location information for public purposes, and to create an environment for the operation of LBS business. The law prevents location information of LBS subscribers from being indiscriminately made use of by the LBS providers. So, any attempt to collect information on the location of LBS subscribers is banned without the approval of those LBS subscribers. To apply such a law on the existing LBS entities, including the LBS platform, it is necessary to adopt new privacy control mechanisms.

The contributions of this paper are as follows: (1) we suggest a way to adopt the new LBS law on the existing LBS platform, (2) we propose the architecture for the LBS platform conforming to the new law in order to support privacy control while tracking moving objects, and (3) we explain each element of the LBS platform and describe scenarios of the tracking of a specific moving object using platform. Especially, we suggest new privacy control mechanisms using a privacy profile, such as time clocking, time blurring, area cloaking, and area blurring. The paper proceeds as follows: Section 2 explains the motivation and requirements to apply the new law for location information protection. Section 3 suggests the architecture of an LBS platform to support privacy control for moving object tracking. The final section provides concluding remarks and offers suggestions for future research.

2 Motivation and Requirements

This section is divided into three subsections. In subsection 2.1 this paper explains the new LBS law for privacy control on location information and in subsection 2.2 this paper suggests the way to apply the new law for regulation of location information. In subsection 2.3, this paper introduces the technology of moving object tracking using the LBS platform.

2.1 New LBS Law for Privacy Control on Location Information

Since location information in databases is subject to a wide range of risks, it requires appropriate privacy and security measures. Even if wireless carriers have long argued that their security solutions were designed to protect subscriber's information from any possible hacking, location information leakage have still occurred. The risks include misuse by insiders, unintentional or mistaken disclosure, and access by unauthorized individuals. In extreme cases, improper disclosure of location information could place a person in physical danger; location information could be misused by stalkers or in domestic violence cases [4]. Actually, one of the large telecommunication companies in Korea experienced a leakage of real-time location information through the internet caused by less-than-rigorous

management of user database [3]. Korea's mobile carriers have amassed a vast of amount of subscriber information as the number of users has surpassed the 32 million mark and such accidents raise questions associated with issues related to electronic and mobile privacy.

To solve the problem which is related with privacy control on location information, the Ministry of Information and Communication (MIC) in Korea announced the new LBS law for location data protection. It is aimed at protecting the privacy of mobile phone users and offering specific guidelines for the industry. The major elements of the law are as follows. First, for the location privacy protection, location information operators, such as a mobile telecommunication company and an LBS provider who provide a location related service using a subscriber's location information, must specify legal obligation on the agreement, when they collect personal location information or provide services, and they must obtain the subscriber's consent of doing this. In other words, they can only use subscriber's private information for agreed purposes and can not provide it to a third party. Furthermore, they should destroy personal location information immediately when they have achieved the specified purpose. Second, for the commercialization of LBS, the location information operator should be granted a license and the LBS provider should report to the licensing authority before setting up business, and finally it is compulsory to destroy the collected location information and not to make further use of it. Third, for the public LBS such as emergency services, the emergency rescue department can request personal location information and give danger warnings to the location information operator and this operator cannot disregard it.

2.2 Applying the New LBS Law on the LBS Platform

As the new LBS law is suggested, applying the new LBS law to the existing LBS technology is essentially required. This paper shows our solution to apply such new law to the LBS platform.

There are four basic but important constituent technologies of the LBS platform: security, authentication, authorization, and having a location privacy portal. First, the LBS platform requires a security mechanism to protect location information from being accessed by unauthorized users. Also, the authentication mechanism is required to permit the location information accessing rights to an LBS provider who supplies subscribers of mobile devices personalized services tailored to their current location. In addition, the LBS platform is in need of an authorization process to decide if a person, program or device is allowed to have access to the location data and system. Lastly, the LBS platform should provide location privacy portal site for users to set privacy related profiles and to track the monitoring records of location requests. To apply the new law to the LBS platform, additional mechanisms of privacy control should be considered, such as detailed access control mechanism to monitor and block the location information by considering time, users, services, and areas. This paper suggests how to efficiently apply the new LBS law on the existing privacy control mechanisms related with location information.

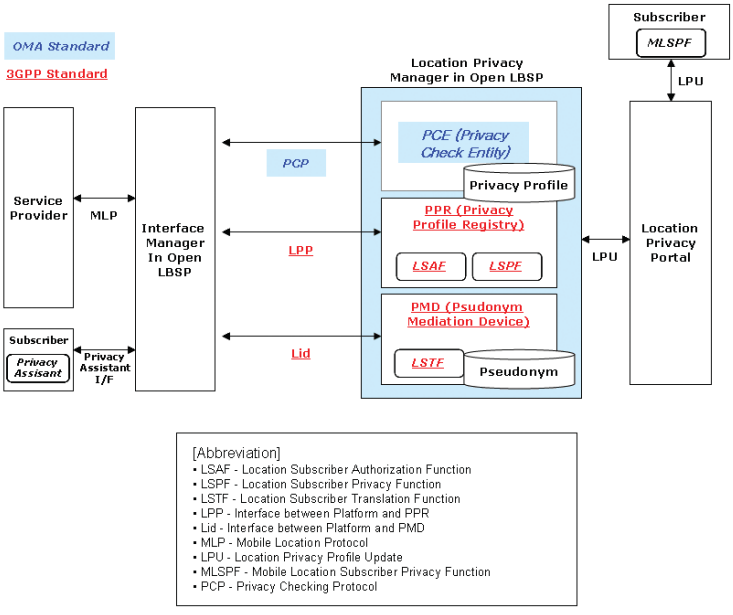


Fig. 1. Convergence LBS platform architecture - Bridging the OMA and 3GPP standards

Fig. 1 shows the suggested convergence architecture of the location privacy manager in the open LBS platform. The OMA (Open Mobile Alliance) [6] and 3GPP (3rd Generation Partnership Project) [7] standards are bridged on LBS platform. Various privacy control entities, including service providers, LBS platforms, and location privacy portal, consist of LBS architecture: highlighted parts represent OMA standard and underlined parts represent 3GPP standards. Basically, the LBS platform needs to integrate heterogeneous LBS applications that run across the Internet on the heterogeneous wireless network. Therefore, supporting standardized privacy control entities could make it possible to achieve interoperability as well as privacy control in the LBS domain.

The location privacy manager in the LBS platform is categorized as three parts: PCE (Privacy Checking Entity), PPR (Privacy Profile Register), and PMD (Pseudonym Mediation Device). In case of PCE, it contains privacy rules for positioning targets and is responsible for checking the privacy settings of the target. The PPR is responsible for maintaining personal privacy profile and interoperate with platforms. This handles pseudonym and verifies pseudonyms to verinyms. Pseudonym is a fictitious identity, which may be used to conceal the true identity (i.e. MSISDN and IMSI) of a target device from the requestor. The brief scenario of privacy control on location information using pseudonym is as follows. First, the terminal requests location services to the proxy. Then the proxy requests the terminal's pseudonym for operator' After

obtaining the pseudonym, the proxy requests services to LBS provider. At present, the LBS provider doesn't know the true identity of the terminal and sends the requested services to the proxy. Then, the proxy obtains verinyms by converting the pseudonym and sending services to the terminal. By doing this, the terminal could conceal own true identity to the LBS provider.

2.3 Moving Object Tracking Using the LBS Platform

In order to make a high quality LBS platform, major functionalities should be efficiently provided. The moving object tracking is one of major functionalities and there are a number of ways such as a multi-modal approach [5]. One approach for moving object location tracking is location polling. This process is classified into server-based location polling and terminal-based location polling, based on the position of the location acquisition agent [8]. The important point in tracking is how to most efficiently capture the positions of a large population of moving objects with precision. Basically, efficient tracking reduces the system load, including network communication and server loads.

