Tai-hoon Kim Hojjat Adeli
Wai-chi Fang Thanos Vasilakos
Adrian Stoica Charalampos Z. Patrikakis
Gansen Zhao Javier García Villalba
Yang Xiao (Eds.)

Communications in Computer and Information Science 265

Communication and Networking

International Conference, FGCN 2011
Held as Part of the Future Generation
Information Technology Conference, FGIT 2011
in Conjunction with GDC 2011
Jeju Island, Korea, December 2011, Proceedings, Part I

Part 1

Springer

Communications
in Computer and Information Science 265

Tai-hoon Kim Hojjat Adeli Wai-chi Fang
Thanos Vasilakos Adrian Stoica
Charalampos Z. Patrikakis Gansen Zhao
Javier García Villalba Yang Xiao (Eds.)

Communication and Networking

International Conference, FGCN 2011
Held as Part of the Future Generation Information Technology Conference, FGIT 2011
in Conjunction with GDC 2011
Jeju Island, Korea, December 8-10, 2011
Proceedings, Part I

Springer

Volume Editors

Tai-hoon Kim
Hannam University, Daejeon, Korea
E-mail: taihoonn@hannam.ac.kr

Hojjat Adeli
The Ohio State University, Columbus, OH, USA
E-mail: adeli.1@osu.edu

Wai-chi Fang
National Chiao Tung University, Hsinchu, Taiwan, R.O.C.
E-mail: wfang@mail.nctu.edu.tw

Thanos Vasilakos
University of Western Macedonia, Kozani, Greece
E-mail: vasilako@ath.forthnet.gr

Adrian Stoica
Jet Propulsion Laboratory, Pasadena, CA, USA
E-mail: adrian.stoica@jpl.nasa.gov

Charalampos Z. Patrikakis
National Technical University of Athens, Greece
E-mail: bpatr@telecom.ntua.gr

Gansen Zhao
Sun Yat-sen University, Guangzhou, China
E-mail: zhaogansen@gmail.com

Javier García Villalba
Universidad Complutense de Madrid, Spain
E-mail: javiergv@fdi.ucm.es

Yang Xiao
University of Alabama, Tuscaloosa, AL, USA
E-mail: yangxiao@cs.ua.edu

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-27191-5 e-ISBN 978-3-642-27192-2
DOI 10.1007/978-3-642-27192-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): C.2, H.4, I.2, H.3, D.2, F.1

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Future generation communication and networking is an area that attract many professionals from academia and industry for research and development. The goal of the FGCN conference is to bring together researchers from academia and industry as well as practitioners to share ideas, problems and solutions relating to the multifaceted aspects of future-generation communication and networking.

We would like to express our gratitude to all of the authors of submitted papers and to all attendees for their contributions and participation.

We acknowledge the great effort of all the Chairs and the members of Advisory Boards and Program Committees of the above-listed event. Special thanks go to SERSC (Science and Engineering Research Support Society) for supporting this conference.

We are grateful in particular to the speakers who kindly accepted our invitation and, in this way, helped to meet the objectives of the conference.

December 2011 Chairs of FGCN 2011

Preface

We would like to welcome you to the proceedings of the 2011 International Conference on Future Generation Communication and Networking (FGCN 2011) — one of the partnering events of the Third International Mega-Conference on Future-Generation Information Technology (FGIT 2011) held during December 8–10, 2011, at Jeju Grand Hotel, Jeju Island, Korea.

FGCN 2011 focused on various aspects of advances in future-generation communication and networking. It provided a chance for academic and industry professionals to discuss recent progress in the related areas. We expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject.

We would like to acknowledge the great effort of the FGCN 2011 Chairs, International Advisory Board, Committees, Special Session Organizers, as well as all the organizations and individuals who supported the idea of publishing this volume of proceedings, including the SERSC and Springer.

We are grateful to the following keynote, plenary and tutorial speakers who kindly accepted our invitation: Hsiao-Hwa Chen (National Cheng Kung University, Taiwan), Hamid R. Arabnia (University of Georgia, USA), Sabah Mohammed (Lakehead University, Canada), Ruay-Shiung Chang (National Dong Hwa University, Taiwan), Lei Li (Hosei University, Japan), Tadashi Dohi (Hiroshima University, Japan), Carlos Ramos (Polytechnic of Porto, Portugal), Marcin Szczuka (The University of Warsaw, Poland), Gerald Schaefer (Loughborough University, UK), Jinan Fiaidhi (Lakehead University, Canada) and Peter L. Stanchev (Kettering University, USA), Shusaku Tsumoto (Shimane University, Japan), Jemal H. Abawajy (Deakin University, Australia).

Last but not the least, we give special thanks to Ronnie D. Caytiles and Yvette E. Gelogo of the graduate school of Hannam University, who contributed to the editing process of this volume with great passion.

We would like to express our gratitude to all of the authors and reviewers of submitted papers and to all attendees, for their contributions and participation, and for believing in the need to continue this undertaking in the future.

December 2011

Tai-hoon Kim
Hojjat Adeli
Wai-chi Fang
Thanos Vasilakos
Adrian Stoica
Charalampos Z. Patrikakis
Gansen Zhao
Javier García Villalba
Yang Xiao

Organization

Honorary Chair

Dae-sik Ko	Mokwon University, Korea

General Co-chairs

Wai-chi Fang	National Chiao Tung University, Taiwan
Thanos Vasilakos	University of Western Macedonia, Greece
Adrian Stoica	NASA JPL, USA

Program Co-chairs

Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Gansen Zhao	Sun Yat-sen University, China
Javier García Villalba	Universidad Complutense of Madrid, Spain
Tai-hoon Kim	GVSA and University of Tasmania, Australia
Yang Xiao	University of Alabama, USA

Workshop Chair

Byungjoo Park	Hannam University, Korea

Publicity Co-chairs

Houcine Hassan	Polytechnic University of Valencia, Spain
Damien Sauveron	University of Limoges, France
Qun Jin	Waseda University, Japan
Irfan Awan	University of Bradford, UK
Muhammad Khurram Khan	King Saud University, Saudi Arabia
Yang Xiao	The University of Alabama, USA
J.H. Abawajy	Deakin University, Australia

Publication Chair

Maria Lee	Shih Chien University, Taiwan

International Advisory Board

Hsiao-Hwa Chen	National Sun Yat-Sen University, Taiwan
Gansen Zhao	Sun Yat-sen University, China
Han-Chieh Chao	National Ilan University, Taiwan
Hamid R. Arabnia	The University of Georgia, USA
Gongzhu Hu	Central Michigan University, USA
Byeong-Ho Kang	University of Tasmania, Australia
Aboul Ella Hassanien	Cairo University, Egypt
Tughrul Arslan	The University of Edinburgh, UK
Jianhua Ma	Hosei University, Japan
Sankar K. Pal	Indian Statistical Institute, India
Xiaofeng Song	Nanjing University of Aeronautics and Astronautics, China
Frode Eika Sandnes	Oslo University College, Norway

Program Committee

Aboul Ella Hassanien	Hsin-Hung Chou	N. Jaisankar
Ai-Chun Pang	Hui Chen	Ning Gui
Aggeliki Sgora	Huirong Fu	Omar Soluiman
Albert Banchs	J. Vigo-Aguiar	P.R. Parthasarathy
Andres Iglesias Prieto	Janusz Szczepanski	Ricky Yu-Kwong Kwok
Andrzej Jajszczyk	Jiann-Liang	Robert Goutte
Antonio Lagana	Jieh-Shan George Yeh	Rui L. Aguiar
Benahmed Khelifa	Jiming Chen	Shun-Ren Yang
Bogdan Ghita	Juha Ro"ning	Soon Ae Chun
Chao-Tung Yang	Kin Keung Lai	Stephen Huang
Chia-Chen Lin	Kwok-Yan Lam	Sun-Yuan Hsieh
Christophe Fouqueré	Li Shijian	Tae (Tom) Oh
Chu-Hsing Lin	Luis Javier García Villalba	Terence D. Todd
Clement Leung	Marc Lacoste	Victor C.M. Leung
Damien Sauveron	Matthias Reuter	Viktor Yarmolenko
Dimitrios D. Vergados	Michel-Marie Deza	Vincent Oria
Don-Lin Yang	Ming-Yen Lin Feng	Vincenzo De Florio
Driss Mammass	Mohammad Moghal	Weili Han
Farrukh A. Khan	Nashwa El-Bendary	Witold Pedrycz
Gianluigi Ferrari	Neveen I. Ghalii	Yung-Hui Li Feng
Hong Sun	Nikolaos Pantazis	Yvette E. Gelogo
Hsiang-Cheh Huang		Ronnie D. Caytiles

Special Session Organizers

Hong Kook Kim
Tae-Young Byun
Y. Byun

Table of Contents – Part I

Table of Contents – Part II

Wireless Multimedia Sensor Networks Testbeds and State-of-the-Art Hardware: A Survey

Muhammad Omer Farooq and Thomas Kunz

Department of Systems and Computer Engineering
Carleton University, Canada
omer.farooq@ymail.com, tkunz@sce.carleton.ca

Abstract. In recent years, Wireless Multimedia Sensor Networks (WMSNs) attracted researchers' attention. This has led to an advancement made at both levels: hardware design and open source libraries to capture and process sound and image on sensor motes. WMSNs have the potential for real time object detection and recognition. It can be used to locate misplaced items, assisted living, helping people with cognitive impairments, species detection, and provide real time images for many other surveillance-based application. The WMSN research domain is relatively new, therefore many people fail to fully recognize the potential of WMSNs. In this paper, we provide a detail description of various state-of-the-art WMSNs testbeds. We categorize these testbeds in two classes: hardware-based testbeds and software-based testbeds. Furthermore, we list state-of-the-art in wireless multimedia sensor motes hardware. The objective of this paper is threefold: unleash the potential application areas of WMSNs, attract new researchers to the WMSNs domain, and provide a guideline for proper hardware selection keeping in mind the type of research that needs to be carried out.

Keywords: Wireless Multimedia Sensor Network (WMSN), Wireless Sensor Network (WSN), Testbed, Multimedia Sensor Mote.

1 Introduction

WMSNs are composed of audio/video sensing devices that are capable of retrieving and transmitting multimedia content such as audio, video and still images. Furthermore, such sensing devices can capture temperature, humidity, light intensity, and many other environment-related readings. The concept of WMSNs can best be understood as the convergence between wireless sensor networks and distributed smart cameras. WMSNs is a multidisciplinary research area and the advancements made in hardware design, signal processing techniques, coding theory, wireless networking, and control theory have made WMSNs a reality.

Ordinary sensor nodes such as MICAz and TelosB are equipped with insufficient capabilities to process multimedia data. Executing Digital Signal Processing (DSP), coding, and compression algorithms on these motes can result in short-lived WMSNs. Furthermore, due to the computational overhead, the effectiveness of the deployed

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 1–14, 2011.

system may be reduced. The stated reasons prompted manufactures and researchers to provide solutions like Cyclops [1], Stargate [2] and Imote2 [2]. These hardware boards provide separate processors to handle multimedia data, moreover they can be interfaced with sensor motes, cameras, and a Personal Computer (PC).

WMSNs need to provide the following functionalities: (a) timely delivery of real time data, (b) differentiated services support, (c) reliability, (d) bulk storage capacity especially at the sink nodes, (e) implementation of state-of-the-art compression algorithms on sensor motes, (f) separate microcontrollers on sensor motes for DSP related processing, and (g) support for a multimedia database at sink nodes.

We have organized this paper as follows. In Section 2, we present a detailed description of different hardware and software-based WMSNs testbeds. State-of-the-art hardware for WMSNs is analyzed in Section 3 and finally, we conclude this paper in Section 4.

2 Wireless Multimedia Sensor Networks Testbeds

Invariably, researchers use mathematical modeling and simulations to evaluate the performance of different networking protocols for WMSNs. Sometimes these evaluation methods are misleading due to a lack of a detailed and accurate communication model. Therefore, to accurately analyze the system behavior and convince people that WMSNs are real, a number of WMSN testbeds have been developed. In this section, we elaborate on existent hardware-based and software-based WMSN testbeds.

2.1 Hardware-Based Testbeds

Hardware-based testbeds deploy hardware devices such as sensing motes and integrate them with other devices e.g., cameras and microphones. Hardware-based testbeds for WMSNs use ordinary sensor nodes, sophisticated sensing platforms, and medium resolution and high resolution cameras. Sophisticated sensing platforms such as Stargate and Imote provide powerful processing and relatively good storage capacities. Some hardware-based testbeds provide libraries to carry out image processing related functions. In the remaining section, we describe various hardware- based testbeds for WMSNs in more detail.

Georgia Institute of Technology BWN-Lab Testbed. The BWN-Lab testbed at Georgia Tech [4] consists of commercial off-the-shelf advanced devices and the purpose is to demonstrate the efficiency of algorithms and protocols for multimedia communication through wireless sensor networks. The testbed comprises of heterogeneous sensors nodes such as iMote and MICAz. The testbed also includes three different types of multimedia sensors: low end imaging sensors, medium quality webcam-based multimedia sensors, and a pan-tilt camera mounted on a mobile robot. The low end cameras are integrated with MICAz motes, the medium end video sensors are interfaced with the Stargate platform. The high end pan-tilt camera is installed on an Acroname GARCIA robotic platform. Each sensor runs a TinyOS-based protocol stack that allows to create a multihop data path to the sink. Sinks are

Stargate boards that receive sensor information through MICAz motes. The mobile sink is able to track the position of any MICAz node deployed in the testbed. If any such node reports abnormal readings, the mobile sink moves to the location of the mote, captures and transmits a video stream.

Characterizing Energy Consumption in a Visual Sensor Network Testbed. The main focus of the work reported in [5] is to characterize the energy consumption of a visual sensor network testbed. Energy consumption depends on the target application of the WMSN. The energy consumption is influenced by the processing, flash memory access, image acquisition, and communication over the network. This work further characterizes energy consumption depending upon different hardware states the system switches through as it executes benchmark applications used in this testbed. The benchmark application uses idle, processing intensive, storage intensive, communication intensive, and visual sensing processes. The testbed consists of eight visual sensor nodes and one sink node. A Dell Inspiron 4000 laptop with PIII CPU, 512 MB memory, and 20 GB hard disk is used as the sink. It runs the Linux (Kernel 2.4.20) and uses an IEEE 802.11b wireless card for communication. A visual sensor node is composed of a Stargate board with an XScale PXA255 CPU (400 MHz), has 32MB flash memory and 64MB SDRAM, and provides PCMCIA and compact Flash connectors on the main board. The Quickcam webcam is used with resolution of up to 640 x 480 pixels.

Explorebots: A Mobile Network Experimentation Testbed. Work presented in [6] demonstrates expandable, vision and sensor-equipped wireless robots called Explorebots. The work is based on MICA motes. The testbed supports experimental analysis of protocols for mobile multi-hop networks. The contribution of the work are: localization and tracking algorithms for WSNs and routing algorithms for hybrid i.e., cellular ad-hoc networks. The major features of Explorebots are: powerful on board microprocessor for accurate robotic movement control, interface with MICA wireless sensor nodes to provide sensing and communication facility, build in electronic compass and ranging device that can be used for navigation, and an on-board camera with wireless transmitter.

A Lightweight Camera Sensor Network Operating on Symbolic Information. The testbed described in [7] primarily focuses on camera sensor networks for behavior recognition. It is based on biologically inspired address event image sensors (the main functionality of such sensors is to pick the useful information from the scene to reduce the bandwidth requirement. Outputs are supplied in an address event scheme that preserves privacy). The application of the resulting system in assisted living is also explained. This testbed makes use of overhead Hitachi KP-D20A imaging devices that operate at a 768 x 494 pixel resolution. It uses Stargate boards along with the MICA2 motes. Additional features include mobile robots and remote reconfiguration.

IrisNet: A Low Cost Embedded Color Vision System. IrisNet [8] is a heterogeneous WMSN platform. This platform allows the user to perform queries over the video sensors distributed worldwide. The hardware components used in this testbed are: Logitech Quick Cam Pro 3000 cameras working at a resolution of 640 x 480 pixels along with the high end Stargate platform. This video sensor network can be queried worldwide, therefore the system maintains a log of events. Once they are collected, IrisNet provides two tools to process and replay the logs for visualization: a trace generator is used for sorting the logged events in ascending order and a trace animator displays the topology.

SenseEye: A Multi-tier Camera Sensor Network. SenseEye [9] is a multi-tier heterogeneous camera-based surveillance network. It makes use of the low fidelity camera sensors at the second tier to view an object. To view an object at comparatively high fidelity, cameras at tier three perform object tracking. The hardware used in this testbed consists of Cycops, CMUCam, and pan-tilt-zoom webcams. The image resolution is set to 352 x 288 pixels. Stargate and MICA2 platforms are used along with mobile robots.

Wireless Line Sensor Network for Distributed Visual Surveillance. The testbed described in [10] is developed for object detection using an efficient distributed visual surveillance system. In WSNs with scarce resources, abundant bandwidth is required for multimedia transmission, moreover the image processing algorithms can consume a significant amount of energy on the sensor motes. Therefore, this testbed implementation uses a line sensor architecture capable of capturing a continuous stream of temporal one-dimensional images. The image processing algorithms for one-dimensional data are much faster, require less memory, and can work in limited-bandwidth environment. This testbed includes hardware and software modules, description of these modules are given below.

This testbed uses the CMOS camera module HV7131GP. This camera module is capable of capturing images at a maximum resolution of 640 x 480 pixels. The camera module contains image processing capabilities such as Gamma correlation, color interpolation and auto white balance. It provides an adjustable frame rate of up to 30 frames per second. The sensor uses an I^2C communication channel to send and receive commands and transfer bit map images to the end devices. The main setup uses the Flex board [8] with the camera attached to a breakout board and a serial to TTL converter.

Hierarchical Character Oriented Wildlife Species Reorganization through Heterogeneous Wireless Sensor Networks. [11] describes the deployment of a WSN based video camera system in Henry Doorly Zoo, Omaha, USA for animal species detection. Animal species recognition is done through heterogeneous wireless sensor motes. The contribution of this work is integration of a WSN with in-network distributive image recognition functionality. The authors have proposed new HIerarchical Scalarized Character Oriented DEtection (HIS-CODE) algorithm and architecture. The HIS-CODE algorithm is implemented on each sensor mote deployed in the testbed. This work leads to the realization of a fully automated camera array system that takes advantage of distributive processing and image recognition within a tiered heterogeneous WSN.

TinyEARS: Spying on House Appliances with Audio Sensor Nodes. Tiny Energy Accounting and Reporting System (TinyEARS) [12] is a system capable of generating device-level power consumption reports. Power reports are being generated based upon the acoustic signature of household devices. This testbed makes use of the iMote2 sensor mote which is integrated with a microphone to capture the sound generated by household appliances. Sensor motes are required to perform signal processing operations e.g., Fast Fourier Transform, feature extraction, and classification. The results obtained through this audio sensor network show that the system is capable of reporting power consumption of household appliances with an error margin of 10%.

Design and Implementation of Dual Camera Wireless Sensor Network for Object Retrieval. [13] presents the design and implementation of a dual camera sensor network. The purpose of the work is to develop a memory assistance tool for assisted living. Its authors have focused on low power camera sensors that can be distributed in a home environment to facilitate retrieval of misplaced objects. The rational presented for this work is that small battery powered cameras are portable, easy to deploy, and can be densely deployed for better coverage. The main contributions of this work are: energy efficient region of object estimation, energy efficient algorithms for object recognition, and incorporation of user feedback. The two cameras in the dual camera system are placed close to each other so that they both share the same Field of View (FoV). User can query the system through a PC or a PDA that is connected to the IEEE 802.15.4 wireless radio.

The operation of the overall system is divided into three main components: object detection, object recognition, and object retrieval. Object detection is done by tier 1 cameras, object recognition is done by tier 2 cameras, on the request of tier 1 camera. Finally, the detection of an object is communicated to the BS using the IEEE 802.15.4 radio. Tier 1 comprises of Cyclops camera connected to a MICAz mote. Tier 2 comprises of an Enalab camera and an iMote2.

Design and Implementation of a Sensor Based Wireless Camera System for Continuous Monitoring in Assistive Environments. The basic idea in the testbed described in [14] is to develop a camera-based surveillance system to monitor someone who may have physical or cognitive impairments. The developed testbed has been deployed in home and office environments to check the system performance under different interference scenarios. The proposed system is called Sensor Integrated Camera Surveillance (SICS). The basic idea is to equip the subject under observation with wearable wireless sensors and deploy the wireless routers in the observed environment. Routers form the mesh network and as the subject moves, automatic handover takes place between the router that was previously serving the subject and the router that is about to start servicing the subject. Cameras are deployed with the wireless routers. The wireless router sends the images to the sink by traversing the wireless backbone. SICS uses an on board image compression algorithm to reduce the bandwidth consumption. SICS allows a human operator to only watch a single screen to track a subject as she moves around. Wireless sensors being worn by the subject provide the location information that helps in selecting the camera. The main contributions of this work are:

a) The first measurement study of a wireless camera network deployed both in residential and office environments.
b) The demonstration of wireless sensor-based localization for a surveillance system.
c) Quantitative performance evaluation of automatic camera hand-off through empirical implementation of the SICS architecture.

The following hardware equipments are being used in the home-based testbed. Each wireless node is an ASUS WL-500g Premium wireless router, with 266 MHz CPU, 8 MB flash, 32 MB RAM, and one Broadcom 4318 IEEE 802.11b/g radio. The ASUS operating system is being replaced with OpenWrt Kamikaze 7.09. The server is an IBM ThinkPad T42 laptop, which has a PCMCIA IEEE 802.11a/b/g card bus adaptor with Atheros AR5213 chipset and Madwifi 0.9.4 wireless driver.

The following hardware equipment are being used in the office-based testbed. Each wireless node is a RouterBOARD 532A device, which has a MIPS 400 MHz CPU, 64 MB RAM, and 2 GB Compact-Flash disk. Each node can have one or two Mini PC wireless radio cards, for which Wistron Neweb CM9 with Atheros AR5213A chipset is used. All wireless nodes use omni-directional antennas that have 3 and 5 dbi gains at the 2.4 and 5 GHz frequencies respectively.

Distributed Search in Sensor Networks. The work reported in [15] focuses on the design and implementation of a distributed search system over a camera sensor network, where each node is a search engine that can sense, store and search information. Furthermore, this research also focuses on local storage, local search, and distributed search. These features are designed to be efficient under a resource constraint WSN. A testbed implementation is provided on a number of iMote2 sensor nodes equipped with low power camera and extended flash storage. The main motivation for this work was: instead of deploying an application-specific camera sensor networks, there is a need for a general purpose image search paradigm that can be used to recognize a variety of objects, including new types of objects, that can be detected. The search engine implementation is made possible by the use of compact image representation called "visterms" for efficient communication and search, and the re-design of fundamental data structures for efficient flash-based storage search. In short, key contributions of this work are: efficient local storage and search, distributed image search, and application case studies.