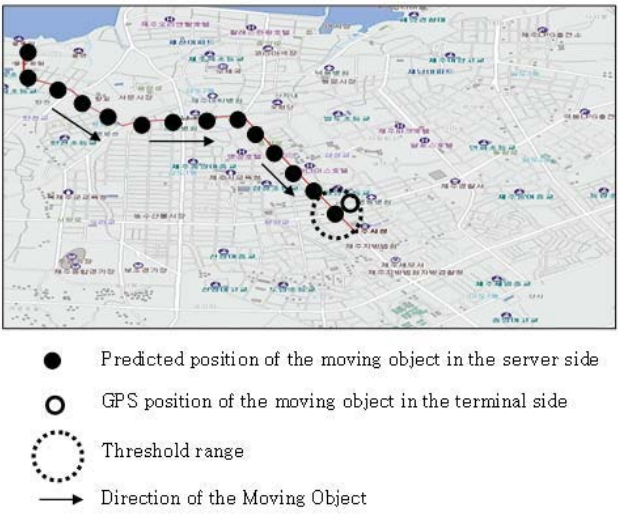


Fig. 2. Road network based moving object tracking

As the capabilities of the terminal increase, many more efficient terminal-based location polling techniques can be introduced [9, 10]. One of the techniques is moving-object tracking based on road networks as shown in Fig.2. That tracking technique is terminal-based location polling and allows the location server to predict the current position of a moving object with minimum communication with the terminal. Initially, the terminal obtains its real position from the GPS receiver and establishes the connection with the location server and sends its real

position to the server. After receiving the terminal's real position, based on the initial position, the server begins to predict the position of the moving object, and the moving object also begins to predict its own position using the same prediction algorithm on the server. The moving object continuously compares the predicted position to its actual position and sends an update to the server when the distance between the two positions exceeds a given threshold, at which time the server restarts its prediction using the newly updated position information. This is more efficient than the traditional terminal-based periodic location polling because the location update from the object to the server is sent only if a predefined threshold is exceeded. By reducing the update rate, this process enhances the overall performance of location tracking. The LBS platform in this paper suggests the way to support both enhanced terminal-based polling and server-based polling techniques.

3 The Architecture of an LBS Platform to Support Privacy Control for Tracking Moving Objects

This section explains our approach of a privacy control for tracking moving object. The LBS platform that we proposed is composed of three main parts: tracking, privacy, and other platform related functionalities such as client interface, service interface, QOS (quality of service), and gateway connectors.

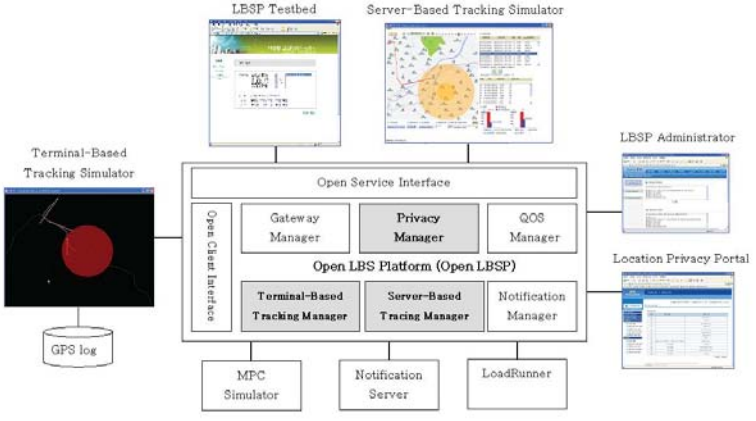


Fig. 3. The architecture of an LBS platform

As shown in Fig.3, this platform supports two kinds of moving object tracking using the Terminal-Based Tracking Manager and Server-Based Tracking Manager. Also, Privacy Manager consists of all privacy control related modules such as PCE, PPR, and PMD. The functionalities of Privacy Manager based on the new LBS law can be classified as follows:

1) Access Control Mechanism for Subscriber's Temporary Veto Right. According to the new LBS law, a subscriber could have a temporary veto right to the use of own location information. To satisfy this rule, we suggest access control mechanisms on location information using a subscriber's privacy profile. This paper suggests four kinds of access control based on the time and area: Time Cloaking, Area Cloaking, Time Blurring, and Area Blurring. In case of Time Cloaking, the subscriber can designate the specific time period to conceal his or her location information during the time. For example, as shown in Fig.4 (a), the user C can hide his or her location information from User A, from t1 to t2. In the same way, the user C can hide his or her location information from User B, from t2 to t3. However, Service A and Service B can always track User C's location regardless the predefined time period. Service A and Service B stand for the location related services such as traffic information services or friend finder services.

(a) Time Cloaking of User C

	Service A	Service B	User A	User B
All	O	O	-	-
[t1, t2]	-	-	X	-
[t2, t3]	-	-	-	X

(b) Area Cloaking of User C

	Service A	Service B	User A	User B
All	O	O	-	-
Geofence 1	-	-	-	X
Geofence 2	-	-	X	-

(c) Time Blurring of User C

	Service A	Service B	User A	User B
All	-	1km	1km	-
[t1, t2]	-	-	1km	-
[t2, t3]	-	-	-	1.5km

(d) Area Blurring of User C

	Service A	Service B	User A	User B
All	-	1km	1km	-
Geofence 1	-	-	1km	-
Geofence 2	-	-	-	2km

O : Allow to provide the location
X : Not allow to provide the location

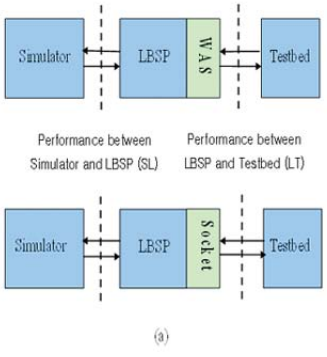
Fig. 4. Proposed access control matrix

On the other hand, using Area Cloaking, the subscriber can conceal his or her location information if he or she enters the predefined area. For example, as shown in Fig.4 (b), the user C can hide his or her location information from User A, if User C exists within geofence 2. In the same way, User C can hide his or her location information from User B, if User C exists within geofence 1.

However, Service A and Service B can always track User C's location regardless of predefined geofence. In case of Time Blurring, if a subscriber doesn't want to provide his or her exact location during the predefined time period, he or she can blur his or her accurate position. For example, as shown in Fig.4 (c), User C can cause a deviation of 1km from an actual GPS position if User A wants to track his or her location, from t1 to t2. On the other hand, using Area Blurring, the subscriber can blur his or her accurate position according to the predefined area. For example, as shown in Fig.4 (d), User C can cause a deviation of 2km from an actual GPS position if User B wants to track his or her location, if User C exists within geofence2. Using these access control mechanism, a subscriber's temporary veto right in the new LBS law can be implemented on the LBS platform. The privacy profiles for each subscriber are maintained in Location Privacy Portal, which is shown in Fig. 3.

2) Pseudonyms/Verinyms Mechanisms for Subscriber's Privacy Protection. According to the new LBS law, the location information operator can use privacy information only for agreed purposes and can not provide it to a third party. To protect a subscriber's location information from a third party, this platform adopted pseudonyms/verinyms mechanisms using the PMD (Pseudonym Mediation Device) in the 3GPP standards. The specific scenario is as follows: 1) if an LBS provider requests another person's location information using his or her mobile phone number, the LBS platform converts the phone number to pseudonym using the PMD. 2) With the pseudonym, the platform obtains location information from the mobile positioning center. Then, 3) the LBS platform obtains verinyms by converting pseudonym and sends location information to the LBS provider. At present, any third party couldn't hack the person's privacy information such as the phone number and location information.

3) Tracking Mechanisms for the Subscriber's Inspection Right. According to the new LBS law, the subscriber has the right to know and request the evidence of operating personal information. At the same time, mobile telecommunication operators and LBS providers have an obligation of automatic recording when using, providing, collecting, and approaching location information. Also, they should dispose location information after achieving the purpose. In our LBS platform architecture, the Location Privacy Portal is responsible for it. The summary of major functionality is as follows: 1) administrator or someone who permits the use of their location information could track the history of location requests. To do this, the accumulated location requests should be maintained for the predefined duration, 2) authentication mechanism using digital signature to the location information should be provided, 3) evidence of providing location information should be maintained for the predefined period. If the time expires, that information should be destroyed automatically. Otherwise, as soon as the location information is used, any related information should be destroyed without being preserved in physical disks. In addition, the subscribers of the Location Privacy Portal could set privacy related profiles and track the historic records of location requests.



Test Configuration				Test Result (LT)		Test Result (LT-SL)
The number of simultaneous requester	Simulator Delay (sec)	The duration of request (min)	The number of request	Response Time (sec)	Success Rate (%)	Response Time (sec)
5	1	10	3978	1.491	100	0.216
5	1	20	8023	1.495	99.9	0.306
5	1	30	12123	1.503	99.9	0.312
10	1	10	7794	1.542	100	0.332
10	1	20	15826	1.516	99.9	0.306
10	1	30	23854	1.510	99.9	0.3
25	1	10	20254	1.484	100	0.204
25	1	20	37606	1.595	99.9	0.315
25	1	30	55612	1.618	99.9	0.338

(b)

Fig. 5. The simulation results of tracking of moving object

Fig. 5 shows the simple simulation result of tracking moving objects using the LBS platform in this paper. As shown in (a), the LT means performance between the LBS platform and testbed and the SL means performance between the gateway simulator and LBS platform. Therefore, the performance of the LBS platform can be represented as *LBSP performance* = *LT* - *SL* and the simulation result is shown in (b).

4 Conclusion

In this paper, we analyze privacy issues on the LBS and introduce the new LBS law for privacy control on location information. We further suggest how to apply the new LBS law efficiently to the existing privacy control mechanisms. Specifically, we propose architecture for an LBS platform to support privacy control during tracking moving objects by conforming to the new LBS law. The suggested LBS platform fulfills three major parts of the law such as subscriber's temporary veto right, subscriber's privacy protection, and subscriber's inspection right. To implement such rules using the LBS platform, this paper proposes four kinds of access control mechanisms using a subscriber's privacy profile, pseudonyms/verinyms mechanisms, and tracking mechanisms. In particular, we have shown practical examples of access control mechanisms based on time and geographical areas: Time Cloaking, Area Cloaking, Time Blurring, and Area Blurring.

The architecture takes advantage of the standardized privacy control entities which could make it possible to achieve interoperability as well as privacy control in the LBS domain. Also, the simulation results could be useful in providing guidelines for adapting privacy control mechanisms on the LBS platform. This proposed approach is expected to be a model for the future privacy control using the LBS platform.

References

1. Jochen Schiller and Agnes Voisard, Location Based Services, 2004, ISBN: 1558609296
2. U.S. Location-based Services Markets-Defining the Enterprise Opportunity, F134-65, 2005 Frost & Sullivan, www.frost.com
3. Ministry of Information and Communication Republic of Korea, http://eng.mic.go.kr
4. Comments of the center for democracy and technology, http://www.cdt.org/privacy/ issues/location/010406fcc.shtml
5. Tae-Hyun Hwang, Seong-Ick Cho, Jong-Hyun Park, and Kyoung-Ho Choi: Object Tracking for a Video Sequence from a Moving Vehicle: A Multi-modal Approach, ETRI Journal, vol.28, no.3, June 2006, pp.367-370
6. Open Mobile Alliance, http://www.openmobilealliance.org/.
7. 3rd Generation Partnership Project, http://www.3gpp.org/.
8. Byung-Ik Ahn, Sung-Bong Yang, Heui-Chae Jin, Jin-Yul Lee: Location Polling Algorithm for Alerting Service Based on Location, W2GIS 2005, pp104-114, 2005.
9. Civilis, A., Jensen, C.S., Nenortaite, J., Pakalnis, S.: Efficient tracking of moving objects with precision guarantees, MOBIQUITOUS 2004, pp164-173, 2004.
10. Civilis, A., Jensen, C.S., Pakalnis, S.: Techniques for efficient road network-based tracking of moving objects, Knowledge and Data Engineering, IEEE Transactions on Volume 17, Issue 5, pp698-712, May 2005.
11. Kam-Yiu Lam, Ulnsoy, O., Lee, T.S.H., Chan, E., Guohui Li: An efficient method for generating location updates for processing of location-dependent continuous queries, Database Systems for Advanced Applications, 2001. Proceedings. Seventh International Conference on 18-21, April 2001.
12. Chartier, E., Hashemi, Z.: Surface surveillance systems using point sensors and segment-based tracking, Digital Avionics Systems, 2001. DASC. The 20th Conference Volume 1, 14-18, Oct. 2001.
13. Kiyoung Moon, Namje Park, Kyoil Chung, Sungwon Sohn and Jaecheol Ryou: Security Frameworks for Open LBS Based on Web Services Security Mechanism, Parallel and Distributed Processing and Applications-ISPA 2005 Workshops, Volume 3759/2005.
14. Minsoo Lee, Jintaek Kim, Sehyun Park, Jaeil Lee and Seoklae Lee: A Secure Web Services for Location Based Services in Wireless Networks, NETWORKING 2004, Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications, Volume 3042/2004.

A New Low-Power and High Speed Viterbi Decoder Architecture

Chang-Jin Choi[1], Sang-Hun Yoon[1], Jong-Wha Chong[1], and Shouyin Lin[2]

[1] Department of Information and Communications, Hanyang University, Seoul, Korea
mibchoi@gmail.com,
{shyoon11,jchong}@hanyang.ac.kr
[2] Department of Electronic and information Engineering, Huazhowg Normal University, Wuhan, china
syliu@phy.ccnu.edu.cn

Abstract. In this paper, we propose a new architecture for low power and high speed viterbi decoder based on register exchange algorithm(RE). In general, the survivor memory unit (SMU) is adopted to RE method for viterbi decoder used in applications that require high speed and low latency. However, the look-ahead trace-back (LATB) method based on the RE method consumes much power due to the frequent switching -activities in register. In this paper, we propose a low power and high speed viterbi decoder that minimizes switching activities of the register used in LATB method to reduce power consumption of viterbi decoder. Because the trace bit of survivor path has a characteristic that the bit value converges into one of 0 or 1, we didn't restore the trace bit to the register of the next stage but to that of the current stage. Simulation results show that the proposed low power and high speed viterbi decoder architecture can reduce switching activities by about 72% in comparison with the conventional LATB architecture using RE method when SNR is 5dB.

Keywords: low-power, viterbi, RE-exchange, look-ahead.

1 Introduction

Generally ubiquitous sensor network terminals need small sized and low power consumption properties. And, in most digital communication system including ubiquitous sensor network, viterbi algorithm having great error-correcting capabilities has been used to achieve low-error-rate data transmission. Generally, It is composed of three units, which are the Branch Metric Unit (BMU), the Add Compare Select Unit (ACSU) and the Survivor Memory Unit (SMU) as in [1][2]. The trace-back (TB) algorithm [3] and the register exchange (RE) algorithm [4] have their own meri1ts and demerits. The implementation of TB algorithm in hard-ware complexity is relatively simple, so that it is applied to many practical application fields. However, the demerit is that the latency is too long (2-3 times longer than the trace-back depth). On the other hand, the RE algorithm has got an ideal latency by using the trace forward scheme, but the demerit is that it has a large resource usages and a high power consumption.