From an implementation point of view the system has three major modules: Image Processing Module (IPM), Query Processing Module (QPM), and Communication Processing Module (CPM). The IPM captures images using the camera periodically, reads the captured image and saves it into the flash memory. The QPM module processes ad-hoc queries and continuous queries from the proxy. CPM handles the communication to other sensors and the proxy.

The developed testbed consists of six iMote2 sensor nodes and a PC acting as a central proxy. Each iMote2 is equipped with Enalab camera and 1 GB SD card, runs arm-Linux and communicates with other motes using the TOSMAC driver, which is compatible with the TinyOS wireless protocols.

Fusion of Audio and Image Information for Efficient Object Detection and Capture. [16] presents a demonstration of an integrated audio-image WMSN for object detection and capture. It has been shown that based on the multi-modal information, it is possible to effectively and efficiently localize objects using relatively cheap audio and camera sensors. The system used for demonstration consists of audio nodes, camera nodes, a base station for centralized control, and a personal computer for real time observations. Each node present in the WMSN uses a wireless mote and DSP for communication and signal processing, respectively. The mote used is the CSIRO developed Fleck™ 3B. The DSP used is an Analog Devices 600 MHz ADSP-BF537 Blackfin processor. The BlueTechnix BF537E board is used which has a CPU with 132 KB SRAM, 32 MB external SDRAM and 4 MB flash. The camera daughterboard contains the sensor chip i.e., an OV7648 Color CMOS from OmniVision Technologies. The audio nodes use an omni-directional, high sensitive electret capsule microphone (a type of condenser microphone which eliminates the need for a polarizing power supply using permanent charged material) for sound capture, and an Analog Devices AD73311 codec IC for analog to digital conversion.

Audio nodes send sound event detection information to the base node. The Python application on the base node performs arbitration required to determine the camera that is closest to the sound source, and then coordinates the camera nodes to effectively localize and capture an image of the object of interest. This demonstration consists of three pairs of co-located audio and camera nodes. Each node is connected to a PC with Ethernet back link.

Demonstration of Image Compression in a Low-Bandwidth Wireless Camera Network. A demonstration of a low bandwidth, wireless camera network is given in [17]. In this demonstration, image compression is performed at each node. Furthermore, this demonstration introduces the Fleck™ hardware platform that has been developed by the authors to run image compression algorithms on each node. This demo shows real time image data coming back to the BS from the deployed WMSN.

2.2 Software-Based Testbeds

Software-based testbeds provide Application Programming Interface (API) libraries and software driver interfaces to facilitate rapid application development and deployment of WMSNs. Software libraries provide an abstraction layer to the researchers and developers. It hides the low-level details of the devices, hence facilitates the rapid evaluation of protocols for WMSNs. In this section, we summarize the software-based testbeds for WMSNs.

WiSNAP: An Image Sensor Network Application Platform. WiSNAP[18] is a MATLAB™ based software testbed for WMSNs. This testbed provides high level APIs that can be used by the developers to evaluate applications and protocols for

WMSNs. These APIs hide the low-level device details and provide users with a powerful framework to interact with sensor motes. Moreover, it provides excellent APIs for image processing. WiSNAP includes device libraries for Agilent's ADCM-1670 camera module, Agilent's ADNS-3060 optical mouse sensor, and Chipcon's CC2420DB IEEE 802.15.4 radio. Since it is an open source architecture, it can be extended for other hardware types. WiSNAP provides two set of APIs: image sensor API and wireless mote API. The image sensor API provides support for capturing images from the sensor mote and other image processing related functionalities. The wireless mote API provides support for initializing a mote, and transmitting and receiving a MAC level packet. In [18], the authors have given two examples for using the provided APIs. One example explains the event detection mechanism and the other demonstrates a node localization mechanism.

Address Event Imagers for Sensor Networks: Evaluation and Modeling. [19] describes a software testbed based on Address Event Representation (AER) of an image. AER is used with the address event sensors. The main goal for developing address event sensors it to preserve energy and bandwidth. Therefore, address event sensor are intelligent enough to extract the required features from the image and then represent them in symbolic form called address event representation. Finally, these symbols are transmitted to the sink. This results in less bandwidth consumption and privacy. The testbed provides an implementation of an emulator of AER imagers that has been written in Visual C++ and runs under the Windows operating system. The purpose of the AER emulator is to take an 8 bit grayscale input stream from a commercial of-the-shelf USB camera and to output a queue of events in a text file.

3 State-of-the-Art Wireless Multimedia Sensing Motes

In this section, we present state-of-the-art multimedia sensing motes. We examine processing capabilities, available memory, multimedia support, and radio technology supported by the contemporary wireless multimedia sensing motes.

3.1 Stargate

Stargate [2] is being manufactured by MEMSIC. It comes with an Intel PXA-255 Xscale processor that operates at 400 MHz. It has 64 MB of SDRAM and 32 MB of flash storage. Its processor can execute advanced image processing algorithms. It supports the embedded LINUX (kernel 2.4.19) operating system. It supports two radios i.e., the IEEE 802.11 and the IEEE 802.15.4. Furthermore, it can be interfaced with MICAZ motes, Ethernet, Serial, JTAG, USB, PCMCIA, and Compact Flash Connector. A Stargate is 3.5 x 2.5 inches in size. The functionalities supported by Stargate make it an ideal choice to act as a base station, especially when it provides the IEEE 802.11 radio that can be used to create a wireless backbone.

Table 1. Comparison of Hardware-based Testbeds

WMSN Testbed	Service Differentiation	Bulk Storage Capacity	Compression Algorithms	DSP Processor	Multimedia Database
BWN-Lab Testbed [4]	NO	YES	NO	YES	NO
Energy Consumption in Visual Sensor Networks [5]	NO	YES	NO	Powerful Single Processor	NO
Explorebots [6]	NO	YES	NO	Powerful Single Processor	NO
Lightweight Camera Network [7]	NO	NO	YES	NO	NO
IrisNet [8]	NO	YES	NO	YES	YES
SenseEye [9]	NO	YES	YES	YES	NO
Wireless Line Sensor Network [10]	NO	YES	YES	YES	NO
Wild Life Species Recognition[11]	NO	YES	YES	Powerful Single Processor	NO

Table 1. (*continued*)

TinyEARS [12]	NO	YES	NO	Powerful Single Processor	NO
Dual Camera WSN for Object Retrieval [13]	NO	YES	YES	YES	NO
Sensor based Wireless Camera for Assistive Environments [14]	NO	YES	NO	Powerful Single Processor	NO
Image Search in Sensor Network [15]	NO	YES	YES	Powerful Single Processor	YES
Fusion of Audio Video Image Information [16]	NO	YES	NO	YES	NO
Image Compression in Low Bandwidth Wireless Camera [17]	NO	YES	YES	YES	NO

3.2 Imote2

Imote2 [2] is developed by the Intel Corporation and it is being manufactured by MEMSIC. It is an advanced platform especially designed for sensor network applications requiring high CPU/DSP and wireless link performance and reliability.

The Imote2 contains an Intel XScale processor, PXA271. It comes with an IEEE 802.15.4 radio (TI CC2420) with an onboard antenna. It has two basic sensor board interfaces, consisting of two connectors. Moreover, it has an advanced sensor board interface, consisting of two high density connectors on the other side of the board. The Imote2 is a modular stackable platform and can be stacked with sensor boards to

customize the system to a specific application. Imote2 provides an additional wireless MMX co-processor. It has 256 KB SRAM, 32 MB Flash, and 32 MB SDRAM. It has an integrated 2.4 GHz antenna. It provides basic and advanced expansion connectors supporting: 3xUART, I2C, 2xSPI, SDIO, I2S, AC97, USB host, Camera I/F, GPIO. To provide direct connection to a PC, a mini USB port is also available. The size of an Imote2 is 48 mm x 36 mm.

3.3 CMUCam3

The CMUCam3 [3] is an ARM7TDMI based fully programmable embedded computer sensor. It is equipped with a Philips LPC2106 processor that is connected to an Omnivision CMOS camera sensor module. Features of CMUCam3 include: CIF resolution (352 x 288 pixels) RGB color sensor, open source development environment for Windows and Linux, MMC flash slot with FAT16 driver support, loads image into memory at 26 frames/sec. It contains LUA (a lightweight scripting programming language) that allows for rapid prototyping, software-based JPEG compression, basic image manipulation library, CMUcam3 image emulation, compatible connector with wireless motes (Tmote Sky, FireFly, 802.15.4), and FIFO image buffer for multiple pass hi-resolution image processing. It supports 64 KB RAM and 128 KB of ROM.

3.4 MeshEye

MeshEye [20] is a hybrid resolution smart camera mote for applications in distributed intelligence surveillance. MeshEye has a unique vision system i.e., a low resolution stereo vision system continuously determines position, range, and size of moving objects entering in its field of view. This information triggers a color camera module to acquire a high resolution image of the object. It contains an ARM7TDMIARM thumb processor. It is a 32 bit RISC architecture that can operate at up to 47.92 MHz. MeshEye has 64 KB of SRAM and 256 KB of flash. It uses a CC2420 transceiver and provides an implementation of the IEEE 802.15.4 standard. It uses the Agilent 2700 VGA camera module.

3.5 WiCa

WiCa [21] is a smart camera mote with high performance vision system. It uses two VGA camera modules (640 x 480 24 bit color), which feed video to IC3D, which is a dedicated parallel processor, running at 80 MHz. WiCa uses an 8051 based Atmel AT89C51 host processor, which runs at 24 MHz, through a 128 KB dual port RAM. For large scale image processing its supports 10 Mbits of RAM. It provides an implementation of the IEEE 802.15.4 standard.

3.6 Cyclops

Cyclops [1] is a small camera device developed for WMSNs. Cyclops can be interfaced with MICA2 and MICAz motes. The Cyclops hardware architecture consists of an Agilent ADCM-1700 CMOS camera module, a Xilinx FPGA and an 8 bit RISC ATMega128 microcontroller.

3.7 FireFly Mosaic

In [22], a vision-enabled image processing framework for sensor motes is presented. It uses FireFly motes coupled with CMUCam3 sensor cameras. The main design objective of FireFly Mosaic is to design a low cost, energy efficient, and scalable camera mote compared to the centralized wireless webcam-based solutions. FireFly Mosaic runs Nano-RK, a real time operating system for WSNs and it uses the networking protocol stack provided by Nano-RK. FireFly mosaic is an integration of FireFly motes with CMUCam3, therefore its hardware architecture is similar to CMUCam3. FireFly motes, CMUCam3, and Nano-RK are developed at Carnegie Mellon University.

3.8 Fleck™

The Fleck hardware platform [17] is designed for low-bandwidth wireless camera networks where image compression is undertaken at each node. Fleck is a robust hardware platform for outdoor use. The current version of Fleck™-3 consists of an Atmegal 128 microcontroller running at 8 MHz. It is equipped with a Nordic NRF905 radio transceiver with a bit rate of 76.8 Kbits/sec. Another promising feature of Fleck is that it is equipped with solar charging circuitry. It provides an additional DSP board which is a Texas Instruments 32 bit 150 MHz DSP processor (TMS320F2812). It has 1 MB of SRAM and an Ominvision 640 x 480 pixel color CCD sensor.

3.9 WiSN

WiSN [23] is a multimedia sensing platform developed at Stanford University. The microcontroller is a 32 bit ARM7TDMI based on the Atmega AT91SAM7S running at 48 MHz. It contains 256 KB flash and 64 KB RAM. Its camera is based on Agilent ADCM 1670 and it supports multiple resolutions. An IEEE 802.15.4 radio is used.

4 Conclusion

In this paper, we have surveyed various hardware and software-based testbeds for WMSNs. Furthermore, we have surveyed the features of the state-of-the-art multimedia sensing nodes. Material presented in this paper suggests that state-of-the- art wireless multimedia sensor nodes have enough capabilities to facilitate deploying WMSNs in different real world phenomenon. Moreover, emerging sensing technologies such as AER based sensing motes and state-of-the-art in coding theory have made it possible to deploy WMSNs in low-bandwidth networks. Therefore, enough support from the hardware side is available to deploy WMSNs in a multitude of scenarios. From the networking protocol perspective, little work has been done to support multimedia applications in low-bandwidth wireless personal area networks. There is a need to employ Quality of Service (QoS) mechanisms to effectively and efficiently utilize the available hardware. Furthermore, to support multiple applications one a single mote there is a need for real time operating system for WSNs. Nano-RK is an example of one such

operating system. Definitely more research is required to support hard and soft real-time applications in WMSNs.

References

1. Rahimi, M., Baer, R., Iroezi, O.I., Garcia, J.C., Warrior, J., Estrin, D., Srivastav, M.: Cyclops: in situ Image Sensing and Interpretation in Wireless Sensor Networks. In: The Proc. of 3rd International Conference on Embedded Network Sensor Systems, SenSys., San Diego, CA, USA, pp. 192–204 (November 2005)
2. MEMSIC, http://www.memsic.com
3. Rowe, A., Goode, A., Dhiraj, G., Illah, N.: CMUcam3: An Open Programmable Embedded Vision Sensor. Carnegie Mellon Robotics Institute Technical Report, RI-TR-07, May 13 (2007)
4. Akyildiz, I.F., Melodia, T., Chowdhary, K.R.: Wireless Multimedia Sensor Networks: Applications and Testbeds. Proc. of IEEE 96(10), 1588–1605 (2008)
5. Margi, C.B., Petkov, V., Obraczka, K., Manduchi, R.: Characterizing energy consumption in visual sensor network testbed. In: Proc. IEEE/CreateNet Int. Conf. Testbeds Res. Infrastructure Development Network Communication (TridentCom), Barcelona, Spain (March 2006)
6. Dahlberg, T.A., Nasipuri, A., Taylor, C.: Explorebots: A mobile network experimentation testbeds. In: The Proc. ACM SIGOPS Oper. Syst. Rev., vol. 35(2), pp. 61–77 (2001)
7. Teixerira, T., Lymberopoulos, D., Culurciello, E., Aloimonos, Y., Savvides, A.: A lightweight camera sensor network operating on symbolic information. In: The Proc. ACM Workshop on Distributed Smart Cameras, Boulder, CO, pp. 76–81 (2006)
8. Rowe, A., Rosenberg, C., Nourbakhsh, I.: A low cost embedded color vision system. In: The Proc. IEEE/RSJ International Conference on Intelligent Robotics System (IROS), Lausanne, Switzerland (October 2002)
9. Kulkarni, P., Ganesan, D., Shenoy, P., Lu, Q.: SenseEye: A multi-tier camera sensor network. In: The Proc. ACM Multimedia, Singapore (November 2005)
10. Chitnis, M., Liang, Y., Zheng, J.Y., Pagano, P., Lipari, G.: Wireless Line Sensor Network for Distributed Visual Surveillance. In: The Proc. of 6th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks, Tenerife, Canary Islands, Spain, October 29-30 (2009)
11. Duran, D., Peng, D., Sharif, H., Chen, B., Armstrong, D.: Hierarchical Character Oriented Wildlife Species Recognition through Heterogeneous Wireless Sensor Networks. In: The Proc. of 18th IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication, Athens, Greece, September 3-7, pp. 1–5 (2007)
12. Taysi, Z.C., Guvensan, M.A., Melodia, T.: TinyEARS: Spying on House Appliances with Audio Sensor Nodes. In: The Proc. of 2nd ACM Workshop on Embedded Sensing Systems for Energy Efficiency in Building, Zurich, Switzerland, November 2, pp. 31–36 (2010)
13. Xie, D., Yan, T., Ganesan, D., Hanson, A.: Design and Implementation of a Dual-Camera Wireless Sensor Network for Objective Retrieval. In: The Proc. of 7th International Conference on Information Processing in Sensor Networks (IPSN), St-Louis, Missouri, USA, pp. 469–480
14. Li, N., Yan, B., Chen, G., Govindaswamy, P., Wang, J.: Design and implementation of a sensor based wireless camera system for continuous monitoring in assistive environments. Personal and Ubiquitous Computing Journal 14(6) (September 2010)

15. Yan, T., Ganesan, D., Manmatha, R.: Distributed Image Search in Sensor Networks. In: The Proc. of 6th ACM Conference on Embedded Network Sensor System (SenSys 2008), Raleigh, NC, USA, November 5-7, pp. 155–168 (2008)
16. O'Rourke, D., Moore, D., Wark, T.: Demo Abstract: Fusion of Audio and Image Information for Efficient Object Detection and Capture. In: The Proc. of International Conference on Information Processing in Sensor Networks (IPSN), San Francisco, California, USA, April 13-16, pp. 401–402 (2009)
17. Karlsson, J., Wark, T., Valencia, P., Ung, M., Corke, P.: Demonstration of Image Compression in Low-Bandwidth Wireless Camera Network. In: The Proc. of International Conference on Information Processing in Sensor Networks (IPSN), Cambridge, Massachusetts, USA, April 25-27 (2007)
18. Hengstler, S., Aghanjan, H.: WiSNAP: a Wireless image Sensor Network Application Platform. In: The Proc. of 2nd International Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities, TRIDENTCOM 2006, Barcelona, Spain, March 1-3, pp. 6–12 (2006)
19. Teixeira, T., Culurciello, E., Park, J., Lymberopoulos, D., Barton-Sweeney, A., Savvides, A.: Address Event Imagers for Sensor Networks: Evaluation and Modeling. In: The Proc. of 5th International Conference on Information Processing in Sensor Networks, IPSN 2006, Nashville, TN, USA, pp. 458–466 (April 2006)
20. Hengstler, S., Prashanth, D., Fong, S., Aghajan, H.: MeshEye: A Hybrid-Resolution Smart Camera Mote for Application in Distributed Intelligent Surveillance. In: The Proc. of 6th International Symposium on Information Processing in Sensor Networks, IPSN 2007, Cambridge, MA, USA, pp. 360–369 (2007)
21. Kleihorst, R., Abbo, A., Schueler, B., Danilin, A.: Camera mote with High-Performance Parallel Processor for Real-Time Frame-based Video Processing. In: The Proc. of IEEE Conference on Advanced Video and Signal based Surveillance, AVSS, Queen Marry, University of London, London, UK, September 5-7, pp. 69–74 (2007)
22. Rowe, A., Goel, D., Rajkumar, R.: FireFly Mosaic: A Vision Enabled Wireless Sensor Networking System. In: The Proc. of 28th International Real Time System Symposium, RTSS 2007, Tucson AZ, USA, pp. 459–468 (December 2007)
23. Downes, I., Rad, L.B., Aghajan, H.: Development of the Mote for Wireless Image Sensor Network. In: The Proc. of Cognitive Systems and Interactive Sensors, COGIS, Paris, France (2006)

An Energy-Efficient Cluster-Based Routing in Wireless Sensor Networks

Seongsoo Cho[1], Bhanu Shrestha[1], Keuk-Hwan La[1], Bonghwa Hong[2], and Jongsup Lee[3]

[1] Department of Electronic Engineering, Kwangwoon University, 26 Kwangwoon–gil, Nowon-gu, Seoul, 139-701, Korea
css@kw.ac.kr, bnu@kw.ac.kr, Khra@kw.ac.kr
[2] Department of Information Communication, Kyunghee Cyber University, Dongdaemun-gu, Seoul, 130-701, Korea
bhhong@khcu.ac.kr
[3] SK C&C,SK u-Tow4er, 25-1, Jeongia-dong, Bundang-gu, Seongnam-si, 463-844, Korea
jsleearmy@sk.com

Abstract. In Wireless Sensor Networks (WSNs), sensor nodes depend on batteries for energy source. The ability to use limited energy efficiently is the key to determining the lifetime of networks and the amount of information transmitted. Low Energy Adaptive Clustering Hierarchy (LEACH) is a representative cluster-based routing protocol designed to ensure energy use efficiency. In this paper, a protocol scheme was proposed wherein member nodes (lower-level nodes) are designed to compare the currently sensed data with the previously sensed one and to switch to sleep mode when a match is achieved. The design is to help improve the transmission energy efficiency. The proposed scheme was tested via simulations and was compared with two existing cluster-based algorithms, i.e., LEACH and Threshold Sensitive Energy Efficient Sensor Network Protocol (TEEN). Performance evaluation was conducted based on the number of surviving nodes in each of the three networks (i.e., LEACH, TEEN, and proposed scheme) over time. The results indicated that the scheme contributed to greater energy efficiency by helping to increase the lifetime of the LEACH network by a maximum of 27%.

Keywords: WSN, LEACH, TEEN, Clustering Algorithm, Energy Efficiency.

1 Introduction

A Wireless Sensor Network (WSN) consists of sensor nodes, each of which includes a microcontroller, a radio transceiver, and a sensing module. Data collected by the sensor nodes are sent to the sink node (the data aggregator) mostly through the multi-hop wireless mesh network. The WSN, deploying a large number of sensor nodes in a specific area, had first started out as the monitoring and patrol application as well as military application for missions involving areas to which humans have limited access. Since then, the network has expanded its application gradually, which now includes environmental monitoring, building hazard diagnosis, and patient monitoring

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 15–22, 2011.

as well as health care service applications [1]. In the sensor networks, MANET (Mobile Ad Hoc Network)-like environment wherein no AP (Access Point) or other fixed infrastructure exists is used, wherein a relatively large number of sensor nodes are deployed throughout a sensor field covering a large area, resulting in various, dynamic topology. Also, autonomous and independent networks are formed between the sensor nodes [2-3]. One of the most important issues about WSN is ensuring the efficiency in energy use, which arises from the energy resource constraints imposed on sensor nodes. For sensor nodes, batteries are the main source of energy, but their operation characteristics do not allow the replacement or charge of batteries. There are three items to evaluate the performance of WSNs: energy efficiency; accuracy of the data; and service quality. Among these items, energy efficiency is the most important one. It acts as a factor that indicates the lifetime of the sensors, as they function, consume energy, and wear out over time [4].

The sensor nodes comprising a WSN are often small and have limited battery capacity. Moreover, the batteries, in most cases, are not replaceable or chargeable due to the characteristics of their operating environment. It is thus crucial that the sensor nodes be designed in such a way that they maximize the efficiency of their energy use and that the process involved be run effectively. One of the solutions to the energy problem is clustering [3]. Networks adopting clustering are hierarchical, and consist of upper-level and lower-level nodes. Because forwarding of data to a remote base station (BS) requires a tremendous amount of energy, transmitting is carried out by only a few upper-level nodes (cluster heads) selected from all the cluster nodes. A big energy spender, upper-level nodes are designed to make sure that each node have an equal chance of becoming a cluster head according to the set probability and transmit data. Taking turns helps extend the lifetime of the sensor network [4]. Low Energy Adaptive Clustering Hierarchy (LEACH), a representative cluster-based routing protocol, assumes that each lower-level node always has data to transmit [5]. The assumption means that in some cases data that were sensed previously may also be sensed afresh, which is a drawback. Under LEACH, nodes react immediately to changes in the sensed data. Threshold Sensitive Energy Efficient Sensor Network Protocol (TEEN), on the other hand, uses threshold values which help prohibit data sensed previously from being sensed afresh or from being sent to upper-level nodes [6]. But this feature can be a drawback to TEEN, since the sensed data failing to meet the thresholds will not be sent to upper-level nodes. Therefore, in this paper, an energy-efficient clustering technique taking into account the way lower-level nodes collect data was proposed.