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 283–291, 2007.

Recently, the method of the combination of TB and RE was introduced in the look-ahead trace-back (LATB) algorithm based on the RE algorithm as in [5]. Large reso- urce usages and long latency from the LATB algorithm have been much improved by eliminating the trace-back operation from the TB algorithm, but the LATB algorithm consumes much power due to the frequent switching activities in the register. Accordingly, in this paper, we propose a new algorithm that can reduce power consumption, which is a problem of LATB algorithm to achieve low latency and low power for real-time mobile communication.

2 TB and RE Algorithm

The viterbi decoder is done by accomplishing Add Compare Select (ACS) which chooses a path with the smallest value when comparing the sum of Path Metric (PM) of the previous stage and Branch Metrics (BM) in each branch, and then it is done again by accomplishing survivor memory(SM). The survivor memory unit (SMU) in viterbi decoder can be generally implemented using the trace-back memory (TB) and register exchange (RE) methods. After the series of operation such as BMU and ACSU, the viterbi decoding method uses the maximum likelihood decoding (MLD) algorithm, which is a method to find out a most likely pattern from the received data in [6]. At the last stage of the trellis diagram (see Fig. 1), the TB method extracts the decoded bits, beginning from the state with the minimum PM, state 0.

After tracing back from the last to the first stage, we can get a reverse ordered decoded sequence. This is indicated by the bold line in Fig. 1. In the RE method,

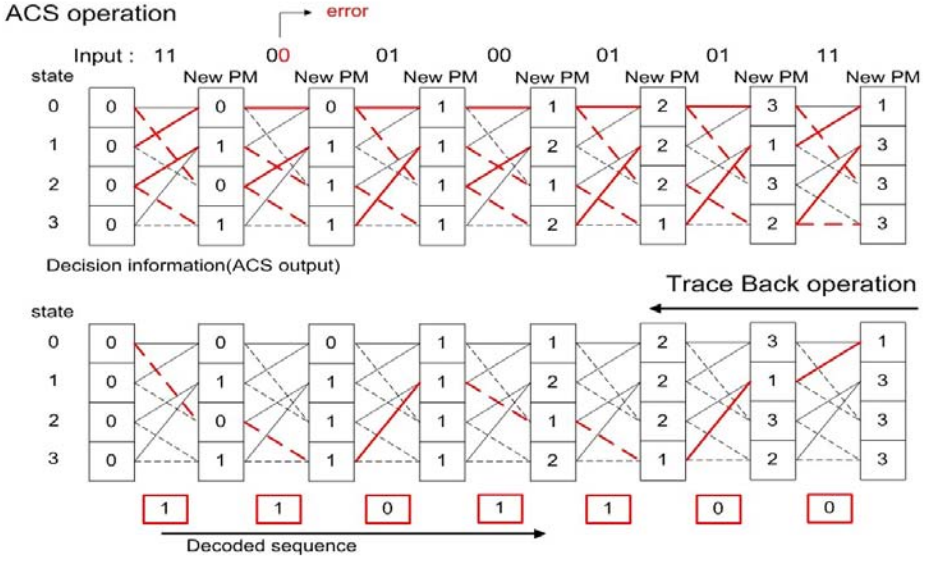


Fig. 1. The operation of the TB algorithm

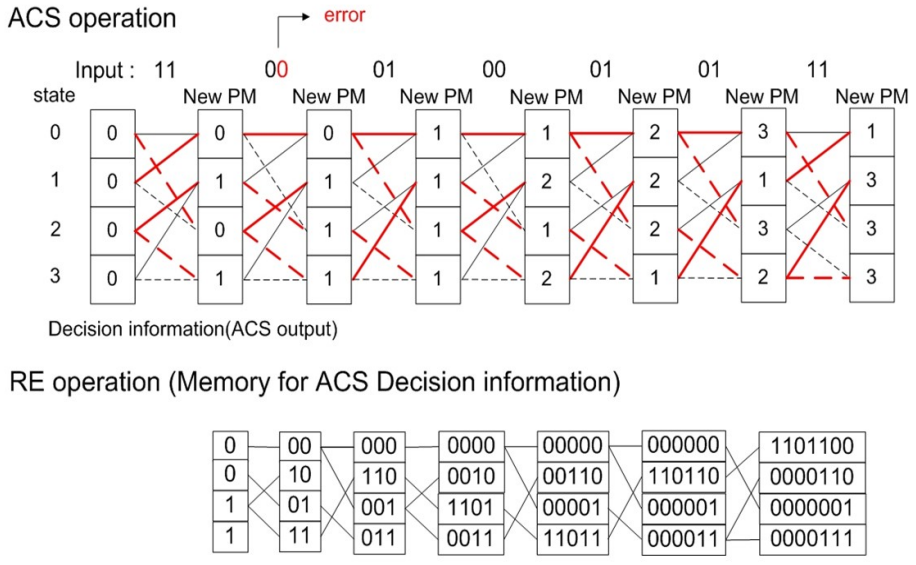


Fig. 2. The operation of the RE algorithm

a register is assigned to each state. The register records the decoded output sequence along the path from the initial state to the final state. This is depicted in Fig. 2. The memory for the decision information is updated and exchanged at every stage.

3 The Look-Ahead Trace-Back Viterbi Decoder

The LATB viterbi decoder based on the RE algorithm doesn't decode data through the trace-back operation after receiving data correspond to the number of trace-back depth like TB algorithm, but it decodes data as soon as receiving the first data by the look-ahead data information. Also LATB algorithm has the same structures as TB algorithm up to ACSU. But it doesn't require trace-back operation because it decodes data as one of the trace bits at the last stage by shifting the trace bits to be decoded to the next stage according to information of survivor path and by storing them at the register just as in the RE algorithm. In other words the LATB scheme without the trace-back operation stores the trace bits and the state addresses of the first trace-forward stage, and their positions are rearranged in the following trace-forward stages. As soon as the trace-forward operation finishes at a survival path length (T) and the survivor path is determined, the trace bit and the state that the survivor path is pointing become the desired decoded bit and the starting state. Fig. 3 shows how the LATB scheme works. A simple trellis diagram is shown in Fig. 3 (a). The trace-forward depth is 5, and the survivor path is drawn with a thick shaded line. Fig. 3(b) shows the change of the contents of the trace memory. The decoding process

is similar to that of the register exchange method [5] except that intermediate values are thrown away, and only the values of the first stage are stored since only the starting state and the corresponding decoded bit are desired. As illustrated in Fig. 3(b), the trace bit and the state address in memory with the minimum PM (path metric) are determined to be '1/10' after the trace-forward, which means the decoded bit is '1' and the starting state is '10'. Only 4 memory units are enough to decode while the register exchange method and the trace-back method need 4 x 4=16 memory units.