2 Related Studies

2.1 Overview of Routing Protocols for Wireless Sensor Networks

A WSN is comprised of nodes that are connected to sensors detecting changes in the data under monitoring. Figure 1 illustrates the architecture of the network.

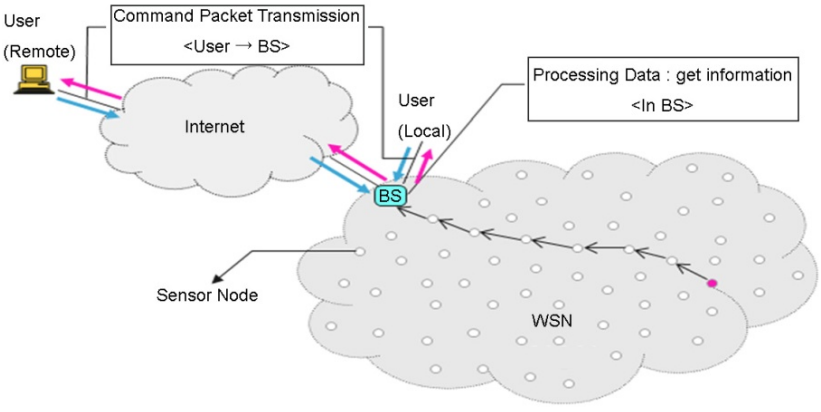


Fig. 1. Wireless sensor network architecture

Depending on network configuration, routing protocols for WSNs are divided into flat routing protocol and cluster-based hierarchical routing protocol. In flat routing protocol, the entire network is considered a unit, and all the nodes therein participate in routing function with equal probability. In comparison, a cluster-based hierarchical routing protocol uses various clustering processes to divide the network into a number of clusters (units) which group the nodes into a hierarchy according to their roles [4]. In this protocol, lower-level nodes collect sensed changes in data and send them to upper-level nodes. The upper-level nodes then aggregate the data and forward them to BS. Some of the best-known cluster-based hierarchical routing protocols are LEACH, LEACH-C (LEACH-Centralized), and TEEN [7].

The routing protocols can also be divided into proactive network protocol and reactive network protocol depending on the network's mode of functioning and type of target application. In the proactive network protocol, the nodes in the field periodically switch on their sensors and transmitters but do so only during the time slots assigned to them. The nodes sense changes in the data and transmit them to the upper-level nodes. The proactive network protocol is suitable for applications that require periodic monitoring of data. LEACH and LEACH-C are among the examples of this protocol. In the reactive network protocol, all the nodes in the field are engaged constantly in sensing changes in the data. The nodes react immediately to the changes, and the data are transmitted immediately to the upper-level nodes. It is better suited for time-critical applications. TEEN is included in this protocol.

2.2 LEACH

A WSN using LEACH is built with several clusters, each of which has a cluster head (CH) and non cluster heads (Non-CHs). The CH controls all the sensor nodes within the cluster, fuses data sent by the sensor nodes, and forwards them to BS. Non-CHs, on the other hand, collect data and send them to the CH. Since it is in charge of aggregating data transmitted from Non-CHs and forwarding them to a remote BS, the CH consumes a lot of energy. Thus, the CH is selected from all the nodes at the beginning of a new round, according to the set probability. This taking turns allows each node to have an equal opportunity to become the CH [5].

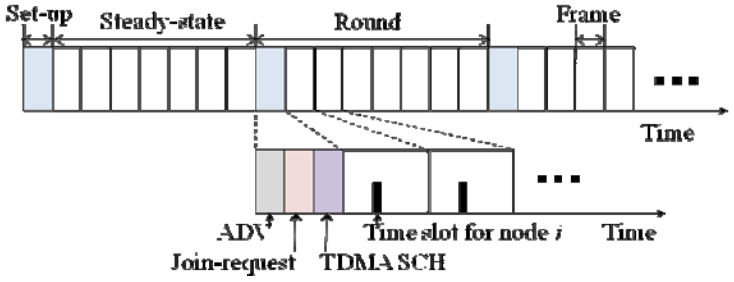


Fig. 2. Timeline showing operation of LEACH

As shown in Figure 2, the operation and structure of LEACH schedule are based on rounds. At each round, Non-CH nodes select their respective CHs (i.e., the set-up stage); data are then transmitted from Non-CHs to the CHs and to BS (i.e., the steady-state stage).[5] Unless it is their time slot, Non-CH nodes remain in sleep mode to save energy. Upon the completion of the current round, a new round begins, with new CHs selected, repeating the aforesaid process all over again. Under this protocol, Non-CH nodes send sensed data to the CHs even if they were the same as the ones sensed previously. In other words, the nodes transmit unnecessary data while consuming energy of the member nodes. Also, in LEACH, CHs are selected based on probability; and clusters are formed based on the location of the selected CHs. This system, therefore, could lead to clusters with non-favorable topology.

2.3 TEEN

Though its cluster structure is basically the same as that of LEACH (proactive network protocol), TEEN is a reactive network protocol where all nodes sense data continuously by reacting immediately to the changes occurring in the data. The reactive network protocol uses two types of threshold in sensing data: hard threshold (HT) and soft threshold (ST).[6] HT is the absolute value of the sensed data. When the value of the data collected by Non-CH nodes is either the same as or less than HT, the data get to be transmitted to the CHs. ST, on the other hand, is a small change in the value of the sensed data. When this value matches or exceeds data collected by Non-CH nodes, the data will get to be sent to the CHs. Once the network starts operating and when the sensed data reach their HT value, Non-CH nodes send the data to the CHs. (The value of the sensed data is stored in the Non-CH nodes.) Next, at the current cluster period, nodes will transmit data only when the current value of the sensed data exceeds HT and at the same time matches or exceeds ST. The purpose of using HT then is to have nodes transmit only the data of importance and to help reduce the number of transmissions conducted by them. In comparison, ST is set to detect minor changes in data when the value of the sensed data is greater than HT. A drawback to TEEN is that the use of thresholds leads to other problem, i.e., Unless the sensed data reach thresholds, communication with upper-level nodes will never occur.

3 An Energy-Efficient Clustering Technique Taking into Account Data Collection Practice

LEACH assumes that each lower-level node always has data to transmit. When each new cluster is created, Non-CH nodes select the closest CH based on the strength of the ADV (Advertisement) signal information transmitted by the CH. By selecting the closest CH, Non-CH nodes are expected to consume the least amount of energy during the transmission. Nevertheless, the consumed energy is still large compared to the one spent during the nodes' sleep mode. Moreover, non-favorable cluster topology, where the distance between Non-CHs and their CHs is far, can result from the LEACH protocol's use of probability in selecting CHs. Previous research on LEACH focuses extensively on how to improve the energy efficiency of CHs, the biggest energy spender, based on the aforesaid assumption of the protocol. In this paper, a method for extending the lifetime of a WSN was explored by focusing on how to reduce the energy consumption at Non-CH nodes, the majority that makes up the cluster. The proposed scheme offers an alternative to help improve the energy efficiency of the entire network while taking into account the data collection occurring at Non-CH level in cluster-based routing protocols. When applied to the current LEACH protocol, the proposed scheme undergoes the process illustrated in Figure 3. In the process, the set-up stage, during which the cluster is created, ends and then triggers a new period, during which Non-CH nodes at once sense data in the environment and compare them with the ones stored in their internal memory during their time slot. In the initial period, no previously collected data exist, and hence no match between data sets. The non-match will next allow the data to be stored in the nodes' internal memory and at the same time to be transmitted to the CHs. Upon the completion of the transmission, Non-CH nodes switch to sleep mode, just as their counterparts in the current LEACH protocol do.

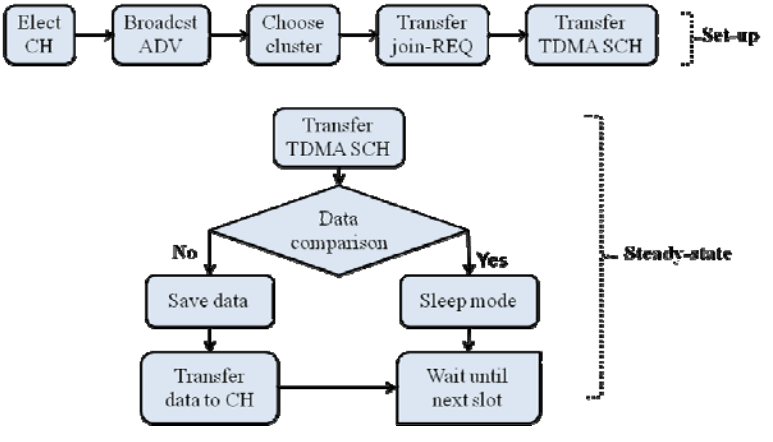


Fig. 3. Operation of the proposed scheme applied to LEACH

When the second period begins, Non-CH nodes come back from sleep mode and start sensing the environment and collecting data during the second time slot assigned to them, according to the TDMA (Time Division Multiple Access) schedule created by the CH of their respective cluster. The data collected this time will be compared with the one stored previously in the internal memory of the nodes, in the same manner adopted during the first time slot. If the previous and current data achieve a match, the nodes will switch to sleep mode to save energy. If the data do not match, the nodes will store the current data in their internal memory and then transmit them to the CHs. This process will be repeated over and over again until the lifetime of the network expires. Note that with the proposed scheme, the advantage of using thresholds in TEEN (i.e., selective collection of data of interest from all the data sensed in the network field) is compromised to a certain degree, because Non-CH nodes in the proposed scheme simply compare the previously sensed data with the currently sensed one and either store and transmit or switch to sleep mode. But the proposed scheme will help solve the big problem in TEEN, i.e., the data collected by Non-CH nodes that fail to reach the thresholds will never get to be transmitted to upper-level nodes. In other words, the proposed scheme will: (a) maintain the characteristics of reactive network protocol by having Non-CH nodes continuously sense the data in the field, react immediately to changes in the data, and transmit them to upper-level nodes; and (b) simultaneously help solve the aforesaid problem of TEEN.

4 Performance Evaluation

4.1 Simulation Environment

For the simulations, it was made sure that the WSN had 100 nodes and a fixed BS. The nodes were randomly assigned to one of the three network algorithms, i.e., LEACH, TEEN, or the proposed scheme. For clustering and other issues, the parameters used in LEACH are listed in Table 1 [5, 8]. To help ease the simulation of the proposed scheme, an environmental scenario was created wherein the temperature was manipulated to fluctuate randomly between 0°C and 200°C at five-second intervals — as applied in TEEN. Regarding energy consumption, incorporating the proposed scheme into an existing routing algorithm for real-life simulation would result in energy consumption that is required for the data comparison process at Non-CH level. But the comparison concerned involves only simple comparison between the previously stored data and the currently sensed one. Thus, the amount of energy consumed for the comparison was judged to be sufficiently small, and it was ignored during the simulation of the proposed scheme.

Table 1. Simulation parameters for the proposed scheme

Parameter	Network grid	Base station coordinate	Number of sensor nodes	E_{elec}	E_{amp}	Initial energy/ node	k	E_{da}	E_{fs}	E_{mp}
Value	From (0,0) to (100,100)	(50, 175)	100	$50nJ/bit$	$100pJ/bit/m^2$	2J	5	$5nJ/bit$	$10pJ/bit/m^2$	$0.0031pJ/bit/m^4$

Another issue to consider was the variability in chances of having the same data in the provided simulation environment wherein temperature changed randomly. Thus, the validity of simulation results needed to be improved. This was achieved by conducting as many simulations as possible (i.e., at least 10 times) and combining the data to help reduce the standard deviation.

4.2 Results and Analysis

Simulations were carried out to compare LEACH and the proposed scheme, and the number of surviving nodes in each protocol was compared over time. As shown in Figure 4, the existing LEACH protocol had nodes that survived a maximum of 630 seconds, whereas the proposed scheme contributed to the survival of nodes for up to 800 seconds. In terms of the maximum lifetime of the network, the proposed scheme increased it by approximately 27% compared to the LEACH model. The increased lifetime is mostly contributable to the reduced energy consumption at Non-CH level. Under the proposed scheme, the CH (upper-level node) consumed less energy, too, because there were less data transmitted from the Non-CH nodes in its cluster, and hence less information to fuse and less energy to spend. This indicates that increasing energy efficiency at Non-CH level is just as important as reducing energy consumption at CH level, i.e., the focus of the majority of the previous studies. Next, simulations were conducted to compare TEEN against the proposed scheme, i.e., to compare the number of surviving nodes over time while implementing TEEN's hard mode (HT) as opposed to implementing the proposed scheme. Under TEEN, the simulation set-up including cluster configuration was mostly similar to that of LEACH. However, the radio electronics model used in LEACH had to be altered so that the model in TEEN would represent both the idle time power dissipation and the sensing power consumption. This was necessary because in TEEN algorithm all the nodes sense the environment continuously.

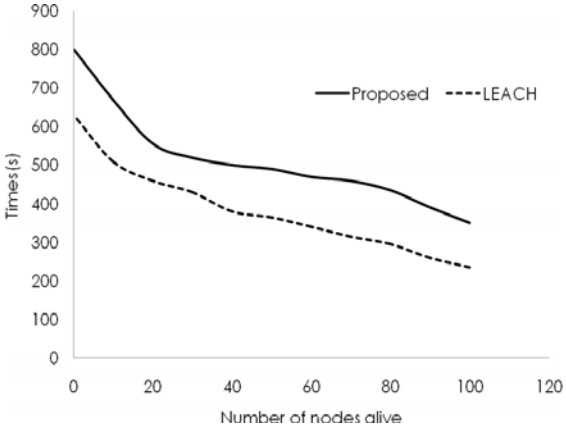


Fig. 4. Comparison of the no. of surviving nodes over time (LEACH vs. proposed scheme)

5 Conclusion

LEACH, a representative cluster-based routing algorithm, assumes that each lower-level node always has data to transmit. The assumption means that in some cases data that were sensed previously and thus need not be sent to upper-level nodes may well be being transmitted afresh. The redundancy problem was what inspired this study. In this paper, a protocol scheme was proposed to help save energy at lower-level nodes that compare the currently sensed data with the previously sensed one during their time slot and decide whether or not to send the data to upper-level nodes (CHs). The proposed scheme was compared with LEACH and TEEN through simulations. The results indicate that the use of the proposed scheme helped increase the energy efficiency of lower-level nodes as well as decrease the energy consumption required during the data aggregation by upper-level nodes (CHs). As a result, the lifetime of the entire LEACH network increased by 26% at the maximum. Compared with TEEN, the proposed scheme performed slightly better TEEN and helped solve the biggest problem it has, i.e., the backfiring of TEEN's use of thresholds to prevent redundancy in data transmission which in some cases acts to prevent communication with upper-level nodes entirely.

References

1. Ian, F., Akyildiz, F., Su, W., Sankarsubramaniam, Y., Cayirci, E.: A Survey on Sensor Networks. IEEE Communications Magazine, 102–114 (August 2002)
2. Hac, A.: Wireless Sensor Network Designs. John Wiley & Sons, Ltd. (2003)
3. Santi, P.: Topology control in wireless ad hoc and sensor networks. ACM Computing Surveys (CSUR) 37(2) (2005)
4. Heinzelman, W.B., Chandrakasan, A., Balakrishnan, H.: Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In: Proc. of the Hawaii International Conference on System Science, pp. 3005–3014 (January 2000)
5. Chandrakasan, P., Heinzelman, W.B.: Application-Specific Protocol Architectures for Wireless Networks. IEEE Transactions on Wireless Communications 1, 660–670 (2000)
6. Manjeshwar, A., Agrawal, D.P.: TEEN: A Protocol for Enhanced Efficiency in Wireless Sensor Networks. In: Proc. of 1st Intl. Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco (April 2001)
7. Heinzelman, W.B., Chandrakasan, A., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
8. Heinzelman, W.R., Sinha, A., Wang, A., Chandrakasan, A.P.: Energy-scalable algorithms and protocols for wireless microsensor networks. In: Proc. IEEE Acoustics, Speech and Signal Processing Conf., vol. 6, pp. 3722–3725 (June 2000)

Event-Driven Test Script Methodology for SOA System

Youngkon Lee

e-Business Department, Korea Polytechnic University,
2121 Jeongwangdong, Siheung city, Korea
yklee777@kpu.ac.kr

Abstract. Electronic collaboration over the Internet between business partners appear to be converging toward well-established types of message exchange patterns that involve both user-defined standards and infrastructure standards. At the same time, the notion of event is increasingly promoted for asynchronous communication and coordination in SOA systems. In collaboration between partners or between components is achieved by the means of choreographed exchanges of discrete units of data - messages or events - over an Internet-based protocol. This paper presents an event-centric test case scripting method and execution model for such systems.

Keywords: SOA, event-driven, process invocation.

1 Introduction

While current Web Service technologies show much progress, current services are mainly limited to atomic services. Thus, they are not adequate to handle the autonomous and complex services in realistic settings. In dealing with this problem, some research works have developed languages to compose the individual Web Services into transactions or workflows. Web Services Flow Language (WSFL) [1] was designed for service compositions in the form of a workflow, and XLANG [2] for the behavior of a single Web Service. However, these works are not sufficient for providing the adaptive web Services generated from a particular context.

Electronic collaborations over the Internet between business partners (e-Business / e-Government) appear to be converging toward well-established types of message exchange patterns that involve both user-defined standards and infrastructure standards. At the same time, the notion of event is increasingly promoted for asynchronous communication and coordination in Event-Driven Architectures (EDA) that are considered as either complementary to or part of SOA systems. In both cases collaboration between partners or between components is achieved by the means of choreographed exchanges of discrete units of data - messages or events - over an Internet-based protocol. Such systems require an event-centric test case scripting markup and execution model.

In e-Business transactions as in EDAs, partners or components must agree on the use of a combination of standards in order to interoperate with each other. Typically, these standards can be classified into three layers:

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 23–30, 2011.

- Messaging infrastructure standards, ranging from transport level to higher-level messaging protocols and quality of service (QoS) including reliability and security, such as those defined as SOAP extensions, or REST (Representational State Transfer).
- Multi-message exchange standards as manifested in business processes and choreographies.
- Business document standards may be business content structure and semantics, taxonomies in use, code lists, semantic rules, or the XML schema modeling style. They are industry-specific (e.g. RosettaNet PIP schemas, AIAG Inventory Visibility and Interoperability schemas), horizontal document standards, or regional guidelines.

There have been conformance and interoperability test suites and testing tools for above layer individually. But the testing of integrations of standards has been ad-hoc, or limited mostly to standards in the messaging infrastructure.

Although the need for testing some form of integration of standards has been well recognized for infrastructure standards, there has been little support for testing integrations that extend to the use of standards specific to a business - e.g. for documents or choreographies. Such integrations can be construed as user-defined profiles. For example, the level of QoS required for a business transaction may depend on the nature of business data being exchanged, or on some property defined by the related business process.

Testing and monitoring these infrastructure layers and their integration also requires that test cases access a combination of contracts - agreements, policies or business transaction patterns - represented by meta-level documents.

This compliance objective goes beyond quality assurance for the messaging function: it requires the monitoring of live transactions in production environments, as well as verifying conformance of business endpoints in operation conditions. This calls for a flexible test execution model that can accommodate performance constraints as well as different timeliness constraints - e.g. where tests are either deferred over log data, or executed on live exchanges in a monitoring mode.

Consequently, the execution model of such test cases or monitoring cases, must accommodate dynamic conditions requiring real-time or near real-time error detection or measurement, allowing to correct and report on business exchanges as they proceed.

The output of a monitoring script also must provide more information than a report of the type pass / fail. Different ways of "passing" or "failing" must be reported on, as well as identifying the types of business transactions. The output must be easy to format and feed to another decision engine, such as a rule engine that will process this input in real-time. For example, a rule may decide to generate an alert if a business transaction lasts too long, depending on the nature of the transaction and on the SLA associated to these business partners.

This paper defines a testing and monitoring model, as well as a test script markup, so that test cases or monitoring cases can be fully automated, and portable across test environments. In section 2, we summarize the related works regarding the web service flow language for testing. Section 3 presents the concept of Event-Centric

Test Case Script (EVEC), section 4 describes the implementation of EVEC, and we conclude in section 5.

2 Related Works

In fact, the automatic or semi-automatic management of service flows over the Web has not been achieved yet. In the Web Services model that is quite different from traditional one, there are a large number of similar or equivalent services which user can freely select and use for their application. Since the service is developed and deployed by the third party, the quality of service is not guaranteed. The services may not be adequate as per service requestor's requirements and kept evolving, without notification to service requestors, according to the provider's requirements and computing environment.

Thus, it is important to provide adaptability to evolving services as well as diverse context of services. Kammer et al. [3] suggested workflow to be dynamic, which allows changes with minimal impact to the ongoing execution of underlying workflow, as well as be reflexive, which provides knowledge about a workflow's applicability to the context and the effectiveness of its deployment evaluated over time. Understanding constraints and context associated with services may affect the quality of service. From this perspective, optimization may occur through the evaluation and refinement of a previous service flow.

Automatic composition of services is challenging, because it is difficult to capture semantics and context of services and measure the quality of services. One exemplary effort that aims for this function is DAML-based Web Service Ontology (DAML-S) [4], which describes the properties and capabilities of Web services.

Workflow technology has been around since a decade ago and has been successful in automating many complex business processes. A significant amount of work has been done in this field, which deals with different aspects of workflow technology process modeling, dynamic workflows, and distributed workflows. Process modeling languages such as IDEF, PIF, PSL or CIMOSA [5] and frame based models of services were used to design process typing, resource dependencies, ports, task decomposition and exception.

Current research on web services paves way for web service based workflows, which has obvious advantages pertaining to scalability, heterogeneity, reuse and maintenance of services. Major issues in such inter-organizational service based workflows are service discovery, service contracts and service composition. Web Services Flow Language (WSFL) was proposed to describe compositions of services in the form of a workflow, which describes the order of service invocation. Service composition aids such as BizTalk [6] were proposed to overcome the limitations of traditional workflow tools which manually specify the composition of programs to perform some tasks. Other industrial initiatives such as BPEL4WS [7], and XLANG concentrates on service representation issues to tackle the problems of service contracts, compositions and agreements. Current efforts are to automate the complete process of service discovery, composition and binding, using the machine

understandable languages. Some other recent advances are WS-Transaction and WS-Coordination which define protocols for executing Transactions among web services. There is a research for modeling QoS of workflows [8], and defining a QoS based middleware for services associated with the underlying workflow [9], but it doesn't take into account QoS factors related to Internet based services. Some researchers describe QoS issues related to web services from the provider's perspective [10]. We believe that the current research has not delved into QoS issues related to Web Service based workflows, and many critical issues related to the availability, reliability, performance and security of Web Services need to be handled. Our approach tactfully utilizes and monitors these QoS parameters to provide a consistent service interface to other applications in the workflow through adaptive QoS based selection, binding and execution of Web Services.