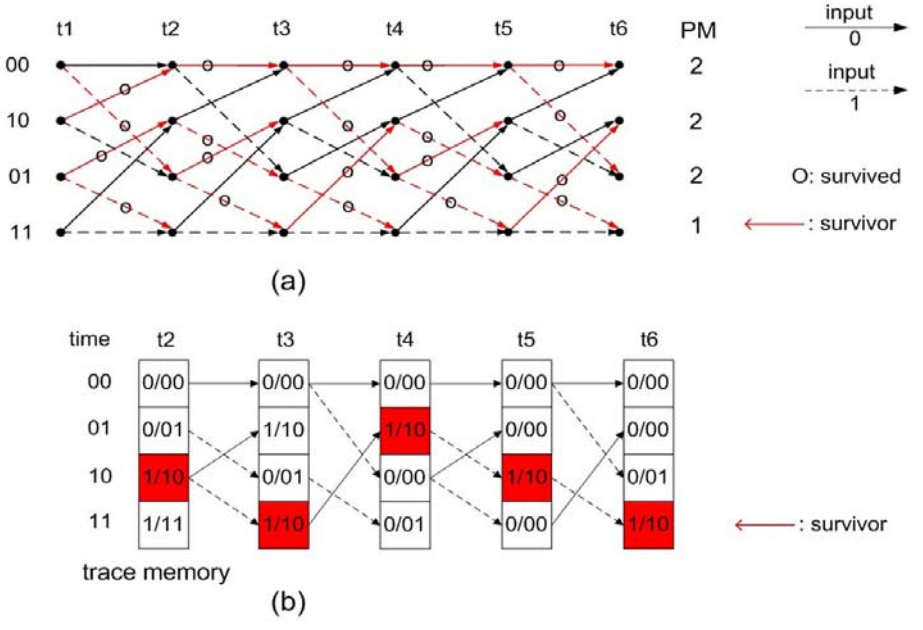


Fig. 3. (a) A trellis diagram (b) The change of memory contents at each trace-forward stage

In view of the trellis diagram, trace bits are a survivor path (0 or 1), which can enter each state in the next stage while performing trace-forward operation from the current stage to the next stage. In other words, the initial value of stage 1 (0 0 1 1) is shifted to the next stage when data is received on the LATB method. Because LATB algorithm decodes data using the trace bits in the smallest PM state at the last stage without trace-back operation, it considerably reduces latency when compared with the conventional TB method.

4 The Low-Power and High Speed Viterbi Decoder

Fig.4 shows the stored data in TB memory while decoding data with convolutional encoder whose code rate R=1/2 and constraint length K=3 in LATB

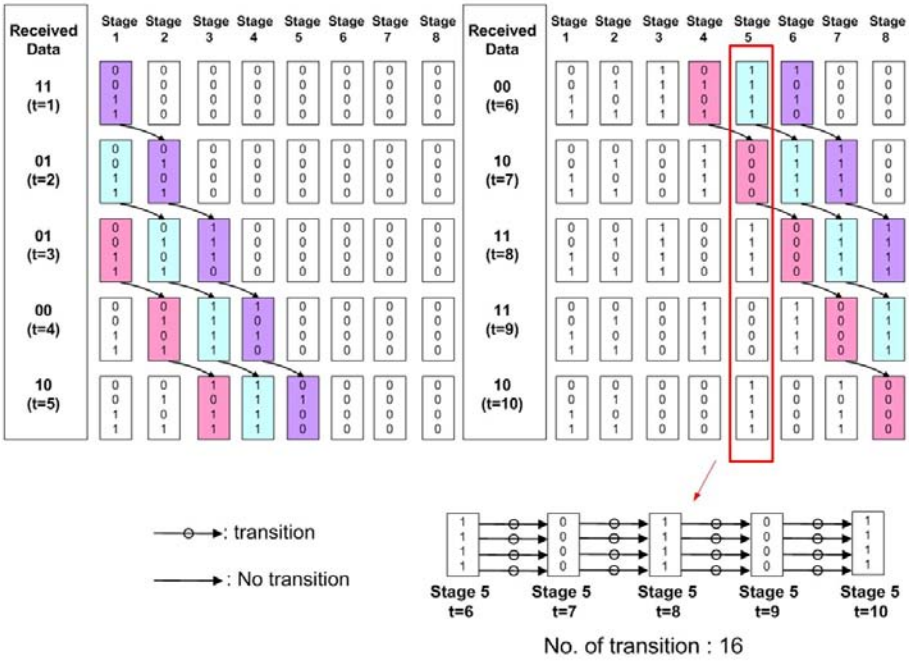


Fig. 4. The stored data in TB memory of LATB decoder

method. In other words, It shows trace bits recorded at the register when the data is received in order of 11, 01, 01, 00, 10,00,10,11,11,10. When the first data 11 is received, the first stage register is initialized to (0 0 1 1).

When the next data 01 is received, the trace bits for decoding the first data 11 (bits 0 0 1 1) are shifted and recorded at the second stage register according to the information of survivor path. At this moment, the first stage register is reset to the initial value(0011). When the third data 01 is received, the trace bits (recorded at the second stage) for decoding the first data 11 are shifted and recorded at the register on the third stage according to the information of survivor path, then the initial value(0011) for decoding the second date received, 01 is shifted and recorded at the second stage register. In the same manner, the first stage register is reset to the initial value by the third data received. Likewise, whenever data is received, trace bits recorded at each register are shifted and recorded at the register on the next stage according to information of survivor path. We need 15 registers to store trace bits to decode data received since the trace bits are to be shifted fifteen times; that is, five times that of constraint length. For convenience, Fig.4 just shows eight stages instead of fifteen. Fig.5 indicates trace bits, which are shifted and recorded in each stage by the information of survivor path for decoding the first and the third receiving data; that is, data11(t=1) and data01(t=3) in Fig.4. Looking at the trace bits stored in each register, we can see them gradually converge into a value (0 or 1). This

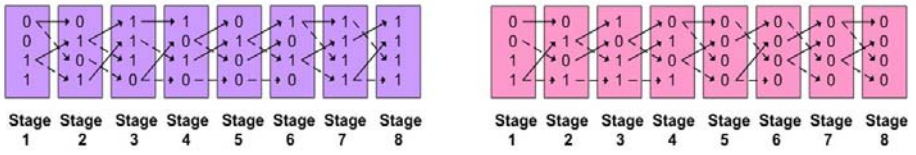


Fig. 5. Condition of trace bits recorded to each stage for decoding one data

is because trace bits tend to converge into the data to be decoded as they are shifted according to the information of survivor path.

Thus, referring to the stage 5 of boxed area in Fig.4, we can see that if the data to be decoded are turned from 1 to 0 then the trace bits recorded at each register would turn from 1 to 0. Thus, more power will be consumed as all trace bits in the registers have to be transitioned,and then too many switching-activities occur.

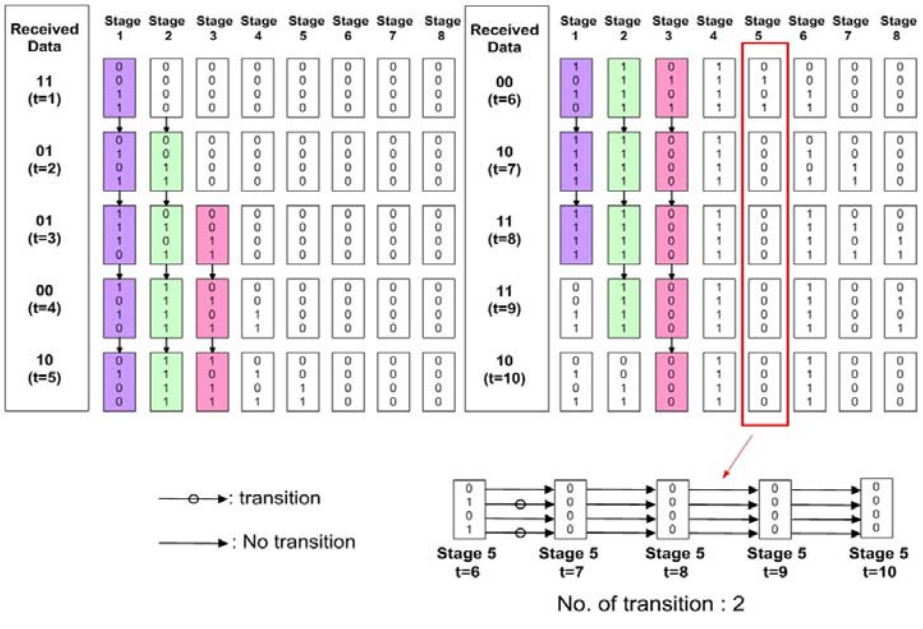


Fig. 6. The stored data in TB memory of the low-power viterbi decoder using low-power and high speed viterbi decoder architecture

The low-power SMU architecture proposed in this paper, therefore, renews and stores trace bits at one Register as shown in Fig.6, instead of shifting and storing them at the register of next stage according to the information of survivor path whenever data is received. In this way, as shown in stage 5 of boxed area in Fig.6, there would be little changes in trace bits as the trace bits will converge

into the values to be decoded, approaching the stages to the trace-back depth. Thus, we can reduce power consumption as transitions of trace bits and switching activities decrease.

5 Result of Simulation

The proposed viterbi decoder reduces power consumption in SMU by decreasing switching activity of register, renewing trace bit according to survivor path information and using its feature that trace bits converge into 0 or 1.

Table 1. The specifications of IEEE.802.11a

Parameter	Value
TB Length	35
General Polynomial	$(133_8, 171_8)$
Code Rate	1/2
Constraint Length	7
Modulation	BPSK
No.Data	10^5

Table 2. The comparison of switching activity of the proposed architecture and the conventional LATB architecture

SNR (dB)	Numbers of switching activity		Ratio of decrease(%)
	LATB architecture	Proposed architecture	
1(dB)	95887650	81152456	15.34
2(dB)	95894749	76586042	20.14
3(dB)	95861283	73237949	23.60
4(dB)	95891248	70953342	26.01
5(dB)	95867375	69496071	27.51

The simulations results were performed according to the specifications of IEEE.802.11a in table 1.We found that the switching activities of the proposed architecture shown in table 2 were decreased by about 72% at 5dB SNR when compared with the conventional LATB architecture.The bigger the SNR is, the larger the ratio of decreased switching activity is, because the bigger the SNR is, the faster the trace bit will converge into the decoding value.

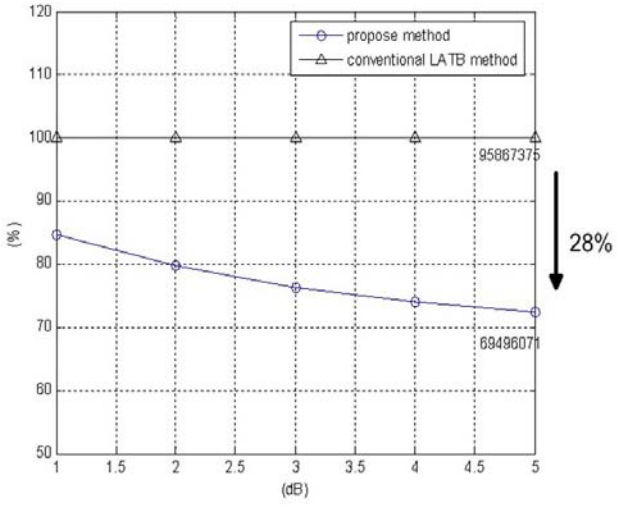


Fig. 7. The number of switching activity of the proposed method and the conventional LATB method

6 Conclusion

In this paper, we proposed a new low-power viterbi decoder architecture. Simulation results showed that the switching activities of the proposed architecture were reduced by 85 to 72 % in comparison with the conventional LATB architecture. This means that we can reduce power consumption, a demerit of LATB method with a short latency and a high-speed communication capability. Therefore, the proposed viterbi decoder can be applied to the IEEE 802.15.4a which is spotlight as a ubiquitous sensor network, because it can make H/W with its low power consumption feature possible. This paper compared the reduction of switching activity of the register of the proposed method with that of the conventional LATB method when the trace-back depth is five times larger than the value of the constraint length, but we can expect that the bigger the trace-back depth is, the larger the ratio of reduction of switching activity will be.

Acknowledgments. " This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment)" , "(IITA-2006-C1090-0603-0019)"

References

1. Viterbi, A. J.: Convolutional Codes and Their Performance in Communication Systems. IEEE Transactions on Communication .Vol.COM-19 No.5. (1971)
2. Forney, G. D., JR.: The Viterbi Algorithm. Proceeding IEEE. Vol. 61(3) .(1973)

3. Wicker, S. B. *Error Control Systems for Digital Communication and Storage.* Englewood Cliffs, NJ: Prentice Hall, 1995.
4. Steinert, M., Marsili, S.: Power Consumption Optimization for Low Latency Viterbi Decoder. International Symposium on Circuits and Systems '04. Vol.2. (2004) 377-380
5. Baek, J. G., Yoon, S.H., Chong, J. w.: Memory Efficient Pipelined Viterbi Decoder with Look-ahead Trace-back. International Conference on Electronics, Circuits and Systems'01. Vol.2. (2001)
6. Forney, G. D.: Convolutional Codes. Maximum-likelihood Decoding, 25(3). Information and Control. (1974)

Dynamic EPG Implementation for Ubiquitous Environment

In Jung Park[1], Duck Je Park[1], and Cheonshik Kim[2]

[1] Dept. of Electronic Eng. Dankook University in Korea
digitallab@kornet.net,fly21c@airport.co.kr
[2] Digital Media Engineering, Anyang University
mipsan@anyang.ac.kr

Abstract. In this paper, we proposed EPG as a suitable method for digital broadcasting in a ubiquitous environment. As an application technique for user custom UI service in a Ubiquitous Environment, a dynamic EPG presentation technique was studied. To accomplish this, a context structure definition and presentation engine to present it was needed. The UI context information consists of various graphic component data, such as position, font, size, shape, etc., used to draw EPG GUI on a screen using UIML language. Context information was defined and the technique for presenting on a screen was studied. Also, the digital broadcasting field was applied for demonstration of Dynamic EPG and finally the paper proved that this can be used as a part of core technique for the ubiquitous dynamic EPG service. Consequently, it was a good idea to change the UI framework configuration based on context information made by someone and send it to a rendering engine for displaying user custom UI.

Keywords: EPG, digital broadcasting, UI.

1 Introduction

The "Ubiquitous" concept was first proposed by Dr. Mark Weiser as "Ubiquitous Computing." Since 1999, it has been expanded to the idea of a "Ubiquitous Network" in Japan. The ubiquitous environment implies that computer chips are embedded in everything. In other words, "Ubiquitous Computing" refers to an IT environment where computers are connected to the network anytime, anywhere. In the long run, Ubiquitous Network will naturally develop into Ubiquitous Computing. This concept has been drawing attention recently because a consensus has formed as a result of the development of silicon technologies and the success of mobile device technologies. The development of digital media based on these technologies is creating a very individualized media environment, moving away from the old community-based structure. For example, hot spots allow us to access the network anywhere, and people use their mobile phones to enjoy TV or download music to their MP3 players from the Internet. This study researched Dynamic UI technology, which extends the mobility of devices in a ubiquitous environment with a focus on the mobile UI (User Interface). Mobility here means

F. Stajano et al. (Eds.): ICUCT 2006, LNCS 4412, pp. 292–300, 2007.

that users have their own UI and can use it anytime, anywhere. Two technologies are required to achieve this: context technology and presentation technology. Context technology relates to the context information structure to be used for mobility, and ultimately it is information storage technology to enable us to use our own UI information anytime, anywhere. This study developed Ubiquitous UI technology based on these two technologies in order to enable custom EPG (Electronic Program Guide) services for users in a mobile environment.

2 Electronic Program Guide

The EPG allows TV viewers to obtain program information from broadcasting stations or other sources, preview, search, filter, sort and select programs. These are the most basic services of EPG, and most EPG service providers offer these services in personalized forms [2][3]. Broadcasting network EPG refers to the display of broadcasting program schedules and related information on TV screens, and in a broad sense, a guide to all products and service offerings by the service provider. EPG can be classified according to structural and functional aspects into the following:

- Grid EPG: The broadcasting stations, programs, and airing times are listed.
- Mosaic EPG: Multiple channels appear simultaneously in one screen.
- Ticker EPG: A part of the screen is used to provide information on broadcasting programs without interfering with the current viewing of a program.
- Mini EPG: Information on all programs is displayed in a part of the screen without interfering with the current viewing of a program.