3 Event-Centric Test Case Script

Event-centric test case script (EVEC) is designed so that the same scripts can be used either in live monitoring mode, or in analysis of past events from a log (referred to as deferred mode in this paper called hereafter the "deferred mode") or yet in mixed situation. Testing and Monitoring of Business Processes as well as more generally of systems the behaviour of which can be traced by events, fall in the following three categories:

- Compliance with specifications. Such specifications may be of a business transaction, business process definition, documents exchanged, or of infrastructure behaviour (e.g. messaging protocol). Enabling the automatic generation of EVEC scripts from such specifications when these specifications are formal – e.g. process definition, choreographies, document schema or rules – is part of the requirements although the methodology to achieve this is out of scope of this document. Some test assertion design and best practices, such as those in Test Assertions Guidelines [11] may be used for deriving scripts from such representations even when automatic generation is not possible.
- Compliance with agreements. Such agreements may be business agreements such as SLAs, or regulatory compliance rules. They may also be infrastructure configuration agreements (e.g. ebXML CPA, WS-Policy). This category of application includes SLA monitoring, business metrics and aspects of business activity monitoring (BAM) that are closest to operations, e.g. for regulatory compliance.
- Business Operation intelligence. Such monitoring is not directly related to compliance, but primarily intended for generating reports and various analytics of business activities. This includes analyzing the logs of processes and business transactions for reports and BI. This category of application includes BAM (business activity monitoring). In its dynamic aspect, this monitoring is addresses the need for visibility in business processes and service-oriented systems, which include problem detection/anticipation, diagnostics and alarm generation. Each one of the above categories may be considered both in a real-time context (e.g. generation of alarms

and notifications during operation) and a deferred, off-line analysis context (periodic generation of reports or metrics with no direct, automatic feedback loop to operations). In both cases, the same input – in form of events – is assumed.

From the viewpoint of script execution semantics, "live" and "deferred" are not distinguished: the same script is executable on input that is either live or logged. To ensure flexibility for handling various monitoring contexts and situations, mixing of both execution modes must be supported:

- A script may start executing "deferred mode" with its events already partially logged, and then catch-up with the on-going logging of events and continue "live".
- Conversely, a script may start live, and if its execution engine is interrupted for some reason, may resume its analysis of events that have already been logged while the test engine was stopped, in deferred mode. Then it may catch-up with events and eventually go live again. When events are consumed in a publish-subscribe mode, a simple queuing mechanism is sufficient to provide the above flexibility. However, EVEC must be able to correlate with past events.

4 Implementation of EVEC

The EVEC script language is designed for testing and monitoring processes or business transactions of various kinds, and more particularly for analyzing and validating event patterns that are generated by these processes. To this extent, EVEC may be described as an event-processing language. The content and structure of these events may be quite diverse, but a common XML wrapper is assumed. The top-level constructs are the script package and the scriplet:

- The script package, or “script”: This is the main unit of execution. The script package contains an “execution context” (<execution-context> element) that defines various global constructs and bindings, e.g. for declaring event boards. The script package also contains one or more "scriplets". The execution context in a script package defines which scriplet to start execution with - or main scriplet. In case other scriplets are defined, the main scriplet is expected to invoke these directly or indirectly.
- The scriplet: A scriplet defines a flow (or thread) of execution for a sequence of atomic operations. Scriplets can execute either concurrently or not (see detailed meaning in the next section), and can be started in a blocking or non-blocking way.

EVEC is designed so that it leverages existing XML script languages for special features such as logical conditions and event selection. The default language for all logical expressions over XML content is XPath, along with its existing function libraries (e.g. advanced set of functions for time and duration management).

The concept of concurrency in EVEC is entirely dependent on the notion of “virtual present time” (VPtime). When a scriplet starts to execute, it is assigned a VP-time which will condition its event consumption and timestamp its event production. The default VP-time assignments are:

- The first scriplet of a script package is assigned the initial VP-time of this script, the default of which is in turn the actual present time (AP-time).
- The VP-time of a scriplet S2 started by a scriplet S1, is the value of VP-time in S1 when [start S2] is executed. These default values can be overriden by the <start> operation, which allows to set the VP-time of the started scriplet (see the start/@vptset attribute in section 4). Inside a scriplet, the VP-time may be progressed by two operations:
- <wait> : will add some predefined duration to the VP-time, or wait until some date, or yet until some other scriplets complete.
- <catch> : when waiting - virtually or really - for some event to occur, will advance the VP-time to the occurring date of events being caught.

Event catching in a scriplet is only considering (by default) events occurring at or after the current VP-time. Besides <wait> and <catch>, the execution duration of other EVEC operations is considered as negligible, as far as the VP-time is concerned: in other words, these operations do not affect the VP-time. The VP-time window of a scriplet execution is defined as the [starting VP-time, ending VP-time] time interval of the execution. Intuitively, concurrent execution is achieved when the VP-time windows of two scriplets overlap. Concurrency is entirely determined by the start VP-time assigned to these scriplets. When a scriplet S1 starts a scriplet S2, it can do so either in a blocking or non-blocking way:

- Blocking invocation: intuitively, the invoking scriplet (S1) will wait until the invoked scriplet (S2) terminates. The next statement in the invoking scriplet S1 (after <start>), will execute at a VP-time that is same as the VP-time set at the end of the invoked scriplet (S2). In other words, the VPtimes of S1 and S2 are “synchronized” after S2 completes (see start/@vptsync="true" in Section 4). More generally, to accommodate the case where a starting time is set at a date/time anterior to the invoking time (@vptset="a past date/time”), the VP-time of the next statement in S1 is either the last VP-time value of S2 or the last VP-time value in S1 (just before invoking S2), whichever occurs the latest.

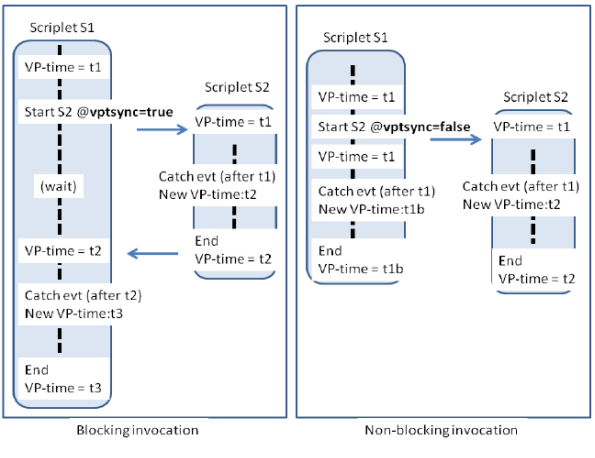


Fig. 1. Blocking and non-blocking modes of scriplet invocation

- Non-blocking invocation: intuitively, the invoked scriplet (S2) will not cause the invoking scriplet (S1) to wait. In other words, the VP-times of S1 and S2 are not "synchronized" after S2 completes. (see start/@vptsync="false" in Section 4). The next statement in the invoking scriplet S1 will execute at a VP-time that is same as the VP-time value just before executing the <start> statement, this regardless of the value of start/@vptset. Non-blocking invocations should not be seen as only useful for actual concurrent (or multi-threaded) processing. In many cases, it makes scripting easier and more intuitive, even when execution is entirely deferred on past (logged) events that could otherwise be processed serially in a single-threaded way. Various cases of blocking and non-blocking invocations are illustrated below. The following figure illustrates both modes of scriplet invocations, and how the VP-time is affected – or not - in the invoking scriplet.

When a scriplet S1 does a blocking invocation of a scriplet S2, the VP-time of the invoked scriplet S2 is initiated at the current VP-time of the invoking scriplet S1 (unless a different VPtime value is given using start/@vptset as illustrated in the next figures). The scriplet S1 is then "blocked" until the VP-time at the end of S2 is known and assigned as current VP-time in S1.

In the non-blocking invocation (see Fig. 1.), S1 will ignore the ending time of S2. A single-threaded execution may still execute S2 first before executing the next statement in S1. The execution semantics would still allow "concurrent" catching (in terms of virtual present) of same events by S1 and S2, as both will select events starting from VP-time t1. In the above figure, S1 is catching an event at time t1b while S2 is catching an event at time t2. Depending on their respective selection expressions, these catches could capture either the same or different event, causing the new VPtime in S1 (t1b) to be either prior or after the new VP-time in S2 (t2).

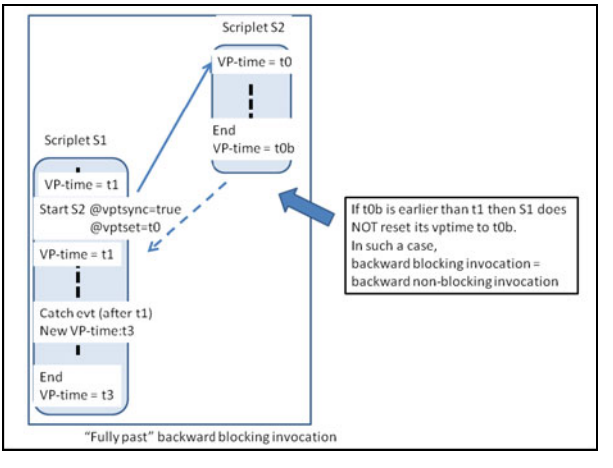


Fig. 2. Fully past backward blocking invocation

In a single-threaded execution of the non-blocking case that started "live" (VP-time = present time), S2 could be executed first live, then the remaining part of S1 can be executed "deferred" on the log of events, starting from time t1 now in the past. Clearly, more fine-grain serialization of S1 and S2 executions would be required if these two scriplets communicate with each other, e.g. if S1 is consuming an event posted by S2 or vice-versa.

5 Conclusion

In this paper, we present an event-centric test case scripting method and execution model, EVEC, for such systems. EVEC enables testing and monitoring of applications and business processes, the behavior of which can be traced and monitored via events. The notion of event in EVEC is broad, thus, it could cover all of the type of SOA business processes. Further study is required to define and classify the detailed test case metadata or artifacts that would complement EVEC in test environments.

References

1. Leymann, F.: Web services flow language, TR WSFL 1.0. IBM Software Group (May 2001)
2. Thatte, S.: XLANG Web Services for Business Process Design (2001), http://www.gotdotnet.com/team/xmlwsspecs/xlang-c/default.htm
3. Kammer, P., Bolcer, G.A., Taylor, R.N., Bergman, M.: Techniques for Supporting Dynamic and Adaptive Workflow. Journal of Computer Supported Cooperative Work (CSCW) 9, 269–292
4. DAML-S Specifications, http://www.daml.org/services/
5. Kosanke, K.: CIMOSA - Open System Architecture for CIM; ESPRIT Consortium AMICE. Springer (1993)
6. Biztalk, http://www.microsoft.com/biztalk/
7. Business Process Execution Language for Web Services, Version 1.0 (July 2002), http://www-106.ibm.com/developerworks/webservices/library/ws-bpel/
8. Cardoso, J., Sheth, A., Miller, J.: Workflow Quality Of Service (2002)
9. Sheth, A., Cardoso, J., Miller, J., Koch, K.: QoS for Service-oriented Middleware "Web Services and Grid Computing. In: Proceedings of the Conference on Systemics, Cybernetics and Informatics, Orlando, FL (July 2002)
10. Mani, A., Nagarajan, A.: Understanding quality of service for Web services, http://herzberg.ca.sandia.gov/jess/
11. OASIS Test Assertions Guidelines (TAG) TC, Test Assertions Guidelines, http://www.oasis-open.org/committees/tag/

Business-Context Based SLA Parameters for SOA Management

Youngkon Lee

e-Business Department, Korea Polytechnic University,
2121 Jeongwangdong, Siheung city, Korea
yklee777@kpu.ac.kr

Abstract. With Web services starting to be deployed within organizations and being offered as paid services across organizational boundaries, quality of service (QoS) has become one of the key issues to be addressed by providers and clients. While methods to describe and advertise QoS properties have been developed, the main outstanding issue remains how to implement a service that lives up to promised QoS properties. This paper provides the service level agreement (SLA) parameters for QoS management applied to Web services and raises a set of research issues that originate in the virtualization aspect of services and are specific to QoS management in a services environment – beyond what is addressed so far by work in the areas of distributed systems and performance management.

Keywords: SOA, business context, SLA parameter.

1 Introduction

Whether offered within an organization or as a part of a paid service across organizational boundaries, quality-of-service (QoS) aspects of services are important in a service -oriented computing environment. Dealing with QoS is a sign of a technology going beyond its stage of initial experimentation to a production deployment and many recent activities related to QoS of Web services indicate that this is becoming an increasingly relevant topic.

Efforts in the past years mainly focused on describing, advertising and signing up to Web and Grid services at defined QoS levels. This includes HP's Web Services Management Framework (WSMF) [1], IBM's Web Service Level Agreement (WSLA) language [2][3], the Web Services Offer Language (WSOL) [4] as well as approaches based on WS-Policy [5]. These efforts enable us to describe quality metrics of services, such as response time, and the associated service level objectives flexibly and in a way that is meaningful for the business needs of a service client.

However, one of the challenging issues is to associate or derive a system configuration that delivers the QoS of a described Web service using the above mentioned approaches. In many cases, this is non-trivial. Sometimes we can rely on experience with tested, dedicated system configurations to decide, for example, the size of a cluster for a particular workload guaranteeing a particular response time for a

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 31–38, 2011.

given percentile of requests. In addition, managing a service at different QoS levels on the same infrastructure is not easy.

While managing QoS in distributed systems is not a novel problem, a number of additional issues arise in the context of a service-oriented computing environment. Those issues arise from the specific properties of Web services. For example, cross-organizational Web services may be accessed through the public Internet and client side QoS metrics have to include network properties in addition to properties of the service-implementing application itself. In addition, composite and recursively aggregated services – and the ability to aggregate is seen as a key benefit of Web services – gain new QoS properties that are not always easily derived from their parts.

The objective of paper is to analyze the main QoS factors of Service-oriented Architecture (SOA) in a business context and to provide the service level agreement (SLA) parameters that affect critically the business performance.

According to OASIS Reference Model for Service Oriented Architecture [SOA-RM] [6], the Service Oriented Architecture (SOA) is a paradigm for organizing and utilizing distributed capabilities that may be under control of different ownership domains. The service within SOA is a mechanism to enable access to one or more capabilities, where the access is provided using a prescribed interface and is exercised consistent with constraints and policies as specified by the service description. This specification further defines the business service level agreement (bSLA) between the service requester and the service provider for the service which is defined in SOA-RM, within the end-to-end resource planning (EERP) technology [7]. The applications of EERP are any kind of business services, and they are not limited to Web Services only. This applies the well-known technique for service discovery and optimization in a novel way to improve business results. It models the business process and the range of potential services, and then guides the selection and deployment of services based on the end-to-end business value. Modeling the business service-level agreements to manage and evaluate services and establishing agreements about the business service is essential to long-term value chain improvement. The bSLA is different from the SLA in the software/IT world. The bSLA is the contact between the service requester and the service provider, and the SLA is the contract between the service provider and the network/system provider. The SLA is network/system oriented agreement that deals with network performance and system availability. The bSLA is a business oriented agreement that deals with price, time to deliver, and the quality/rating of the service.

In section 2, we summarize the related works about web service selection based on the QoS metrics. Section 3 presents the service process model which enables service performance optimization in the business respect and section 4 details the bSLA model including parties, parameters and obligations, and we conclude in section 5.

2 Related Works

The first step to manage Web service's quality is to define it. While this is important for Web services as well as in traditional distributed systems, explicit definition is

particularly important in an environment transcending organizational boundaries. Quality is expressed referring to observable parameters relating to a non-functional property, for example, the response time of a request. A level of quality is agreed upon as a constraint over those parameters, potentially dependent on a precondition. Hence, the party offering a Web service, in agreement with its customers and users, will define the QoS parameters and the particular instances of the service to which these parameters relate. In the case of a Web service, a parameter such as response time can relate to an individual invocation of an operation or a class of operations, all having the same (individual) quality properties of having an aggregate property, e.g., the average response time of this class of operations or another stochastic metric.

A further step in managing Web services QoS is the definition of the semantics of the QoS parameters. A Web service and its subscribers and users must understand what it is meant. It is important what is measured where. For performance-oriented metrics this can be at different points, as Fig. 1 illustrates.

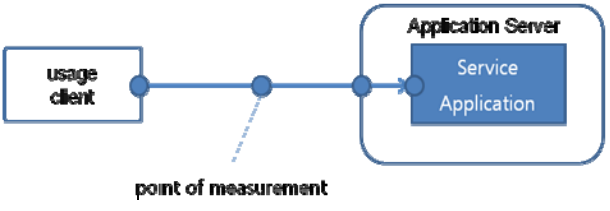


Fig. 1. Points of measurements defining semantics of metrics

The definition of QoS parameters corresponds to the establishment of ontology between a service provider and its clients. Ontology can be established in two approaches. (1) It can be a definition of terms and, potentially, or the semantics of the relationships between them, as facilitated by DAML and OIL [8]. This approach results in a fixed set of well understood terms – in our case the QoS parameters. (2) Another approach uses constructive ontology. Based on a set of well-know defined terms (as in 1) and a set of well-know composition operators, new terms' (QoS) parameters can be defined by composing new parameters out of existing ones using the operators.

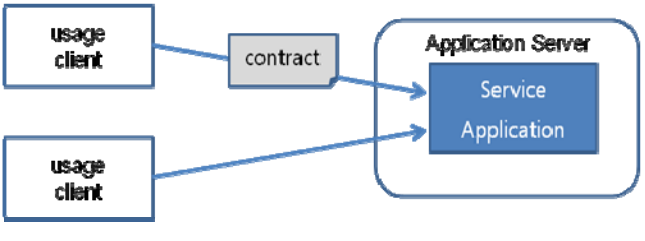


Fig. 2. Contracts defining the scope of quality guarantees

Having established common understanding of quality of service parameters and the associated guarantees given by the provider, it also has to be established to which relationships between a client and a server a QoS guarantee applies. A service may provide the same quality to all requesting clients, to each client individually or to a defined set of clients that a provider organization and an organization requiring a QoS level for multiple clients agree upon in a contract, which is also called an SLA.

Clients will refer to the contract when requesting the service according to a particular quality level. The different scoping approaches of QoS guarantees require different means of establishing a particular quality level for a client: If a QoS level is associated with a service, a client searches for a suitable service in a directory, e.g., UDDI and retrieves its quality definition, e.g., stated as a WS-Policy expression. In the case of an individual client or a contract, a negotiation mechanism, which can be very simple, must be provided. Once the contract is established, the provider organization must provision a service-implementing system such that it behaves as it has been agreed upon. This involves deriving the amount of resources needed and the runtime management of resources.

However, this is not simple. While we have developed – improvable – approaches to the issues raised above, the issue of provisioning and runtime managing a Web service-implementing system is not equally well understood yet. In the next section, we discuss what distributed systems and performance management approaches can provide.

A number of performance management technologies, such as workload managers and network dispatchers, have been developed to control response times of individual systems and clusters and various availability management approaches. However, it is not straight-forward to configure, for example, workload managers to satisfy response time goals for a set of different scopes of quality – for Web services as well as for any distributed system. In this section, we outline some typical approaches of how today's QoS parameters are managed in distributed systems.

3 Service Process Model

This section describes the service process model conceptually. Fig. 3 shows the conceptual model, and of messages flows with brief descriptions. We also include timeline and sequence diagrams Fig. 4 to show how an implementation would use service messages and build a continuous business process improvement loop. In Figure 3, the business quality of Service is abbreviated as bQoS, business rating is abbreviated as Rating, and business service level agreement is abbreviated as bSLA.

The service requester is the client system who tries to find an optimal solution provided by service optimization portal (SOP). Service providers provide business services. Each service provider may offer the same service but with different bQoS and Ratings. Services may be running on different platforms with different implementations, but they all support message exchanges of bQoS, Rating, and BSLA information in the XML formats.

The SOP accepts the request from the Service requester, performs bQoS and rating queries, calculates optimal solution(s), and then returns the result to the service

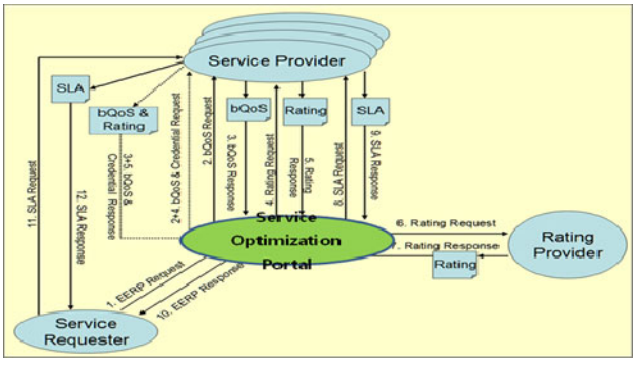


Fig. 3. Service Process Model

requester. The Rating Provider is a party unaffiliated with either the requester or the target of the rating request, such as a third party rating organization, given a reference to a particular business service and provider, issues either a number or a classification description.

There can be another way to implement the service optimization without the SOP. For example, there can be a case for some services providers and service consumers using SOA Registry-Repository to find each other, to exchange business quality of services information among them, and to begin negotiations for establishing Service Level Agreements (SLAs). The results of messages 2 through 9 in Figure 4 are used to calculate the optimal deployment for a given set of services requests. A list of alternatives might be returned in message 10. Each step in the process would have a service provider determined for each service and for each alternative. Messages 11 and 12 are exchanged between the service requester and the selected service providers to define the BSLA.

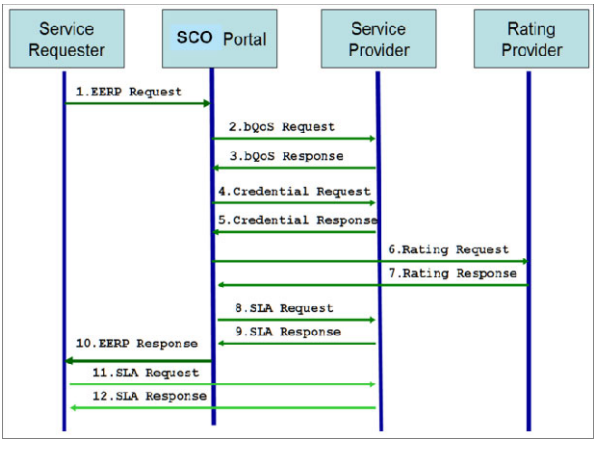


Fig. 4. Service message sequence without optional messages

4 bSLA Model

The bSLA model is for representing SLA parameters in the respect of business value. The BSLA is the root element for EERP- Business Service-level agreement (bSLA). The bSLA is a formal contract between a service provider and a client guaranteeing quantifiable business quality of service (bQoS) at defined levels. It can have one or more of the following elements:

```
<sla:BSLA xmlns:sla="..." xmlns:bqos="..." ...>
  <sla:SLAParties ...>sla:SLAPartiesTypeType</sla:SLAParties>
  <sla:SLAParameters ...>sla:SLAParametersType</sla:SLAParameters>
  <sla:SLAObligations ...>sla:SLAObligationsType</sla:SLAObligations> ?
  <sla:SLATerms ...>sla:SLATermsType</sla:SLATerms> ?
</sla:BSLA>
```

Fig. 5. XML schema for bSLA

The following describes the attributes and elements listed in the schema outlined above:

- /sla:BSLA is a root element of Business Service-level agreement (bSLA) for EERP.
- SLAParties is a required element in bSLA that defines parties invoked in this bSLA for the service. SLAParties element has both the service provider and services requester elements.
- /sla:BSLA/sla:SLAParties/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the SLAParties element in the future. Unrecognized attributes may cause a fault or be silently ignored.
- /sla:BSLA/sla:SLAParameters is defined monitoring of bQoS metrics, including service profile uri, operations and other optional elements. It is a required element that uses sla:SLAParametersType.
- /sla:BSLA/sla:SLAParameters/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the SLAParameters element in the future.
- /sla:BSLA/sla:SLAObligations is agreed bSLA obligations aspect of the service, including obligations, action guarantees. It is a optional element that uses sla:SLAObligationsType.
- /sla:BSLA/sla:SLAObligations/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the SLA Obligations element in the future.
- /sla:BSLA/sla:SLATerms is agreed bSLA terms aspect of the service, including bSLA term elements.
- /sla:BSLA/sla:SLATerms/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the SLATerms element in the future.
- /sla:BSLA/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the root BSLA element in the future.

- /sla:BSLA/sla:BSLAExtension is an optional element that keeps different (extensible) elements to be specified in the future.
- /sla:BSLA/sla:BSLAExtension/{any} is an extensibility mechanism to allow different (extensible) elements to be specified in the future.

The SLAParties describes the list of parties invoked in the bSLA for the service. There should be one SLAParties element present in the bSLA of service. The following describes the attributes and elements listed in the schema outlined above:

- /sla:SLAParties, bSLA Parties aspect of the service, is for parties invoked in the bSLA for the service, including both service provider and service requester elements.
- /sla:SLAParties/sla:ServiceProvider represents the provider for parties. It is a required element for bSLA Parties.
- /sla:SLAParties/sla:ServiceProvider/sla:ServiceUri is a required element for Service Provider.
- /sla:SLAParties/sla:ServiceProvider/sla:ServiceProviderName is the name of the service provider. It is also a required element for Service Provider.
- /sla:SLAParties/sla:ServiceProvider/sla:ServiceProviderName/@languageID is an optional attribute in the ServiceProviderName element, using xsd:language type.

```
<sla:SLAParties xmlns:sla="..." ...>
<sla:ServiceProvider ...>sla:ServiceProviderType
<sla:ServiceUri ...>sla:SlaUriType</sla:ServiceUri>
<sla:ServiceProviderName
languageID="...">sla:ServiceProviderNameType</sla:ServiceProviderName>
</sla:ServiceProvider>
<sla:ServiceRequester ... >sla:ServiceRequesterType
<sla:ServiceRequesterUri ... >sla:SlaUriType</sla:ServiceRequesterUri>
<sla:ServiceRequesterName
languageID="...">sla:ServiceRequesterNameType</sla:ServiceRequesterName>
</sla:ServiceRequester>
...
</sla:SLAParties>
```

Fig. 6. XML schema for bSLA parties

- /sla:SLAParties/sla:ServiceProvider/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to ServiceProvider element in the future. /sla:SLAParties/sla:ServiceRequester represents requester for the service, including requester's name and the URI that represents the requester. It is a required element for bSLA Parties.
- /sla:SLAParties/sla:ServiceRequester/sla:ServiceRequesterUri represents the requester's identifier in URI format for the service requester. It is a required element for Service Requester.
- /sla:SLAParties/sla:ServiceRequester/sla:ServiceRequesterName is requester's name for the service requester. It is a required element for Service Requester.

- /sla:SLAParties/sla:ServiceRequester/sla:ServiceRequesterName/@languageID is an optional attribute in the ServiceRequesterName element.
- /sla:SLAParties/sla:ServiceRequester/@{any} is an extensibility mechanism to allow additional attributes, based on schemas, to be added to the serviceRequester element in the future.
- /sla:SLAParties/{any} is an extensibility mechanism to allow different (extensible) elements to be specified in the future.

5 Conclusion

In this paper, we proposed a new concept of bSLA, whose items are proper to evaluate SOA service performance in the respect of service business benefit. The bSLA includes the service actor information, bSLA parties and the quality information, bSLA parameters, and bSLA obligations of the service parties. We also devised a service optimization portal which provides the best service composition by evaluating the value of bQoS of each service chain. Further study is required to define and classify the quality factor group for business case by case.

References

1. Catania, N., Kumar, P., Murray, B., Pourhedari, H., Vambenepe, W., Wurster, K.: Web Services Management Framework, Version 2.0, Hewlett-Packard, July 16 (2003), http://devresource.hp.com/drc/specifications/wsmf/WSMF-WSM.jsp
2. Ludwig, H., Keller, A., Dan, A., King, R., Franck, R.: A service level agreement language for dynamic electronic services. Electronic Commerce Research 3, 43–59 (2003)
3. Ludwig, H., Keller, A., Dan, A., King, R., Franck, R.: Web Service Level Agreement (WSLA) Language Specification, Version 1.0, IBM Corporation, January 28 (2003), http://www.research.ibm.com/wsla/WSLASpecV1-20030128.pdf
4. Tosic, V., Pagurek, B., Patel, K.: WSOL – A Language for the Formal Specification of Classes of Service for Web Services. In: Proc. of ICWS 2003 (The 2003 International Conference on Web Services), Las Vegas, USA, June 23-26, pp. 375–381. CSREA Press (2003)
5. Box, D., Curbera, F., Hondo, M., Kale, C., Langworthy, D., Nadalin, A., Nagaratnam, N., Nottingham, M., von Riegen, C., Shewchuk, J.: Web Services Policy Framework (WSPolicy) May 28 (2003), http://www.ibm.com/developer-works/library/ws-policy
6. Mackenzie, C.M., et al.: Reference Model for Service Oriented Architecture 1.0. OASIS Committee draft (August 2006)
7. Cox, W., et al.: SOA-EERP Business Quality of Service Version. OASIS Committee draft (November 2010)
8. Connolly, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: DAML+OIL (March 2001) Reference Description, W3C, http://www.w3.org/TR/daml+oil-reference (December 18, 2001)

Prescription-Level Based Test Assertion Model for SOA

Youngkon Lee

e-Business Department, Korea Polytechnic University,
2121 Jeongwangdong, Siheung city, Korea
yklee777@kpu.ac.kr

Abstract. This paper presents a design method for business-centric SOA test framework. The reference architecture of SOA system is usually layered: business process layer, service layer, and computing resource layer. In the architecture, there are so many subsystems affecting the system's performance, which relates with each other. As a result, in respect of overall performance, it is meaningless to measure each subsystem's performance separately. In SOA system, the performance of the business process layer with which users keep in contact usually depends on the summation of the performance of the other lower layers. Therefore, for testing SOA system, test cases describing business process activities should be prepared. We devised a business-centric SOA test assertion model which enables to semi-automatic transform test assertions into test cases by the concept of prescription level and normalized prerequisite definition. The model also minimizes the semantic distortion in the transformation process.

Keywords: SOA, business process, test assertions, test cases.

1 Introduction

Service Oriented Architecture (SOA) is generally defined as a business-centric IT architectural approach that supports integrating businesses as linked, repeatable business tasks, or services. SOA enables to solve integration complexity problem and facilitates broad-scale interoperability and unlimited collaboration across the enterprise. It also provides flexibility and agility to address changing business requirements in lower cost and lesser time to market via reuse.

SOA has a lot of promises of interoperability, however, at the cost of: lack of enterprise scale QoS, complex standards which are still forming, lack of tools and framework to support standards, and perform penalty. Recently, as SOA has been widely adopted in business system framework, performance issues in SOA are raised continuously from users and developers.

SOA system is generally composed of various subsystems, each of which relates intimately with others. Therefore, if performances are issued, it is very difficult to find out the reason clearly. For example, if a business process in SOA system has longer response time than before, there could be various reasons: cache overflow in a business processor, wrapping overhead in service interface, or exceptions in computing resources, etc. One thing clear is that the performance of business process

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 39–46, 2011.

layer depends on the lower layer and measuring the performance of business layer includes indirectly measuring the performance of all the lower layers. But, most of the test frameworks developed focus on measuring SOA messaging performance, as we present in chapter 2. They almost adopt batch-style testing where all the test cases are executed in a sequence.

OMG recommended a standard SOA reference model, MDA (Model Driven Architecture) [1]. It is widely adopted in real world because it is normative and enables SOA system to be implemented in a business-centric approach. In the MDA, a business process is designed firstly in a way for satisfying business requirements and later services are bounded to the activities in the business process. Business processes are described in a standardized language (e.g. WSBPEL) and they are executed generally on a business process management (BPM) system.

For testing SOA systems implemented according to the MDA reference model in business-centric way, test harness should have business process simulation functionality so that it can behave as BPM and test overall performance at the same time. This means that the test harness can execute business process, perform tests, and gather metric values. The performance of the business process layer with which users keep in contact usually depends on the summation of the performance of the other lower layers. Therefore, for testing SOA system, test cases describing business process activities should be prepared.

In SOA system, test assertions may help develop tighter test cases which could be used as an input for SOA test harness. Any ambiguities, contradictions and statements which require excessive resources for testing can be noted as they become apparent during test assertion creation. Test assertions should be reviewed and approved to improve both the quality and time-to-deployment of the test cases. Therefore, best results are achieved when assertions are developed in parallel with the test cases.

Test assertions provide a starting point for writing conformance test cases or interoperability test cases. They simplify the distribution of the test development effort between different organizations while maintaining consistent test quality. Test assertions improve confidence in the resulting test and provide a basis for coverage analysis.

In section 2, we present some related works. Section 3 provides the concept of test assertion. In section 4, we describe a test assertion model. Section 5 presents cases of the test assertion and section 6 shows complex predicates of test assertions. Conclusions are presented in the last section.

2 Related Works

This section presents some test frameworks and script languages developed or proposed for SOA system.

Web Services Quality Management System

This system has been developed by NIA in order to measure Web services' quality on the criteria of WSQM (Web Services Quality Model) quality factors [2]:

interoperability, security, manageability, performance, business processing capability, and business process quality. This system contributes to consolidate the quality factors of SOA. However, it requires expanding its architecture to apply SOA, system, because it targets only to Web services system.

ebXML Test Framework

This framework has been implemented by NIST and KorBIT for testing ebXML system according to OASIS IIC Specification [3]. It could test packaging, security, reliability, and transport protocol of ebXML messaging system implemented by ebMS specification [4]. The main purpose of this framework is to test conformance and interoperability of ebXML messaging system, and it is not proper for testing service oriented systems. Besides, it cannot test ad hoc status resulting from various events, because it is not event-driven but batch-style test framework.

JXUnit and JXU

JXUnit [5] and JXU [6] is a general scripting system (XML based) for defining test suites and test cases aimed at general e-business application testing. Test steps are written as Java classes. There is neither built-in support for business process test nor support for the event-driven features. However, as a general test scripting platform that relies on a common programming language, this system could be used as an implementation platform for general e-business tests.

ATML (Automatic Test Mark-Up Language)

In its requirements, ATML provides XML Schemata and support information that allows the exchange of diagnostic information between conforming software components applications [7]. The overall goal is to support loosely coupled open architectures that permit the use of advanced diagnostic reasoning and analytical applications. The objective of ATML is focusing on the representation and transfer of test artifacts: diagnostics, test configuration, test description, instruments, etc.

Test Choreography Languages

These are standards for specifying the orchestration of business processes and/or transactional collaborations between partners. Although a markup like XPDL [8] is very complete from a process definition and control viewpoint, it is lacking the event-centric design and event correlation / querying capability required by testing and monitoring exchanges. Also, a design choice has been here to use a very restricted set of control primitives, easy to implement and validate, sufficient for test cases of modest size. Other languages or mark-ups define somehow choreographies of messages and their properties: ebBP, WS-BPEL, WS-Choreography[9]. The general focus of these dialects is either the operational aspect of driving business process or business transactions, and/or the contractual aspect, but not monitoring and validation. Although they may express detailed conformance requirements, they fall short of

covering the various aspects of an exhaustive conformance check, e.g. the generation of intentional errors or simulation of uncommon behaviors. In addition, the focus of these languages is mainly on one layer of the choreography – they for instance, ignore lower-level message exchanges entailed by quality of service concerns such as reliability, or binding patterns with the transport layer.

3 Concept of Test Assertion

A test assertion is a testable or measurable expression for evaluating the adherence of an implementation (or part of it) to a normative statement in a specification.

A set of test assertions may be associated with a conformance clause in order to define more precisely what conformance entails. Test assertions lie between the specification and any suite of tests to be conducted to determine conformance (See Fig. 1). Such a test suite is typically comprised of a set of test cases. These test cases may be derived from test assertions which address the normative statements of the specification.

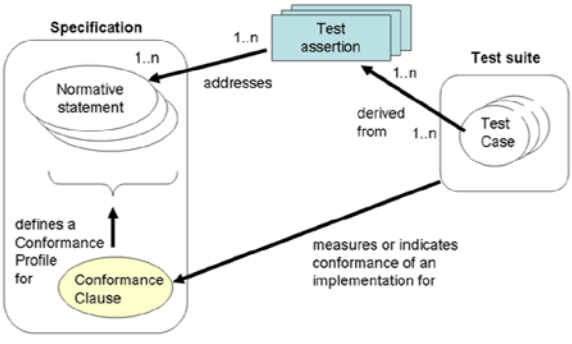


Fig. 1. Role of Test Assertion

Judging whether the test assertion is testable may require some knowledge about testing capabilities and resource constraints. Sometimes there is little knowledge of what actual testing conditions will be. In such cases the prime objective of writing test assertions is to provide a better understanding of what is expected from implementations, in order to fulfill the requirements. In other cases, the test assertions are designed to reflect a more precise knowledge of testing conditions. Such test assertions can more easily be used as a blueprint for test suites.

4 Test Assertion Model

This section aims to cover the simpler aspects of test assertions. Some more complex aspects are covered later in this section. Fig. 2 below shows the anatomy of a typical test assertion, and how its parts relate to the specification being addressed, as well as to the implementations under test. Some optional parts are not shown in the figure.

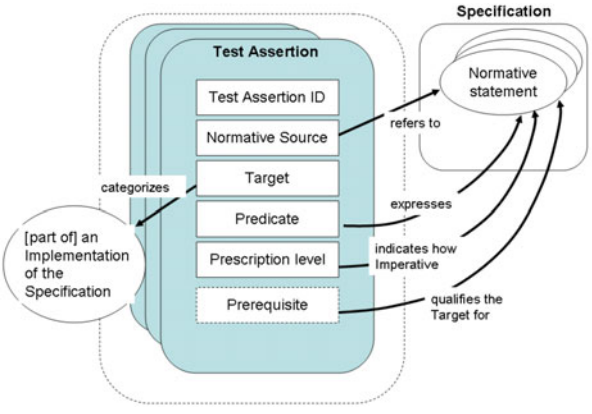


Fig. 2. General Anatomy of a Test Assertion

Some of the elements which comprise a test assertion are considered core while others are optional. A test assertion includes, implicitly or explicitly:

Identifier
A unique identifier of the test assertion facilitates tools development and the mapping of assertions to specification statements. It is recommended that the identifier be made universally unique.

Normative Sources
These refer to the precise specification requirements or normative statements that the test assertion addresses.

Target
The target categorizes an implementation or a part of an implementation of the referred specification, which is the main object of the test assertion and of its Normative Sources.

Predicate
A predicate asserts, in the form of an expression, the feature (a behavior or a property) described in the specification statement(s) referred by the Normative Sources. If the predicate is an expression which evaluates to "true" over the test assertion target, this means that the target exhibits this feature. "False" means the target does not exhibit this feature.

In addition, a test assertion may optionally include following components.

Description
This is an informal definition of the role of the test assertion with some optional details on some of its parts. This description must not alter the general meaning of the test assertion and its parts. This description may be used to annotate the test assertion with any information useful to its understanding. It does not need to be an exhaustive description of the test assertion.

Prescription Level
This is a keyword that indicates how imperative it is that the Normative Statement referred to in the Normative Source, be met. See possible keyword values in the Glossary.

Prerequisite
A test assertion Prerequisite is a logical expression (similar to a Predicate) which further qualifies the Target for undergoing the core test (expressed by the Predicate) that addresses the Normative Statement. It may include references to the outcome of other test assertions. If the Prerequisite evaluates to "false" then the Target instance is not qualified for evaluation by the Predicate.

Tags
Test assertions may be assigned 'tags' or 'keywords', which may in turn be given values. These tags provide you with an opportunity to categorize the test assertions. They enable you to group the test assertions; based on the type of test they assume or based on their target properties.

Variables
Test assertions may also include variables for convenience in storing values for reuse and shared use, as well as for parameterization.

As a test assertion has parts that can be evaluated over a Target instance (i.e. the Prerequisite and the Predicate), the following semantics apply to a test assertion:

- "Target not qualified": if the Prerequisite (if any) evaluates to "false" over a Target instance.
- "Normative statement fulfilled [by the Target]": if the Prerequisite (if any) evaluates to "true" over a Target instance, and the Predicate evaluates to "true".
- "Normative statement not fulfilled [by the Target]": if the Prerequisite (if any) evaluates to "true" over a Target instance, and the Predicate evaluates to "false".

5 Case Study of Test Assertion

Consider the following statement in the widget specification:

[requirement 101] “A widget of medium size MUST use exactly one AA battery encased in a battery holder.”

There are actually two requirements here that can be tested separately:

(requirement 101, part 1) A medium-size widget MUST use exactly one AA battery.
(requirement 101, part 2) A medium-size widget MUST have a battery holder encasing the battery.

Because of this it is possible to write two test assertions:

- **TA id:** widget-TA101-1a
  Normative Source: specification requirement 101, part 1 Target: medium-size widget

Predicate: [the widget] uses exactly one AA battery. Prescription Level: mandatory

and

– **TA id:** widget-TA101-1b
Normative Source: specification requirement 101, part 2 Target: medium-size widget
Predicate: [the widget] has a battery holder encasing the battery. Prescription Level: mandatory

The granularity of a test assertion is a matter of judgment. A single test assertion instead of two can be written here, with the predicate: "[the widget] uses exactly one AA battery AND has a battery holder, encasing the battery".

This choice may later have an impact on the outcome of a test suite written to verify the conformance of widgets. With a single test assertion, a test case derived from this test assertion will not be expected to distinguish between the two failure cases. Using two test assertions - one for each sub-requirement - will ensure that a test suite can assess and report independently about the fulfillment of each sub-requirement. Other considerations such as the different nature of tests implied or the reuse of a test assertion in different conformance profiles [VAR], may also lead to the adoption of "fine-grained" instead of "coarse-grained" test assertions. Usage considerations will dictate the best choice.

6 Complex Predicates

Recall the previous example of [requirement 101]. The target can be defined as "a medium-size widget" or as just "a widget". The latter is a natural decision if the specification requirement uses the wording: "[requirement 101] If a widget is medium size, then it MUST use exactly one AA battery and be encased in a battery holder." For the simplicity of this example, if the two test assertion predicates for widget-TA101-1a and widget-TA101-1b are combined into one example, one possible outcome is:

TA id: widget-TA101-2a
Normative Source: requirement 101 Target: widget
Predicate: if [the widget] is medium-size, then [the widget] uses exactly one AA battery AND the battery is encased in a battery holder.
Prescription Level: mandatory

The target category is broad, but the predicate part is really of interest only for a subset of this category (the medium-size widgets). Usage considerations should again drive the decision here: a test suite that is designed to verify all widgets, and does not assume a prior categorization of these into small / medium / large sizes, would be improved with test assertions that only use "widget" as the target, such as widget-TA101-2a.

A test assertion predicate may, then, be a Boolean expression - a composition of atomic predicates using logical operators AND, OR, NOT. A test assertion predicate may also be of the kind: "if (condition) then (expression)".

The predicate is worded in an abstract way, still close to the wording of the specification. No indication of what kind of test procedure will be used, such as how to determine the number and type of batteries, is given. Detailed criteria for the condition evaluation, such as what kind of battery holder is acceptable, are also not provided. These details are normally left to the test cases that can be derived from the test assertions. These test cases will determine the precise criteria for conforming to the specification. However, if a precise criterion for interpreting the battery holder requirement is provided in an external specification, either referred to directly by the widget specification or by a related conformance clause, then a test assertion must use this criterion in its predicate. Such a test assertion must then refer not only to the specification requirement in its reference property, but also to the external specification or to the conformance clause that refers to this specification.

Another case where a predicate is more complex is when its conditional expression involves more than one part of an implementation (or implementations). In some cases it is clear which one of these objects must be considered the target, while others are just accessory objects. Consider the following predicate: "the [widget price tag] matches the price assigned to the widget in its [catalog entry]", where price tags and catalog entries are both items that must follow the store policy (in effect the specification). In this case it may be reasonably assumed that the "catalog" content is authoritative over the price tag. The price tag can then be considered as the test target, while the accessory object may be identified by a variable which is then used in the predicate.

7 Conclusion

We presented a SOA test assertion model, which facilitates to make test cases in normalized form. In the model, we devised the concept of prescription level and normalized prerequisite for preparing test cases. By the concepts, test assertion can be transformed into test cases without semantic distortion. The model minimizes the human intervention in preparing test cases by automating some processes for translating test assertions into test cases. We showed two cases of complex predicates. Further studies are required to develop a test framework to check out automatically that test cases are conformed to test assertions.

References

1. Miller, J., Mukerji, J.: MDA Guide Version 1.0.1 OMG (June 2003), http://www.omg.org/docs/omg/03-06-01.pdf
2. Lee, Y., et al.: Web Services Quality Model 1.1. OASIS WSQM TC (October 2008)
3. Durand, J., et al.: ebXML Test Framework v1.0. OASIS IIC TC (October 2004)
4. Wenzel, P., et al.: ebXML Messaging Services 3.0. OASIS ebMS TC (July 2007)
5. Java XML Unit (JXUnit), http://jxunit.sourceforge.net
6. JUnit, Java for Unit Test, http://junit.sourceforge.net
7. ATML, Standard for Automatic Test Markup Language (ATML) for Exchanging Automatic Test Equipment and Test Information via XML, IEEE (December 2006)
8. XPDL: XML Process Definition Language (Workflow Management Coalition) Document Number WFMC-TC-1025: Version 1.14, October 3 (2005)
9. OASIS, Business Process Specification Schema 1.0.1 (May 2001); and ebBP, v2.0.4 (October 2006)

Multithreaded Power Consumption Scheduler Based on a Genetic Algorithm*

Junghoon Lee, Gyung-Leen Park, and Hye-Jin Kim

Dept. of Computer Science and Statistics
Jeju National University, 690-756, Jeju-Do, Republic of Korea
{jhlee,glpark,hjkim82}@jejunu.ac.kr

Abstract. This paper presents a multithreaded power consumption scheduler and measures its performance, aiming at reducing peak load in a scheduling unit. Based on the observation that the same genetic algorithm leads to a different solution for a different initial population, the proposed scheduler makes each thread generate its own initial population and independently run genetic iterations for a better solution. Judging from the performance measurement result obtained from a prototype implementation, multithreaded version can reduce the peak load even with small population size without loss of accuracy. After all, the threaded scheduler improves the computation speed, which is inherently dependent on the population size of a genetic scheduler mainly consist of sorting and selection procedures.