3 Proposed Design Methods of EPG

3.1 Proposed Concept of Dynamic EPG

Dynamic EPG means EPG in a mobile environment. It also refers to the provision of personalized media services in a ubiquitous environment and personalized EPG services related to digital broadcasting. Dynamic EPG allows users to freely control EPG layout on the screen. In other words, users can change the position to display time, program, and summaries on the screen, as well as altering font size and colors.

Fig. 1 shows a user, who is using EPG services by connecting to MediaServer #1, and then moving to another place and connecting to MediaServer #2 where he can use his own EPG services for viewing broadcasting programs.

3.2 Context Information for Presentation

To present dynamic EPG on the screen, three types of context information are used: EPG context, User context, and UI context. Fig. 2 shows the overall flow of the dynamic EPG output through these 3 types of context information. The

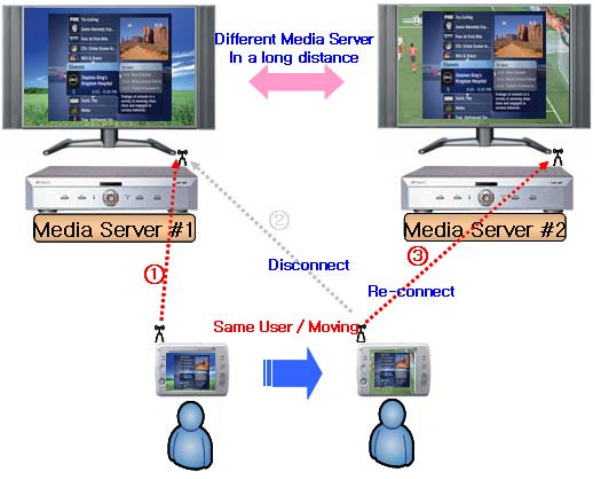


Fig. 1. Same users are moving to another location

context information in this study was limited to digital broadcasting. Therefore, the context information for dynamic EPG was expressed in XML language, which is appropriate for our research. Dynamic EPG uses three types of context for presentation of information. The first is EPG context, which converts the EIT information of the transmission stream from the broadcasting network into XML format. The second is user context, which stores information about the user's favorite TV programs and device-related information. The third is the UI context which contains the values for various presentations, including the positions and properties of objects in dynamic EPG and event links for activation of external events. This can create a variety of screen layouts preferred by the user with scripts composed in a predefined format. It is similar to wallpaper on a PC, and allows users to build their own GUI environment for broadcast receiver using scripts.

3.3 Screen Composition by Rendering Engine

The rendering engine (Fig. 3) composes EPG screens using the three types of context information, i.e., EPG context, user context and UI context. Fig.2 shows that three layers are required to compose a full screen. These are the event layer, presentation layer, and background layer. These three layers are projected and combined to compose the actual screen viewed by the user. Each layer is described in more detail below. First, the Event Layer accepts external inputs through input devices such as a mouse, keyboard, or remote controller, and uses them to control the dynamic EPG. Inputs are connected to event handling in the actual graphic API, and the GUI on the screen changes according to input values. For input devices to generate events, a mouse or remote controller can be used. Second, the Presentation Layer uses UI context information to

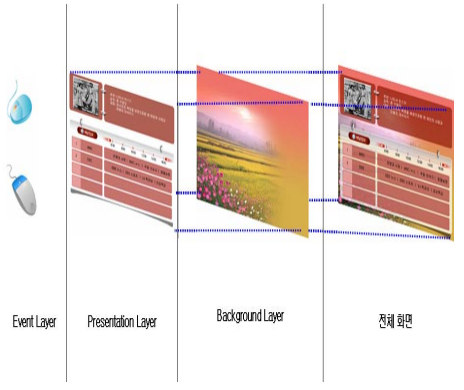


Fig. 2. A compunction screen composes a rendering engine

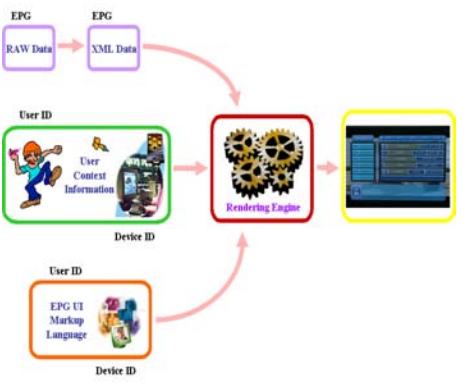


Fig. 3. Three EPG output structure charts which used different context information

display the EPG context on the screen. The rendering engine interprets object information expressed in markup language through DOM objects, links them with graphic components, and then changes the properties of each component to display GUI on the screen. Various types of information such as the position of each component and the presentation values can be obtained from these DOM objects. Third, the Background Layer is like the canvas of the EPG screen, and users can create their unique EPG feeling with images, or display primary color images without a background image. This presentation on the background is carried out by the Presentation Layer. The final image that is viewed by user shows the full screen to which each layer is projected.

4 Proposed User Interfaces

This chapter introduces a method to extract EPG information transmitted through digital broadcasting streams, context information structure, and a method of analyzing and handling this information using the rendering engine.

4.1 Context Information Technology

The dynamic EPG described in this study uses the EIT information extraction method that is employed in terrestrial and cable broadcasting networks. In terrestrial and cable broadcasting networks, PSIP information, which contains various pieces of broadcasting-related information such as time and channels, is sent through MPEG-2-based transmission streams [5]. Among many tables, the most important ones are STT, MGT, VCT, and EIT. They contain the information required to extract and present EPG data. STT (System Time Table) sends the current system time, and MGT (Master Guide Table) provides the version, size, PID, etc., of the other tables except for STT. VCT (Virtual Channel Table) contains the properties of all virtual channels in the current transmission stream,

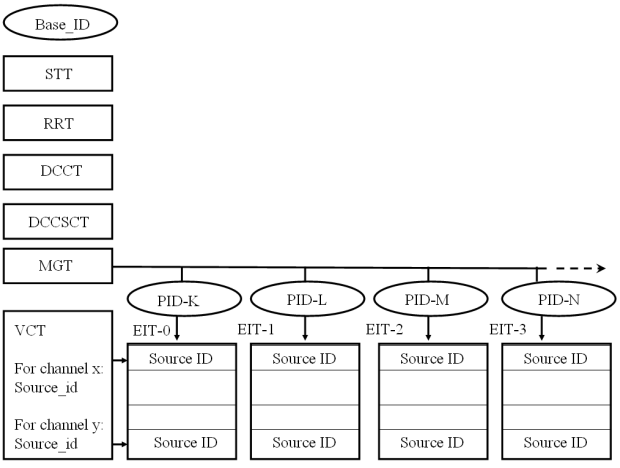


Fig. 4. A PSIP table structure

and EIT (Event Information Table) provides information on the events of the virtual channels in time sequence [6][7].

Fig. 4 shows the process of analyzing EIT in a stream and extracting the desired image. First, STT is extracted, and the RRT, DCCT, MGT, and VCT tables are analyzed in this order. Next, EIT information is extracted according to the PID value. These values can be saved in the memory and used for other purposes. The actual tables that can be used to obtain data for EPG are PMT, VCT and EIT; the other tables play the role of guides in extracting these tables.

User context information can be divided into four types, which are shown in Fig. 5. These are User ID, Device ID, User System Information, and Device Information. User ID is used to check whether the user context information transmitted via the network is that of an authenticated user. Device ID indicates the category of the other media terminal when mobility is added to the dynamic EPG. Device information consists of basic control information for the device so that a media terminal can be set up with the user's custom setting values when the user accesses the terminal by applying dynamic EPG to it. For example, people with normal hearing and vision will use normal volume, fonts, and colors. However, people with abnormal hearing and vision will use volume, font size, or colors adapted to them. Device information is used to set the default controls of the other terminal through parameters based on specific user information.

Fig. 6 shows the basic structure of broadcast EPG. Broadcast EPG typically has a few common basic structures, which are described below. EPGDate shows the current broadcasting time, EPGTimerInterval shows the time shift when a program is selected, EPGStation shows information on broadcasting stations and channels, EPGProgram shows broadcasting programs by hour, EPGDescription shows a simple review of each broadcasting program, and EPGBack decorates these elements to make them look good on the screen. This structure is not

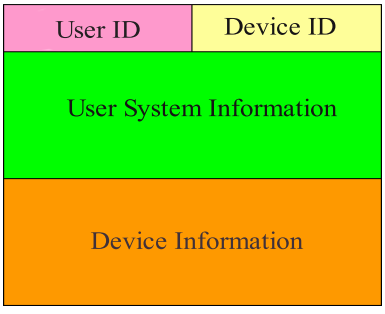


Fig. 5. A user context information formation table

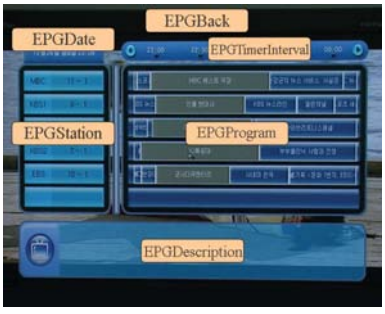


Fig. 6. A broadcast EPG basic structure

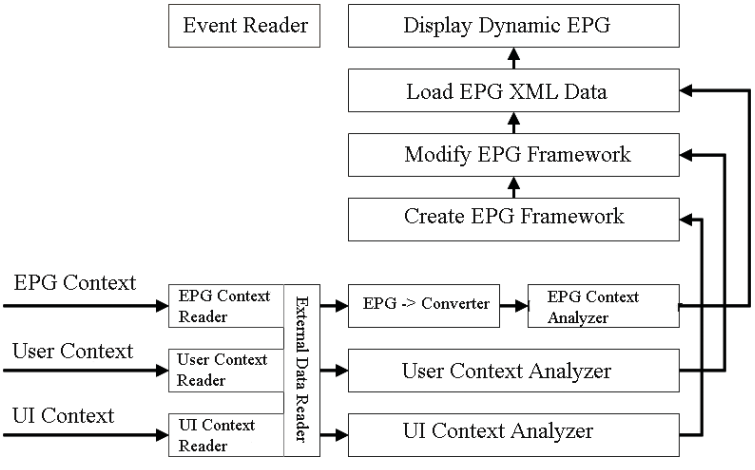


Fig. 7. Dynamic EPG rendering engine block diagram

essential for the total EPG, but rather a common structure. These terms are not used in the actual broadcasting EPG, but defined to create the EPG markup language in this study.

Fig. 7 shows the rendering engine structure, which is a core technology for dynamic EPG.

5 Implementation of Proposed Method

5.1 Method of Experiment

Experiments for this study were performed using three types of networks and different context information for each. In particular, TIP-30 model USN, wireless

LAN for IEEE 802.11 b/g, and front-end for receipt of terrestrial/cable digital broadcasting were used. Fig. 8 shows the experiment diagram. As the dynamic EPG is displayed on the screen after being processed in a media server, only the media server structure and experiments will be described here. The broadcasting network signals are input to the front end of the media server, and the event information is extracted using the PID values set by user. For PID values, the values defined in [6] and [7] are used. Information obtained from this process is hexadecimal EPG raw data, which needs to be transformed to XML format so that it can be used by the rendering engine. The middle converter converts raw data to XML format.

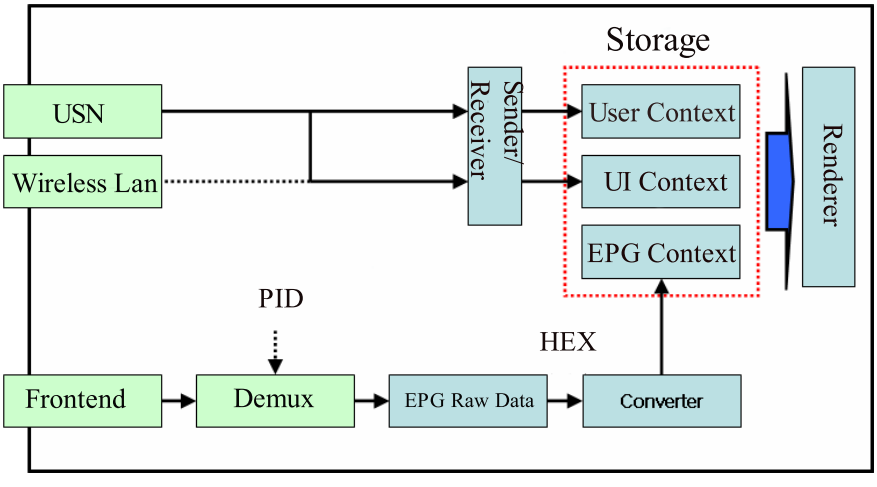


Fig. 8. Architecture of system for the parts

The test with the TIP 30 model, with values set by the experiment, revealed that stable transmission was possible for 10 Kbytes or lower. Therefore, UI information of 10 Kbytes or more is transmitted via the wireless LAN, and information smaller than that is transmitted through the USN. Further, user identification, device identification, and user custom information for the EPG are transmitted through the USN.

5.2 Experiment Results

The experiments obtained outputs from the rendering engine in the media server using the three types of context information: user context, UI context, and EPG context information, which is the core technology. Fig. 9 shows a general output screen using EPG context information without any changes to the screen. This

Fig. 9. A general EPG output screen

Fig. 10. A change of font size from 12 to 18

is an EPG of a general structure that can be seen in typical digital broadcasting, and used as a reference screen for comparison with other output screens.

6 Conclusions

In this paper, we designed a dynamic EPG service structure, defined a few types of context information that can be provided for it, and implemented EPG screens using this information to demonstrate the validity of the definitions. In particular, this paper defined three types of context information for dynamic EPG service: EPG context information for programs, user context information, and UI context information. To provide dynamic EPG, this paper conducted research to identify a rendering engine that can effectively present it. The rendering engine interprets UI context information in order to interpret and present the screens desired by the user. In conclusion, as the dynamic EPG technology proposed by this paper is not based on hardware technology, but composed of software components, it can be restructured to provide ubiquitous services, applied to various devices by developing extended rendering technology, and used to provide personalized UI services appropriate for the ubiquitous environment.

References

1. Jae-yun kim, Ubiquitous Computing: Business model and view, Samsung economy research institute, 2003
2. Basic, R. and Mocinic, M., "User's requirements for electronic program guide (EPG) in interactive television (iTV)", VIPromcom-2002, pp16-19, June 2002.
3. Bobbie, W., "Interactive electronic programmer guides," IEE Half-day Colloquium on navigation in Entertainment Services, Jan., 1998.
4. Product Solution, www.itmg.co.kr
5. Siryong Yu etc, MPEG systems, Dae-young, 1997.

6. D. R. Tarrant, "An open european standard for an electronic programme guide", International Broadcasting Convention, pp441-446, Sept., 1997.
7. ATSC Standard: Program and System Information Protocol for Terrestrial Broadcast and Cable (Revision B), Mar., 2003.
8. ATSC Recommended Practice: Program and System Information Protocol Implementation Guidelines for Broadcasters, June,2002.

Author Index