Keywords: Power consumption scheduler, multithreaded version, initial population, execution time, peak reduction.

1 Introduction

The smart grid ensures reliability, reduces energy consumption, and minimizes environmental impact in the electric power delivery system taking advantage of advanced information technologies including sensors, communications, and computational ability [1]. As a great deal of heterogeneous components are involved in the power system, an efficient interaction mechanism between them is the most fundamental building block, while ever-growing communication and distributed computing techniques can provide solutions to such a requirement. Hence, energy providers and consumers can exchange necessary information such as price change, current power load, and reservation status, while the consumer can even select a preferred supplier dynamically. After all, power consumption can be smarter using the smart grid communication framework.

From the consumer's side, DSM (Demand Side Management) functions embrace the whole demand side activities such as device operation planning,

* This research was supported by the MKE, Republic of Korea, under IT/SW Creative research program supervised by the NIPA (NIPA-2011-(C1820-1101-0002)) and also through the project of Region technical renovation.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 47–52, 2011.

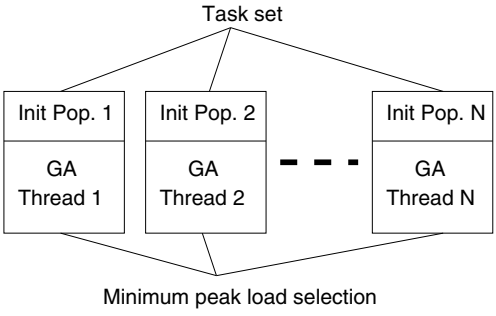


Fig. 1. Threaded computing for the power consumption scheduler

evaluation, implementation, and monitoring [2]. Particularly, in DSM, power consumption scheduling can achieve cost saving, energy efficiency, and peak load reduction. Peak load reduction is important, as pricing policies are highly likely to decide the power rate according to the amount of power consumption for each time unit. A power consumption schedule decides when to initiate, sometimes suspend and resume a specific electric device operation, or a task, for the given task set. Scheduling is a kind of time-intensive optimization problems greatly sensitive to the number of tasks. Moreover, if scheduling is conducted on an IHD (In-Home Display) equipped with a low-speed CPU, it can take too much time to generate a schedule.

Practically, instead of an optimal schedule, suboptimal schedules are desirable if they can be generated within an acceptable delay bound. Genetic algorithms are good candidates for this purpose. Based on principles of natural selection and genetics, each evolutionary step generates a population of candidate solutions and evaluates the population according to a fitness function to select the best solution and mate to form the next generation [3]. At each evolution, bad traits are eliminated while the good traits survive and are combined with other good traits to make better candidates. Its performance greatly depends on how to select initial populations. In the mean time, threaded execution is very popular even in personal computers and hand-held devices. It can benefit from parallel processing on the multiprocessor architecture, significantly enhancing the computing time [4].

In this regard, this paper is to present a threaded power consumption scheduler built upon a genetic algorithm and measure its performance. In genetic algorithms, different initial populations, namely, the set of feasible solutions, lead to different solutions, irrespective of whether a guideline for initial population selection is available or not. If we employ a large population to attempt a variety of initial populations, the execution time can grow too much, as the genetic iteration includes sorting and several selection steps. Hence, it will be advantageous to initialize populations differently, allocate populations to each thread, and make respective threads run the genetic iteration independently. Then, the coordinator thread collects the respective results to select the best solution as shown in Figure 1.

2 Scheduler Design Based on Genetic Algorithms

Task T_i can be modeled with the tuple of $<F_i, A_i, D_i, U_i>$. F_i indicates whether T_i is preemptive or nonpreemtive [5]. A_i is the activation time of T_i, D_i is the deadline, and U_i denotes the operation length. The load power profile characterizes the power consumption behavior of each appliance, while the power consumption pattern for each electric device is aligned to the fixed-size time slot. The length of a time slot can be tuned according to the system requirement on the schedule granularity and the computing time. In power schedule, the slot length can be tens of minutes, for example, 20 minutes. From now on, N denotes the number of tasks and M denotes the number of slots in the scheduling window.

Genetic algorithms have been successfully applied to find acceptable solutions to problems in business, engineering, and science within a reasonable amount of time [3]. It begins with an initial population set and iteratively runs selection and reproduction steps. The population contains a set of feasible schedules and the quality of those schedules gets better iteration by iteration. First of all, the schedule must be encoded to apply genetic operations. Here, a chromosome corresponds to a single feasible schedule, and is represented by a fixed-length string of integer-valued vector [6]. A value element denotes the start time for nonpreemptive tasks. As they cannot be suspended once they have begun, just the start time is enough to describe their operation plan. Here, if the consumption profile of a task is (3, 4, 5, 2), and the vector element is 2, the allocation for this task will be (0, 0, 3, 4, 5, 2, 0, 0,...). This allocation representation consists of M numbers, each of which is bound to a slot.

As contrast, for a preemptive task, possible slot allocations are generated in advance. Suppose that A_i is 2, D_i is 7, and U_i is 3. In this example, $D_i - A_i$ is 5, so the number of possible allocations is ${}_5C_3$. That is, 3 out of 5 slots must be selected for the task operation. Some examples of feasible maps can be (0, 0, 1, 1, 1), (0, 1, 0, 0, 1), and (1, 1, 1, 0, 0), where 1 means the task operation is assigned. If the mapping vector is (0, 1, 0, 1, 1), each profile entry is mapped to the position having 1 one by one from the start time, namely, 2. Hence, the allocation will be (0, 0, 2, 3, 0, 4, 0, ...). Each allocation vector is taken as a binary number and its decimal equivalence is used for a gene. The allocation vector can be transformed to the allocation table which has N rows and M columns to assess the quality of a schedule. For each allocation, the scheduler can calculate per-slot power requirement and peak load.

The iteration consists of selection and reproduction. Selection is a method that picks parents according to the fitness function. The Roulette Wheel selection gives more chances to genes having better fitness values for mating. Reproduction, or crossover, is the process taking two parents and producing a child with the hope that the child will be a better solution. This operation randomly selects a pair of two crossover points and swaps the substrings from each parent. Reproduction may generate the same gene with the existing ones in the

population. It is meaningless to have multiple instances of a single schedule. So, they will be replaced by new random chromosome. Additionally, mutation exchanges two elements in a gene. In our scheme, the meaning of the value is different for preemptive and nonpreemtive tasks. Hence, the mutation must be prohibited. It is true that the appliance scheduler is subject to time constraint. However, this constraint can be always met, as the scheduler selects the start time only within the valid range and the precalculated combination index.

3 Performance Measurement

This section implements the multithreaded power consumption scheduler based on genetic algorithms with Visual C++ 6.0, making it run on the platform equipped with Intel Core2 Duo CPU, 3.0 GB memory, and Windows Vista operating system. For a task, the start time is selected randomly between 0 and M, while the operation length is also selected randomly lest the task operation should get out of the scheduling window. All tasks have the common deadline, namely, M. In addition, the power level for each time slot has the value of 1 through 5. This can be adjusted according to the target scheduling group. The experiment focuses on the effect of the number of tasks, population size, and iterations to the peak load according to the number of threads. For each experiment parameter setting, 20 task sets are generated and their results are averaged.

The first experiment measures the effect of the number of tasks to the peak load when the number of threads ranges from 1 to 10. This experiment runs 1,000 iterations, with the population size set to 50. Figure 2 plots two cases when the number of tasks is 5 and 10, respectively. Among these, 2 are preemptive. Preemptive tasks can be suspended and resumed during its operation even after they began their operations, so they can be flexibly placed in the slots having low load. As each task has the same average power consumption, more tasks will have larger peak load. But the peak load is not exactly proportional to the number of tasks as the scheduler reshapes the load. Figure 2 shows that the peak

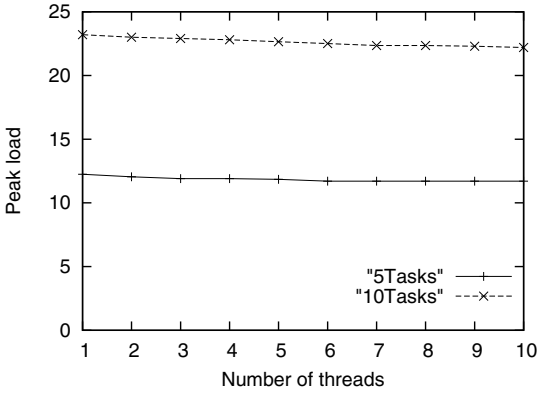


Fig. 2. Peak load according to the number of tasks

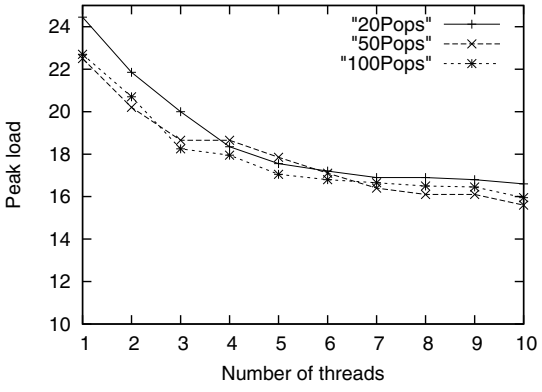


Fig. 3. Peak load according to population size

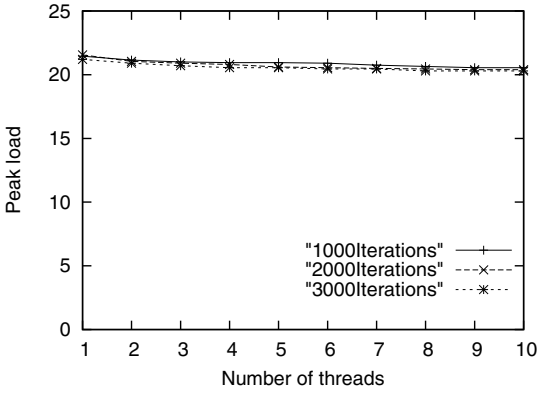


Fig. 4. Peak load according to iterations

load can be reduced as the number of threads increases for both cases. The peak load reduction is around 4.5 % when the number of threads is 10.

The second experiment measures the effect of the population size to the peak load when the number of threads ranges from 1 to 10. This experiment runs 1,000 iterations, with the number of tasks set to 10. Here again, 2 tasks are preemptive. Figure 3 plots the cases when the population size is 20, 50, and 100, respectively. This graph shows the threaded execution can significantly contribute to reducing the peak load especially when the population size is 20. Even if the small population improves the computation speed, it suffers from accuracy loss. However, multiple threads can overcome this problem, indicating that threaded computing can achieve both computation speed and accuracy.

The third experiment measures the effect of the number of iterations to the peak load when the number of threads ranges from 1 to 10. This experiment sets the population size to 50 and the number of tasks to 10 including 2 preemptive tasks. Figure 4 plots the cases when the number of iterations is 1,000, 2,000, and 3,000, respectively. Actually, even if the case of 3,000 iterations shows a

little bit better peak load, the reduction is not so significant. It takes less than a second to run 1,000 iterations in our prototype platform which corresponds to an average-performance computer. After all, for power scheduling, multiple threads can work very efficiently just with small size population and just 1,000 iterations. Essentially, the number of iterations is a tunable parameter according to the number of tasks and the power load profile.

4 Concluding Remarks

This paper has presented a multithreaded power consumption scheduler and measured its performance, aiming at reducing peak load in a scheduling unit and meeting time constraint of each electric operation. Based on the observation that a genetic algorithm leads to a different solution according to a different initial population, our design makes each thread generate its own initial population and independently run genetic algorithm loops, and then a coordinator thread collects the result to find the best one. From the performance measurement result obtained from the prototype implementation, multithreaded version can reduce the peak load even with small population size. After all, the threaded scheduler cannot just reduce the peak load by around 4.5 %, but also achieve both computation speed and accuracy.

References

1. Gellings, C.: The Smart Grid: Enabling Energy Efficiency and Demand Response. The Fairmont Press (2009)
2. Bonneville, E., Rialhe, A.: Demand side Management for Residential and Commercial End-Users (2006), http://www.leonardo-energy.org/Files/DSM-commerce.pdf
3. Katsigiannis, Y., Georgilakis, P., Karapidakis, E.: Multiobjective Genetic Algorithm Solution to the Optimum Economic and Environmental Performance Problem of Small Autonomous Hybrid Power Systems with Renewables. In: IET Renewable Power Generation, pp. 404–419 (2010)
4. Kang, M., Kang, D., Crago, S., Park, G., Lee, J.: Design and Development of a Run-time Monitor for Multi-Core Architectures in Cloud Computing. Sensors 11, 3595–3610 (2011)
5. Derin, O., Ferrante, A.: Scheduling Energy Consumption with Local Renewable Micro-Generation and Dynamic Electricity Prices. In: First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods, and Tools (2010)
6. Lee, J., Park, G.-L., Kwak, H.-Y., Jeon, H.: Design of an Energy Consumption Scheduler Based on Genetic Algorithms in the Smart Grid. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 438–447. Springer, Heidelberg (2011)

Design of a Composite Sensor Node in Agricultural Ubiquitous Sensor Networks*

Junghoon Lee[1], Gyung-Leen Park[1], Hye-Jin Kim[1], Ho-Young Kwak[2], Seongjun Lee[3], Jong-Heon Lee[4], Bong-Soo Kang[4], and Yun-Hyuk Kim[4]

[1] Dept. of Computer Science and Statistics,
[2] Dept. of Computer Engineering,
Jeju National University, 690-756, Jeju-Do, Republic of Korea
[3] EZ Information Technology
[4] InToBe Inc.

Abstract. This paper first designs and implements a composite sensor node capable of reliably collecting sensor data mainly and delivering control actions to actuators in the agricultural sensor network, and then presents the corresponding ubiquitous sensor network architecture that integrates an intelligent information technology. Consist of a main control unit, a sensor interface base board, an actuator interface & control board, a global network interface module, and a sensor signal acquisition board, the composite node extends the data transmission range, regulates power according to the device operation type, exploits lightweight 6LowPAN and IPv6 protocol, and makes it possible to share pin connections between a group of sensor devices. Our design can contribute to building an efficient sensor data processing framework which can seamlessly combine XML specification, an expert system, a filtering engine, and sensor query lifetime management.

Keywords: Ubiquitous sensor network, composite sensor node, power supply, reliable data exchange, sensor network gateway.

1 Introduction

The main task of ubiquitous sensor networks, or USN in short, is monitoring sensor values, deciding the control actions, and triggering appropriate actuators [1]. To this end, a lot of sensors are installed over the wide target area, while the sensor network must handle a great volume of sensor data, detect events, and compute the real-time control action, regardless of sensor applications. Unfortunately, the sensor network generally lacks the computing power, so a sophisticated data analysis and intelligent control logic can be performed at the remote high-performance computing server. This problem highlights the importance of an efficient and robust data transmission mechanism between the sensor network

* This research was supported by the MKE (The Ministry of Knowledge Economy), through the project of Region technical renovation, Republic of Korea.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 53–58, 2011.

and the computing server. The transmission mechanism must consider also a reliable power supply to communication nodes, packet conversion, signal sharing, and typical gateway functions.

The goal of USNs is different for each application, as the sensor device and the corresponding data analysis scheme will be different. More importantly, according to the required reliability level, the system design selects appropriate sensors, necessary actuators, and communication protocols. In farms, livestock farms, and fisheries, it is necessary to continuously monitor the environmental change in temperature, humidity, lightness, CO_2 and NH_2 levels in addition to the biosensors which captures the disease of a livestock. With this data, we can promptly detect the malicious condition and send control command to actuate fans, pumps, heaters, and the like. Also in this scenario, the reliable data exchange is the most critical requirement. In this regard this paper is to design a composite sensor node capable of exchanging a great volume data between two domains, namely sensors and servers, based on IP-USNs, which can possibly incorporate many intelligent information technologies such as XML specification, middleware, filtering engine, and sensor query lifetime management.

2 Background and Related Work

Under the research and technical project named *Development of convergence techniques for agriculture, fisheries, and livestock industries based on the ubiquitous sensor networks*, our project team has designed and developed an intelligent USN framework [2]. This framework provides an efficient and seamless runtime environment for a variety of monitor-and-control applications on sensor networks. Much focus is put on how to handle the sensor data stream as well as the interoperability between the low-level sensor data and application clients. Systematic middleware coordinates the interaction between the application layer and low-level sensors, for the sake of analyzing a great volume of sensor data by filtering and integrating to create value-added context information. Then, an agent-based architecture is proposed for real-time data distribution to forward a specific event to the appropriate application, which is registered in the directory service via the open interface.

On this framework, we have designed an efficient cyclic routing scheme for network management messages, aiming at improving productivity and profit in those industries [3]. Instead of multiple point-to-point message transmissions for status indication collection, the management message traverses the specific set of nodes of interest one by one. The management application collects routing information from the neighbor table commonly available in the current sensor protocol to calculate the communication cost between each pair of target nodes. Based on this topology view, a genetic algorithm solver decides the traversal sequence of a cyclic management path. The experiment result discovers that the multithreaded version, in which each thread runs its own initial population, can find a much better solution, efficiently escaping local traps.

For the sake of efficiently and accurately detecting the specific events out of the great amount of sensor data which may include not just erroneous terms but also correlative attributes, the middleware module embeds an empirical event patterns and knowledge description [4]. For the filtered data, data mining module opens an interface to define the relationship between the environmental aspect and facility control equipments, set the control action trigger condition, and integrate new event detection logic. Finally, the remote user interface for monitoring and control is implemented by on Microsoft Windows, Web, and mobile device applications.

3 USN Node Design

Figure 1 illustrates the USN node architecture we develop in our sensor network framework. To begin with, the MCU, or main control unit, has interfaces to the wireless sensor network (WSN) and the sensor interface base board, or base board in short from now on. To the WSN side, the Sensornet communication protocol is employed for message exchange via multi-hop mesh networking. An RF power amplifier extends the transmission range of USN nodes up to 500 m. To the base board side, three connections are implemented, namely, a digital input-output connection, a universal asynchronous receiver/transmitter, and an analog-digital interface. The MCU can selectively connect to a specific device via the base board to collect sensor data and give a command to an actuator. This framework can easily integrate an additional component for better efficiency. As an example, for the set of given control actions, an energy consumption scheduler can manage their operations to save power, reduce cost, and so on [5].

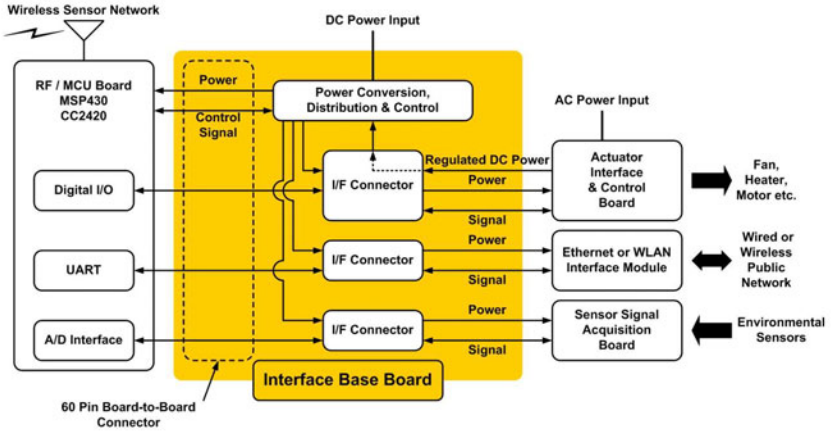


Fig. 1. Sensor interface base board

The base board plays a key role in our IP-USN integrated remote control module, intermediating between MCU and other heterogeneous modules including an

actuator interface & control board, an Ethernet or WLAN interface module, and a sensor signal acquisition board. The base board converts the voltage level for module driving, distributes the power, and controls power sources. To interface with actuators mainly working on AC power, it regulates DC power, and more importantly, it can give control commands to fans, heaters, and motors. Next, the base board provides power to the Ethernet and WLAN interface module to connect to the IP-based public network, working as an embedded USN gateway. It is the connection to a sensor signal acquisition board, or a sensor board in short that constitutes a main data stream path. A base board provides power to sensor boards for individual sensor activation and allocates signal lines for sensor signal acquisition. Hence, the sensor board can work without its own power input as long as it is connected to a base board.

The proposed sensor board design is aiming at the reduction of power consumption and the minimization of the number of signal lines by sharing them between as many sensor devices as possible. First, for power consumption reduction, the power provided from the base board is distributed by a demultiplexor and a load switch to make only one sensor part active at each time instant. Second, due to the limitation in the number of connector pins, the number of signal lines is a restricted resource. In our implementation, a single signal line can collect sensor signal from up to 8 sensors, while a sensor board can accommodate up to 8 sensor boards. After all, a base board can cover up to 64 sensor devices. Such signal line sharing is made possible taking advantage of the combination of a multiplexor and a low resistance analog switch for sensor device control as well as the combination of a multiplexor and an analog switch for signal line allocation.

The USN node can easily integrate an actuator board for better flexibility in a variety of sensor applications as shown in Figure 2. The actuator board, embedding its own MCU, can autonomously control the board operation, even in the place where neither network connection nor remote control is available. Moreover, the MCU can compensate for the limitation in the number of pin connections to the sensor interface board by alternating a group of signal transmissions for better reliability and diversity of control actions. After all, the actuator board consists of a connector part to the sensor interface board, a control signal level converter, a signal isolator to separate lower power DC signal from high power AC signal, already-mentioned embedded MCU circuit, and an AC-DC power converter to provide power to the whole remote control module. Noticeably, it is not necessary to install an additional DC power supply module even for an AC power control application.

4 USN Architecture

With the USN node described in the previous section, our IP-USN can be configured as shown in Figure 3. The wireless sensor network is built using the legacy IEEE 802.15.4 communication technology. The sensor data and control actions can be delivered through USN gateway. Each sensor node installs TinyOS-2.X

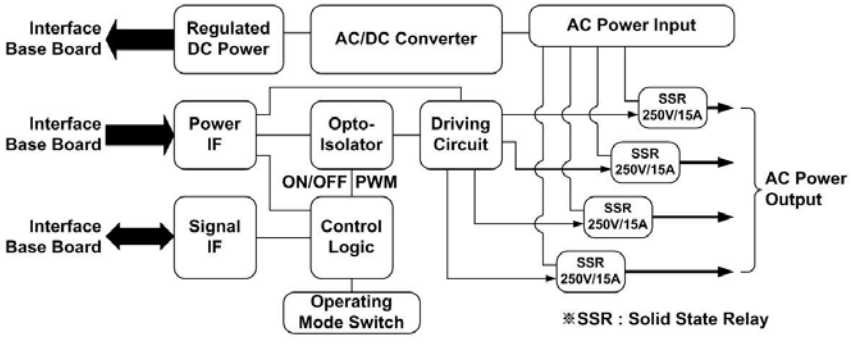


Fig. 2. Actuator control board

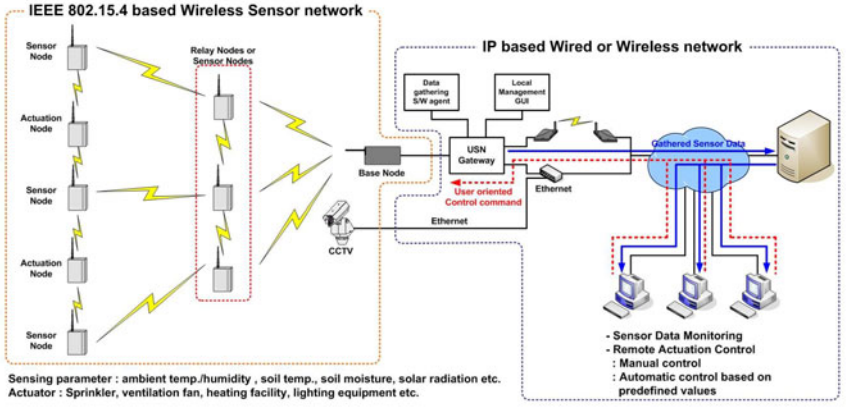


Fig. 3. Proposed ubiquitous sensor network architecture

[6], exchanges message via the Sensornet protocol [7], and is placed in farms, livestock farms, and fisheries. For the gateway connection, the lightweight 6LowPAN and IPv6 protocol stack is exploited. This protocol stack implements USN-to-IPv6 tunneling to convert 6LowPAN packets received from the UART interface to IPv6 packets. Beyond the gateway, it totally belongs to the global Internet domain where the sensor data can be analyzed with a sophisticated inference engine while the control logic can be built by an intelligent agent [8].

5 Concluding Remarks

This paper has first designed and implemented a composite sensor node capable of reliably collecting sensor data and delivering control actions to actuators, mainly for agricultural farms, livestock farms, and fisheries. Consist of a main control unit, a sensor interface base board, an actuator interface & control board, a global network interface module, and a sensor signal acquisition board, the

composite node extends the data transmission range, regulates power according to the device operation type, exploits lightweight 6LowPAN and IPv6 protocol, and makes it possible to share a pin connection between a group of sensor devices. Based on the composite node implementation, this paper also has presented a ubiquitous sensor network architecture that can seamlessly integrate high-end information technologies such as XML specification, an expert system, a filtering engine, and sensor query lifetime management to build an efficient sensor data processing framework.

As future work, we are planning to design an efficient power consumption scheduler for a series of control actions generated from the control logic, as the power issue in the sensor network is becoming more and more important. This scheduler can reduce the power consumption or reshape the peak load, while meeting the time constraint of each control action.

References

1. Culler, D., Estrin, D., Srivastava, M.: Overview of Sensor Networks. IEEE Computer 37, 41–49 (2004)
2. Lee, J., Park, G., Kwak, H., Kim, C.: Efficient and Extensible Data Processing Framework in Ubiquitous Sensor Networks. In: International Conference on Intelligent Control Systems Engineering, pp. 324–327 (2011)
3. Lee, J., Park, G., Kim, H., Kim, C., Kwak, H., Lee, S., Lee, S.: Intelligent Management Message Routing in Ubiquitous Sensor Networks. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 537–545. Springer, Heidelberg (2011)
4. Lee, J., Kim, H.-J., Park, G.-L., Kwak, H.-Y., Kim, C.M.: Intelligent Ubiquitous Sensor Network for Agricultural and Livestock Farms. In: Xiang, Y., Cuzzocrea, A., Hobbs, M., Zhou, W. (eds.) ICA3PP 2011, Part II. LNCS, vol. 7017, pp. 196–204. Springer, Heidelberg (2011)
5. Derin, O., Ferrante, A.: Scheduling energy consumption with local renewable microgeneration and dynamic electricity prices. In: First Workshop on Green and Smart Embedded System Technology: Infrastructures, Methods, and Tools (2010)
6. http://www.tinyos.net
7. Cuevas, A., Urueña, M., Laube, A., Gomez, L.: LWESP: Light-Weight Exterior Sensornet Protocol. In: IEEE International Conference on Computer and Communications (2009)
8. Esposito, F., Basile, T.M.A., Di Mauro, N., Ferilli, S.: A Relational Approach to Sensor Network Data Mining. In: Soro, A., Vargiu, E., Armano, G., Paddeu, G. (eds.) Information Retrieval and Mining in Distributed Environments. SCI, vol. 324, pp. 163–181. Springer, Heidelberg (2010)

Millimetric Waves Technologies: Opportunities and Challenges

Jahangir Dadkhah Chimeh and Saeed Bashirzadeh Parapari

Research Institute for ICT, Tehran, Iran

Abstract. Providing an available wideband and better antenna beam forming are two good profits of millimeter wave (mmWave) technology. MmWave technology makes radio systems lighter and smaller and radars more precise. Today, commercial MmWave equipment work below 90GHz frequencies. MmWave radios work to transport Internet traffic in the backhaul of communication networks. There is a challenge in mmWave technology since the prices of equipment increases as the frequency increases. In this paper we study the applications of mmWave technology, its products, standards and compare it with other wireless technologies.

Keywords: millimeter wave, products, standards.

1 Introduction

MmWave technology was initially used for aviation support systems but quickly became the foundation for other innovative solutions including security imaging applications and wireless communications systems. Narrow band services are offered within 2GHz and lower bands and include voice and low data rate services. Today, we need a wide bandwidth link in the backhaul to access the Internet. This link can be exploited in the mmWave frequencies. To access to a very high bandwidth link we need a very high frequency band, e.g., 5 Gb/s bandwidth can be provided in 60GHz frequency. On the other hand, mmWave frequencies can also provide special properties in the indoor and outdoor communications, e.g., signals are attenuated while encountering to space particles, trees and walls and gas absorption [1, 2]. In addition to free space loss, the air's oxygen attenuates the mmWaves 15dBm/Km in 60 GHz frequency.

Due to the high noise level in the wide bandwidth mmWaves and also high propagation loss, the mmWave signals travel in the short range. This, results in small cells which provides frequency reuse pattern capability in the system. Antenna sizes are reduced in the mmWave bands. Besides, in the mmWave band, the antenna directivity increases which makes the multipath interference to decrease. In addition, due to the very small wavelength we may implement antennas better. Besides, we may use small MMICs in the system.

On the other hand, the mmWave may be used in the point-to-point (P-P) and point-to multipoint (P-mP) communications in the short range relative to microwave bands.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 59–66, 2011.

MmWaves can also provide fiber optic and satellite link connections to the land subscribers. Therefore we pay attention to mmWave application in section 2, then compare the mmWave technology with other technologies in section 3. In section 4 we survey two company's products and in section 5 we study the mmWave standards. In the end we draw the conclusion.

2 MmWave Applications

Here we study the important applications of mmWave technology in the four following main categories [3].

• Teaching Applications
Elementary and high schools, colleges and universities need high capacity links. They can use wireless mmWave links instead of wired links. We should note that most teaching centers are in distinct campus buildings and their wired connections are with low speed and high cost.

• Disaster / Weapon Recovery Solutions
Today, the networks carry more critical data than ever before and sometimes human's life / death and success / failure depend on these data. MmWave solutions are often used as alternative and redundant paths to traditional cable and leased lines between two locations. Besides, to counter terrorism weapon, recovery solutions can be established by this technology.

• Healthcare Solutions
MmWave radio productions are a solution for hospital and medical groups needing real-time access to medical imaging and patient records. By this technology, bandwidth-intensive patient information such as x-rays and digital images may be handled. Transmitting high resolution MRI images or x-rays, for example, over slow T1/E1 connections can take several minutes which is unacceptable when dealing with urgent patient diagnosis and treatment.

• Financial/Banking Solutions
MmWave radio products are ideal for financial institutions such as bank i.e., securely transmitting classified data between locations. Instead of using traditional leased lines or low-security unlicensed radio frequency solutions to connect remote locations, financial companies can now link via a high capacity, high secure wireless point-to-point solutions.

• Transportation Solutions
MmWave radio products are ideal for transporting and accessing data and video surveillance images in transportation environment. These links can connect cameras in remote locations or buildings to centralized monitoring centers.

• Video Surveillance Solutions
Whether to track remote locations, monitor suspicious activity or scan for terrorist threats, the use of mmWave video technology has been a key to ensuring the safety and security of homes, businesses and other sites.

• Military Solutions
Since this technology can easily and fast be deployed and redeployed and also has a very wide bandwidth it is applicable to military aspects like high definition real-time supervisory operations in the battlefields.

Besides, the following four categories are introduced in mmWave products:

• Point-to-Point Products
These products are used for straight (line of sight) communications. Those products are used for data communication between two or more distinct networks in the backhaul of the cellular networks and so on. These products include antennas, outdoor and indoor equipment. Outdoor equipment also includes transceiver and antennas and indoor equipment has an interface to the wired network.

• Radar Products
These products are used for tracking, fire control, level detection, detection of hidden weapons, etc. in the range of 18-140 GHz. Their output power is between 50-100 mw and their range resolution is about 0.5 to 1.5 meter.

• Wireless Personal Communication Products
These products are used for small distances with high data rates. The following networks use these products:

- Ad-hoc networks which are used for headphone, keyboard and medical sensors, …and are used for below 10m distances.
- WLAN with new 60Ghz carrier frequency This network is used for communication between laptops and data networks ranging more than 10m.

• Medical Products
These products are extremely high-frequency therapy equipment, through direct exposure to extremely high-frequency electromagnetic wave on the lesion site. These products may also improve blood circulation in the body [4].

• Security Products
Now, more than 70 airports in USA and many airports in UK are equipped with mmWave body scanner in order to find cold and warm weapons.

3 mmWave Technologies Comparison

Here, we compare the properties of mmWave technologies with other wireless technologies. Fig. 1depicts bit rates of different technologies versus coverage areas [4].

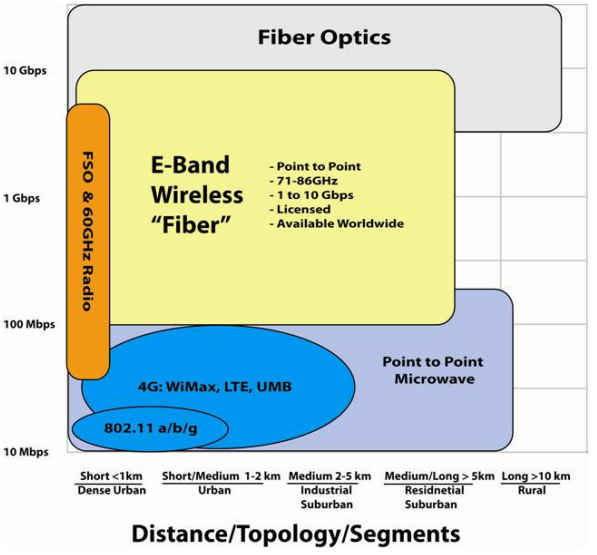


Fig. 1. mmWave bit rate versus coverage area [4]

As it is shown mmWave has 10Gbps bit rate and up to 5 km coverage area. Besides, Table 1 shows different parameters of some different technologies.

4 MmWave Products

To introduce mmWave equipment, here, we selected randomly Gigabeam and Loea vendors and describe two products of them.

4.1 Gigabeam Company [5]

This company has different mmWave equipment. One of them is Gi-core product which transmits Ethernet information within 70/80 GHz with 1.25 Gb/s bandwidth. Besides, GI-FELEX product can transmit 300Mb/s in the 4-60 GHz frequencies. Gigabeam's products are applicable for trade centers, backhaul, metro Ethernet and wideband applications out of the cities. For example Gi-core has the following specifications:

- Full duplex
- SNMP MP for device management
- SSH and Radius security protocol
- Low latency
- 188dB gain
- +21 dB output power
- Threshold -65dBm
- Standard EN302 217
- Parabolic 60 cm antenna

Table 1. Comparison of wireless technologies

	WiFi	3G/4G	Microwave	60 GHz	FSO	Fiber	E-band
Data rate	Variable, typically 1 Mbps	Variable, typically 10 Mbps	2 to 311 Mbps	100 Mbps to 1 Gbps	100 Mbps to 1 Gbps	To 100 Gbps	100Mbps to 3 Gbps today, to 10 Mbps in the future
Typical distance for carrier class performance	20 Yards	2 miles	5 miles	500 Yards	200 Yards	20-60 km	1-3 miles
Spectrum availability and licensing	Freely available for unlicensed use	Spectrum very scarce	Usually available for are licensing from country regulators	Varies country by country	Spectrum freely available as technology not regulated	N/A	Available worldwide
Guaranteed interference and regulatory protection	No	Usually	Yes	No	No	Yes	Yes
Relative cost of ownership	Low	Medium	Medium	Medium	Medium	High	Medium
Install and commissioning time	Hours	Months / Years	Weeks/Months	Hours/days	Hours/days	Months / Years	Hours/days

As it is shown transmission bit rate in the mmWave technology is greater than the other technologies and in addition, installation and commissioning times are less than the others.

4.2 Loea Products [6]

This company offers point-to-point products in the 70/80 band. These products are confirmed by NTIA, FCC and TIC. Two main products are radio L1000 for short range and radio L2710 for medium range. L2710 is a point-to point radio capable of transporting up to 1.5 Gbps of full duplex traffic. Its specifications are as follows:

- Low latency
- BER less than 10e-11
- Management interface RS-232, RJ-45
- Transmit power 100mw (20dBm)
- Receiver sensivity -53 dBm
- Antenna size 61 cm cassegrain
- Beamwidth <0.4deg

5 Standards

We pay attention to various mmWave standards in which ETSI EN 302 217 is the most complete one. Many different European standards up to 86 GHz have been merged to contribute to this standard. Besides, , since 60 GHz and E-band are very important, in addition to the above standard, they have been discussed distinctly in two different sections.

5.1 60 GHz Frequency Band Standards

60 GHz frequency band is an unlicensed band and thus has many applications and also many organizations have defined different standards for this band. We mention some of them here.

5.1.1 ITU Standard

ITU-R M.1452-1 standard is one of ITU-R's suggestions for the mmWave applications which has been payed attention to the mmWave intelligent transport system. According to it, 57-60 GHz band is for inter-automobile communications and 60-61, 76-77 and 77-81 GHz are for radar and satellite communications.

Since Oxygen has a high absorption property on 60 GHz band, it is used for low range applications [7].

On the other hand, since wavelength of 60 GHz mmWave is small, the designed antennas in this band are more smaller than the antennas in lower frequencies. The received antenna power is dependent to distance, bandwidth and received signal angle. Besides, frequency and pulse modulation have been proposed for mmWave band by ITU-R.

5.1.2 ETSI

ETSI has defined some different applications for 57-66 GHz frequency band. Those are depicted in Table 2, [8].

5.2 60 GHz Industrial Standards

5.2.1 ARIB

ARIB has been defined two standars ARIB-STD T69 and ARIB-STD T74 for 60 GHz band [9, 10].

The first one is for mmWave video transmission equipment with low power radio stations (point-to-point system), while the second one is for very high speed mmWave WLAN with low power radio stations (point-to-multipoint systems). Both standards cover 59-66 frequency band in Japan.

5.2.2 IEEE 802.15.3c (TG3c)

With increasing interest to 60 GHz band, a workgroup for wireless personal network (WPAN) was constituted in IEEE 802.15. In Mars 2005 IEEE 802.15.3c workgroup began to develop a new physical layer for mmWave band on the IEEE 802.15.3-2003 WPAN. To support 2 Gb/s rate, this workgroup developed PHY layer with 3 Gb/s rate [11].

5.2.3 Wireless HD Standards

In 2007, wireless HD (High Definition) issued a standard including uncompressed high definition voice and video with both 5 Gb/s rate and up to 90 meter transmission range for an unlicensed 60 GHz radio. Wireless HD products have been developed 60 GHz mmWave technology somehow that can coexist with wired products [12].

5.2.4 ECMA International Standards

This association established a workgroup TG48 (TC32-TG20) for MAC and Phy layers in the low-range and 60GHz band. According to ECMA-387 a 10Gb/s data rate is provided for the transmission of the different data for WPAN applications (including point-to-point system). The specifications of 60 GHz transceiver has been described in the report of this standard completely [13].

5.3 70/80 Frequency Band (E-band)

As mentioned before, absorption capability of Oxygen in the atmosphere varies with frequency. This absorption will be maximized in the 60 GHz frequency and the resulting signal loss will become 15dBm/km. This will result in limitation in signal transmission range. In the frequencies higher than 60 GHz the signal loss will be lowered and in the 120 GHz frequency that loss will be maximized again. Therefore 70-110 GHz frequency band may be used for data transmission in long-range. In 70/80 band, two frequency bands 71-76 and 81-86 GHz bands are important for data transmission. These bands are used for fixed and mobile terrestrial and satellite services.

Since this band is a licensed one the standards have described that less than the ordinary 60 GHz frequency band. However, we have presented some common E-band ETSI standards in the Table 2.

Table 2. Common 70/80 GHz bands in ETSI [8]

Frequency band (GHz)	Application	ECC report number	Standard number
71-71	Satellite, fixed, mobile	ECC/REC (05)07	EN 302 217
74-75.5	Fixed, mobile, space research	ERC/REC 70-03 ECC/REC/(05)07	EN 302 372 EN 302 217
75.5-76	Satellite, Fixed, mobile, space research	ERC/REC 70-03 ECC/REC/(05)07	EN 302 372 EN 302 217
81-84	Satellite, fixed, mobile	ERC/REC 70-03 ECC/REC (05)07	EN 302 372 EN 302 217
84-86	Satellite, fixed, mobile, compressed links	ERC/REC 70-03 ECC/REC (05)07	EN 302 372 EN 302 217

6 Conclusion

MmWaves haven't x-rays disadvantages and thus this technology can be used as radio, security, medical, etc., applications. Besides, mmWave equipment is small and can provide up to 10 Gbps rates in the range 5km. In addition, their installation and commissioning times are small relative to the other technologies. On the hand, their challenges include small elements and high path loss.

References

1. Yong, S.K., Chong, C.C.: An Overview of Multi gigabit Wireless through Millimeter Wave Technology: Potentials and Technical Challenges. EURASIP Journal of Wireless Communications and Networking (2007)
2. Lehpamer, H.: Millimeter-Wave Radios in Backhaul Networks, Communication Inrastructure Corporation (2008)
3. http://www.e-band.com/
4. http://hkxcz.en.alibaba.com/
5. http://www.gigabeam.com/
6. http://www.loeacom.com/
7. Wells, J.: Multi-gigabit Microwave and Millimeter-Wave Wireless Communications. Artech House pub. (October 2010)
8. ECC Recommendation (05)02: Use of the 64 - 66 GHz Frequency Band for Fixed service
9. ARIBSTD-T69, Millimetre-Wave Video Transmission Equipment for Specified Low Power Radio Station (July 2004)
10. ARIBSTD-T74, Millimetre-Wave Video Transmission Equipment for Specified Low Power Radio Station (Ultra High Speed Wireless LAN System) (May 2001)
11. IEEE 802.15.3c: Millimeter-wave based Alternative Physical Layer Extension for IEEE Standard 802.15.3-2003 titled, Part 15: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High Rate Personal Area Networks (WPANs)
12. http://WWW.WirelessHD.Com
13. ECMA. International, High Rate 60 GHz PHY MAC and PALs. In: ECMA-387, 2nd edn. (December 2010)

A Reduced Complexity Subcarrier Switching Scheme for PAPR Reduction in OFDM System

Sabbir Ahmed[1] and Makoto Kawai[2]

[1] Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu city, Japan
gr045054@ed.ritsumei.ac.jp
[2] Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu city, Japan
kawai@is.ritsumei.ac.jp

Abstract. High values of Peak to Average Power Ratio or PAPR is a fundamental problem of Orthogonal Frequency Division Multiplexing (OFDM) systems. Literature survey shows that recent research works on PAPR reduction are proposing schemes focusing on IEEE standard based systems. One such technique is the null and data subcarrier switching method- a scheme that can reduce PAPR without any need of side information, causes no distortion and can complement other existing methods. However, the associated computational complexity of this method can be very high. In this paper, we propose a modified technique based on the subcarrier switching method that can reduce computational complexity to a great extent with very little or no loss in the PAPR reduction. In addition, we also present further PAPR reduction capability of our scheme with low computational overhead by making compromise on the no-side information advantage.

Keywords: OFDM, PAPR, Subcarrier switching.

1 Introduction

Orthogonal Frequency Division Multiplexing (OFDM) is a well known technique for high speed data communication. OFDM exhibits robust performance in combating inter-symbol interference (ISI) and also gives very good spectrum efficiency. Exploiting these advantages, multiple access schemes like Orthogonal Frequency Division Multiplexing Access (OFDMA) and Multicarrier Code Division Multiple Access (MC-CDMA) have been proposed. MC-CDMA and OFDMA both have strong potentials of becoming future generation de-facto multiple access scheme. In fact, OFDM is being used in various contemporary communication systems including digital audio and video broadcasting (DAB and DVB) [1],[2] and wireless LANs [3], whereas OFDMA has been standardized for the PHY layer of IEEE 802.16 wireless metropolitan area networks and the downlink of 3GPP-LTE [4],[5].

Besides all its advantages, OFDM comes with a fundamental problem of generating occasional very high peaks in its transmit signal which is often quantified

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 67–76, 2011.

by the parameter called "Peak to Average Power Ratio" or PAPR. As power amplifier are generally operated near the saturation point, if PAPR is high, e.g., the peak power of the signal crosses the operating range of the amplifier, it may get driven well into the saturation region causing severe bit error ratio (BER) degradation due to non-linear amplification. On the other hand, increasing operating range of the amplifier or operating it with high back-offs is not practical for many power constrained devices, specially for hand-held mobile terminals. Hence PAPR reduction is a crucial issue for any system that uses OFDM.

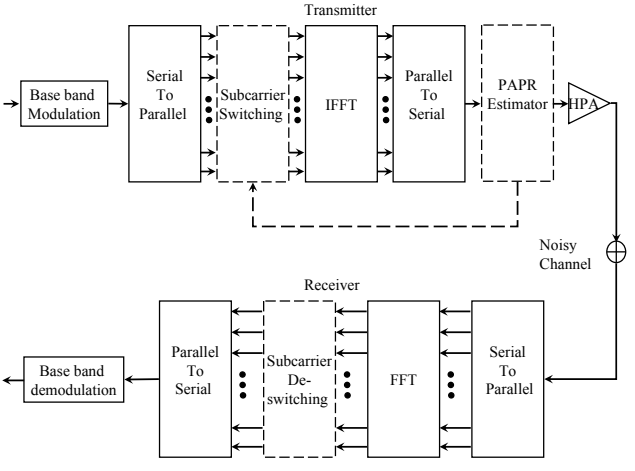


Fig. 1. OFDM system block diagram with subcarrier switching

Literature survey shows that the investigation of PAPR reduction has been a very active field of research for the last couple of years and quite a large number of techniques have been proposed. Amongst them, some of the fundamental and generic proposals for PAPR reduction include signal clipping, selected mapping (SLM) and partial transmit sequences (PTS) etc. [6]-[8]. While signal clipping is a very simple and effective PAPR reduction technique, it can exert a negative impact on the spectral containment of the transmitted signal and thus cause inter-channel interference (ICI) [9]. The PTS and SLM techniques are distortion less methods but they require transmission of extra information known as side information to the receiver, which decreases the data throughput.

Recently, other than considering generalized architecture, PAPR reduction investigations are focusing on systems that are based on different standard specifications, e.g., WiMaX (IEEE 802.16) or wireless LAN (802.11) [10]-[12]. In particular, the scheme described in [11] known as the null and data subcarrier switching method shows significant PAPR reduction by switching data subcarriers with null subcarriers specified in the IEEE802.11a standard specifications [3]. Unlike signal clipping, it does not distort the transmission signal and also there is no need for side information transmission as required by SLM or PTS techniques. But the problem of this method is its high computational overhead.

In this contribution, we propose a modified technique of null and data subcarrier switching which requires much less computations compared to the original method [11] and yet achieves almost same level of PAPR reduction. Then we extend our method of low computational complexity for further PAPR reduction by making a compromise on the advantage of no-side information. We show that to achieve similar PAPR reduction, the original method with no-side information requirement is almost impractical due to its extremely high level of computational overhead. Finally, we take into consideration a non-linear power amplifier and a noisy channel and demonstrate the comparative BER performance of our method with the original one.

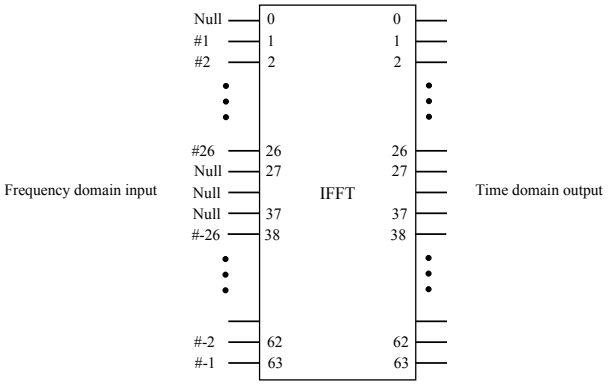


Fig. 2. Inputs and outputs of IFFT

We organize the rest of the paper as follows. In the following section, we show and discuss the system model that we have considered. Detail discussion of our method is given in section 3. Section 4 deals with the performance analysis results and related discussions. And finally in section 5, we make some concluding statements.

2 System Model

Figure 1 shows the model of an OFDM system where blocks with solid lines and dashed lines represent the fundamental components and the components for PAPR reduction scheme respectively. In this model, at first binary random input data is baseband modulated that generates input symbol given by $x[i] = x[0], x[1], ..., x[N-1]$. They are then converted from serial to parallel and fed into the IFFT module. The IFFT module performs the task of multicarrier modulation. Output of the IFFT is parallel to serial converted and the resultant discrete time domain symbols is given by Eqn. 1.

$$X[n] = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} x[i] e^{j\frac{2\pi ni}{N}}, 0 \leq n \leq N-1 \tag{1}$$

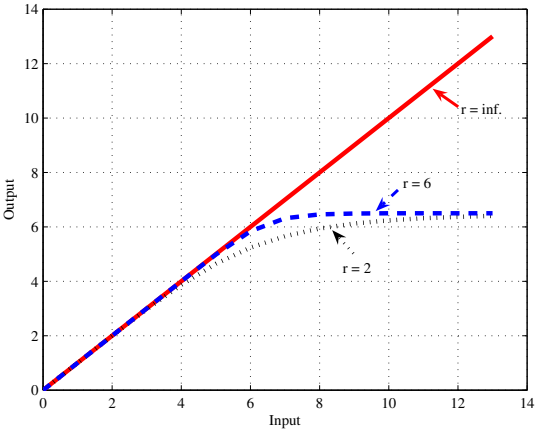


Fig. 3. Non-linear power amplifier input-output response

The time domain signal is then amplified by a high power amplifier (HPA) and transmitted to the noisy wireless channel. The receiver perform the opposite operations and finally detects the transmitted symbols.

Now, for PAPR reduction purpose a subcarrier switching module along with PAPR estimator block are inserted in the transmitter whereas a subcarrier de-switching module is placed in the receiver. The subcarrier switching module performs the operation of switching between null and data subcarriers. We describe in detail its operation in the following section. The PAPR estimator block determines and stores the PAPR value of all the switching combinations and transmits the signal with the lowest PAPR. PAPR in dB is expressed by the following Eqn.

$$PAPR(dB) = 10\log_{10}\frac{max|X[n]|^2}{E\left[|X[n]|^2\right]} \tag{2}$$

where $E[\cdot]$ denotes expectation.

As mentioned before, our PAPR reduction scheme is based on the IEEE 802.11a WLAN standard [3]. This standard specifies a total of 64 subcarriers that includes 12 null subcarriers. The remaining 52 subcarriers are used for data (48 subcarriers for user data and 4 subcarriers for pilot data) transmission. Among the 12 null subcarriers, 6 subcarriers work as guard-band at the low-frequency edge of the spectral band of the subcarriers while 5 subcarriers serve as guard-band at the high-frequency edge. The remaining one is placed at the middle to avoid direct current energy. The IFFT input-output lay out is shown in Fig. 2 [3].

Finally, for simulating the HPA, we consider the Solid State Power Amplifier (SSPA) model given in [13] where the AM-to-AM conversion characteristics is expressed by Eqn. 3.

$$F[x] = \frac{x}{[1+(x/A)^{2r}]^{1/2r}} \tag{3}$$

where x is the amplitude of the input signal, A is the saturated output level and r is the non-linearity level. This model only considers AM-to-AM non-linearity. The parameter r can be used to tune the level of non-linearity and it is depicted in Fig. 3. A large value of r turns this amplifier into a linear one where as very small values make it behave as a simple clipping amplifier.

Step: 1 *Select* $\{NC_1, NC_2, .., NC_P\}$, P *number of null subcarriers out of available* L *for switching with total available* $N - L$ *data subcarriers.*
Step: 2 *Construct data subcarrier combination taking* P *number of subcarriers at a time from data subcarrier index vector of length* $N - L$.
Step: 3 *Switch the* P *null subcarrier with data subcarriers generated in Step 2 such that the order of null subcarrier indices and data subcarrier indices are same.*
Step: 4 *Apply* IFFT *for OFDM operation and calculate PAPR from the time domain signal.*
Step: 5 *Repeat Step 2 to Step 4,* $C_P^{N-L} - 1$ *times for all other possible subcarrier combinations and record the lowest PAPR.*
Step: 6 *Apply the data subcarrier combination corresponding to the lowest PAPR as switched data subcarriers for signal transmission.*

Fig. 4. Original null and data subcarrier switching

3 PAPR Reduction Method

The main contribution of this paper is a proposal to reduce the computational overhead involved in searching low PAPR producing data subcarrier switching positions proposed in [11]. The additional components required to implement the switching technique is marked by dashed lines in Fig. 1 and we explain the working principle here. If, we consider an N-size IFFT module with L number of total null subcarriers, i.e., $\{NC_1, NC_2, .., NC_L\}$, as per [11], once the number of null subcarriers for switching is determined, say P, the null subcarriers for switching $\{NC_1, NC_2, .., NC_P\}$ are needed to be switched with all possible data subcarrier indices combinations for finding the lowest PAPR. And the underlying constraint is that the order of data subcarrier and null subcarrier indices must be same so that no side information is necessary for the receiver to perform the de-switching operation. As a result C_P^{N-L} number of search operations are involved for every input data block, which in turn means that many IFFT operations are required. For example, a system with $N = 64$ and $L = 12$, if $P = 4$ is considered, $C_4^{52} = 270725$ number of search operations are required. The time involved in such calculation may make this technique impractical for many delay sensitive applications.

On this basis, we were motivated to consider modified approach for searching that can reduce computational burden without making significant sacrifice in PAPR reduction. In our proposed method, we apply incremental search by considering one null-data subcarrier switching at a time. For example, for $P = 2$,

we at first only consider NC_1 and search C_1^{N-L}, i.e., $N-L$ times to look for the data subcarrier position, say D_1 yielding lowest PAPR when switched with NC_1. After this, we apply the same operation for P_2. But in order to keep the order of null subcarriers and switched data subcarriers, we first remove all the data subcarriers positions $\geq D_1$ from the search space. Thus the size of the data subcarrier search space for $NC_i, i \geq 1$ is dependent on the outcome of search with NC_{i-1}. Step-by-step by descriptions of the original null and data subcarrier switching and our proposed methods are listed in Figs. 4 and 5 respectively.

Step: 1 *Select* $\{NC_1, NC_2, .., NC_P\}P$ *number of null subcarriers out of available* L *for switching with total available* $N-L$ *data subcarriers and set* $i = 1$.
Step: 2 *Select a data subcarrier from data subcarrier index vector.*
Step: 3 *Switch the the null subcarrier* NC_i *with the data subcarrier generated in step* 2.
Step: 4 *Apply* IFFT *for OFDM operation and calculate PAPR from the time domain signal.*
Step: 5 *Repeat Step 2 to Step 4 for for all other data subcarriers one at a time, record the lowest PAPR and mark the corresponding data subcarrier as* DC_i.
Step: 6 *Remove all data subcarrier indices* $\geq DC_i$ *from data subcarrier index vector and increase* i *by 1.*
Step: 7 *Repeat Step 2 to Step 6* $P-1$ *times.*
Step: 8 *Apply the data subcarrier combination* $\{DC_1, DC_2, .., DC_P\}$ *as switched data subcarriers for signal transmission.*

Fig. 5. Proposed null and data subcarrier switching

The computational complexity reduction is depicted in Table 1. As we can see, for $P = 2$, maximum computations required by our method is only about 7.76% of the original one, i.e., a reduction of around 92.24%. Again for the original method, with higher values of P, one expects more reduction in PAPR. But at the associated computation cost can be extremely high. For example, as mentioned earlier, for $P = 4$, 270725 number of searching is required which makes it almost unfeasible for practical implementation. As a matter of fact if significant PAPR reduction can be achieved with $P = 2$, higher values of P may not be required. Thus primarily we compare our method with the original one for $P = 2$.

Table 1. Comparative computational complexity

No. of switched null subcarriers, P	No. of Searching	
	Original	Proposed (Maximum)
2	1326	103
4	270725	202
6	20358520	207
8	752538150	388

Our method with $P = 4$ is also computationally efficient compared to the original one for higher values of P. But due to reduction of the search space for every increment in P (Step 6), there remains a possibility of the search space getting very limited for $P_i, i \geq 1$. This problem becomes very much probable when $P \geq 2$. In order to avoid this problem, we propose another searching where we relax the need for keeping the order of the switched null and data subcarriers by modifying step 6 as below,

Step: 6 Remove data subcarrier index DC_i from data subcarrier index vector and increase i by 1.

Table 2. Simulation parameters

Modulation scheme	BPSK
IFFT size	64
Number of data subcarriers	52
Number of switched null subcarriers	$2, 4$
Total number of OFDM symbols	10^4
Oversampling factor	4
HPA Model	SSPA
HPA level of non-linearity, r	2
Channel model	AWGN

This eventually also removes the constraint of switched null subcarriers and data subcarriers having the same order. As a result the receiver can not perform the opposite operation of switching unless it has some extra information. Thus the need for side information comes into effect. We think this trade-off is worthy of, since some side information (e.g. control signal) is always needed irrespective of whether any PAPR reduction scheme is applied or not.

4 Performance Evaluation

Our simulation is based on the IEEE802.11a standard specification. However, without distinguish between user data and pilot data subcarriers, we used all the 52-not null subcarriers for switching with null subcarriers. In contrast, one may recall, 48 data subcarriers was used in [11]. We considered BPSK modulation and investigated scenarios for both $P = 2$ and $P = 4$. For accurate estimation of PAPR, we considered an over sampling factor of 4 [14]. All the pertinent simulation parameters are listed in Table 2.

Figure 6 shows the complementary cumulative distribution function (CCDF) of PAPR without any PAPR reduction scheme, the original null and data subcarrier switching scheme and our proposed scheme, the later two with $P = 2$. In this figure, we refer to the original null and data subcarrier switching as "SC Switching" and our method as "Reduced complexity (No side info required)" respectively. To make a fare comparison, we selected the null subcarriers indexed at ± 27 (pls. refer to Fig. 2) as was done in [11]. As can be seen from

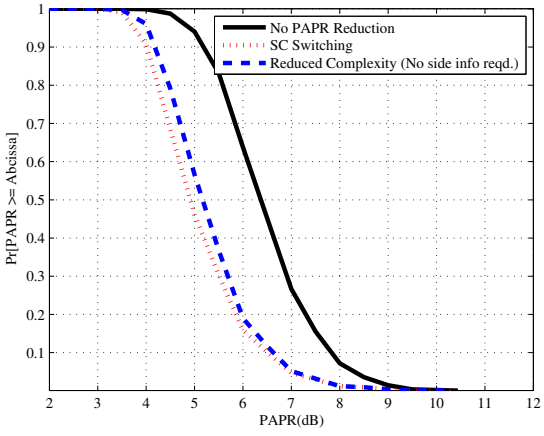


Fig. 6. Comparative PAPR distribution for $P = 2$

Fig. 6, compared to no PAPR reduction scheme, our method achieves significant PAPR reduction which is slightly inferior to the original method. For example the probability of PAPR equalling or exceeding $6dB$ is around 0.6 when no PAPR reduction scheme is applied. Compared to this, the corresponding probabilities of our proposed method and the original method are approximately 0.2 and 0.18 respectively. Again in other terms, with a 0.2 probability, our method and the original method shows PAPR of approximately $6dB$ and $5.9dB$ respectively. It means our method shows a 1.66% degradation in PAPR reduction. This very little loss is more than compensated when we consider 92.24% reduction in computational complexity as was mentioned before.

In Fig. 7, we show the result with $P = 4$. Here, we selected null subcarriers positioned at $\pm 27, \pm 28$ for switching. Our methods of both requiring side-info and no side-info are shown. If we sacrifice some throughput by sending small amount of side-info, our method can achieve good PAPR reduction with very little computations compared to the original method. Here, we need to send the switched data subcarrier index to the receiver with which it will exchange the values of the null subcarrier. Since $log_2 64 = 6$ bits is required to represent a subcarrier index, four $6 - bit$ words will be needed for this purpose. In this figure we can see with a 0.2 probability, our method (side info reqd.) and the original method shows PAPR of approximately $5dB$ and $4.7dB$. And the associated number of searchings are 270725 and 202 (maximum) respectively. Which means our method shows 6% higher PAPR and requires additional side-info but reduces 99% computational complexity.

Finally, in Fig. 8, the BER result for our proposed system ($P = 2$) under the influence of AWGN channel and non-linear HPA is depicted. We fix the non-linearity level $r = 2$ and the saturation power level at $3dB$. This result shows the combined effect of the noise contamination by the wireless channel and the

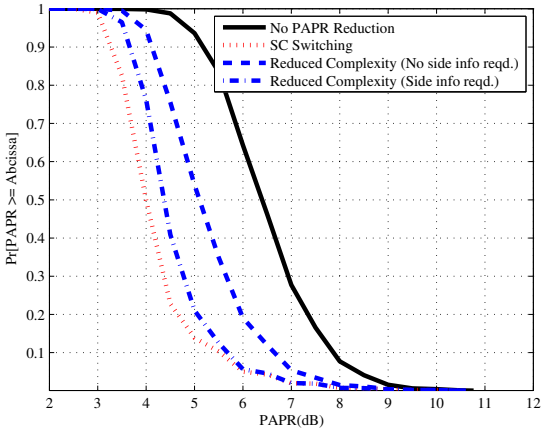


Fig. 7. Comparative PAPR distribution for $P = 4$

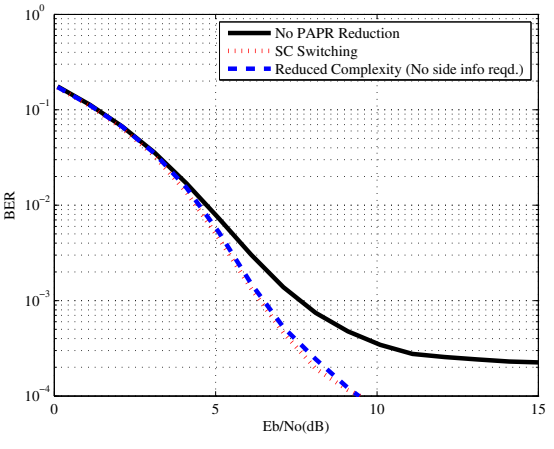


Fig. 8. Comparative BER performance for $P = 2$

non-linear amplification due to high PAPR. As we can see from this figure, our proposed method shows almost similar BER as the original one. And this performance is better compared to OFDM system with no-reduction strategy applied.

5 Conclusions

Null and data subcarrier switching scheme is a very recent technique of PAPR reduction which appears to be an effective solution for IEEE 802.11a OFDM system. But the associated computational overhead can be very high specially when higher number of null subcarriers are used for switching. In this paper, we presented a modified technique for searching the data subcarriers positions to be

switched with null subcarrier with reduced complexity. We demonstrated that when 2 null subcarriers are used for switching, the proposed method achieves very high computational complexity reduction yet achieves similar PAPR compared to the original method. And the associated BER performance is also very good. For higher number of null subcarriers, we compromise on the no-side information requirement to achieve acceptable PAPR with very low computational overhead.

References

1. ETSI, Radio broadcasting systems: Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers, ETS 300 401 ed.2 (1999)
2. ETSI, Digital video broadcasting systems: Framing structure, channel coding, and modulation for digital terrestrial television. European Telecommunications Standard, EN 300–744 (1997)
3. IEEE, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications. High Speed Physical Layer in the 5GHz Band, IEEE Std 802.11a (1999)
4. IEEE, Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE std. 802.16-2004 (Revision of IEEE Std. 802.16-2001) (2001)
5. http://www.3gpp.org
6. Li, X., Cimini, L.J.: Effects of clipping and filtering on the performance of OFDM. IEEE Communications Letters 2(5), 131–133 (1998)
7. Bauml, R.W., Fischer, R.F.H., Huber, J.B.: Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping. IEE Electronics Letters 32(22), 2056–2057 (1996)
8. Cimini, L.J., Sollenberger, N.R.: Peak-to-average power ratio reduction of an OFDM signal using partial transmit sequences. IEEE Communications Letters 4(3), 86–88 (2000)
9. Thompson, S.C., Proakis, J.G., Zeidler, J.R.: The effectiveness of signal clipping for PAPR and total degradation reduction in OFDM systems. In: Proc.: IEEE Global Telecommunications Conference, GLOBECOM 2005, St. Louise, Montana, USA, pp. 2807–2811 (2006)
10. Chaokuntod, S., Uthansakul, P., Uthansakul, M.: Fast Dummy Sequence Insertion method for PAPR reduction in WiMAX systems. In: Proc.: Asia Pacific Microwave Conference, Singapore, pp. 2168–2171 (2010)
11. Wong, K.T., Wang, B., Chen, J.-C.: OFDM PAPR reduction by switching null subcarriers and data-subcarriers. Electronics Letters 47(1), 62–63 (2011)
12. Hussain, S., Guel, D., Louet, Y., Palicot, J.: Performance comparison of PRC based PAPR reduction schemes for WiLAN Systems. In: Proc.: European Wireless Conference, Aalborg, Denmark, pp. 167–172 (2009)
13. Rapp, C.: Effects of HPA-nonlinearity on a 4-DPSK/OFDM signal for a digital sound broadcasting system. In: Proc.: Second European Conference on Satellite Communications, Liége, Belgium, pp. 179–184 (1991)
14. Tellambura, C.: Computation of the Continuous-Time PAR of an OFDM signal with BPSK subcarrier. IEEE Communications Letters 5(5), 185–187 (2001)

An Indoor Location-Aware System Based on Rotation Sampling in Positioning

Chu-Hsing Lin[1], Jung-Chun Liu[2], Chien-Hsing Lee[3], and Tang-Wei Wu[4]

Department of Computer Science, Tunghai University,
407 Taichung, Taiwan
{chlin[1],jcliu[2],g96257004[3],g99350031[4]}@thu.edu.tw

Abstract. With the rapid development in hardware and software of wireless communications, applications of real-time positioning systems have been widely used; among them the Global Positioning System is the most popular [1]. However, the satellite signal is obstructed by the sheltering effect of the building and becomes too weak to be applicable for indoor positioning. Instead, the wireless sensor networks techniques are suitable for indoor positioning, for which many related researches and applications have been proposed. Based on the ZigBee wireless technology operating on the IEEE 802.15.4 physical radio specification, we use the Received Signal Strength Indicator together with the Signal Patent Matching method for positioning. Furthermore, to enhance accuracy of positioning, we adopt a sampling method to compare on-line samples as well as the rotation sampling positioning method.

Keywords: Wireless sensor networks, Indoor positioning, Rotation sampling, Received Signal Strength Indicator, Signal Patent Matching.

1 Introduction

In recent years, due to rapid development of wireless communication techniques and increasing popularity of mobile devices, measurement systems tend to be wireless, miniaturized, modular, web-based, economy-oriented, and power-saving; and so the real-time locating services technique based on wireless networks becomes one of the most active research topics. On the other hand, because the wireless network has advantages such as power-saving, large network nodes, and low cost, the wireless sensing techniques have been widely used in the home life, environmental monitoring, commercial, and military applications. This paper studies indoor positioning. With the aim to enhance accuracy of positioning, the Received Signal Strength Indicator (RSSI) as well as the Signal Patent Matching method is used to calculate the possible position of targets.

The rest of this paper is organized as follows. Section 2 briefly describes related works of indoor positioning techniques. Section 3 introduces the experimental environment, provides experimental results, and performs analysis. Finally, Section 4 concludes the paper.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 77–81, 2011.

2 Related Works

In recent years, extensive research has been conducted on the Wireless Sensor Network (WSN), in which positioning of sensors is a very important issue. Currently, positioning algorithms used in WSN can be divided into two categories: Range-base and Range-free. The former required some information to measure the distance, such as the distance, angle and time difference of the location node and location tag. It is more accurate but need to pay additional cost of equipments. The related methods include Time of Arrival (TOA) [2][3], Time Difference of Arrival (TDOA) [4][5][6], Angle of Arrival (AOA) [7][8][9], and Received Signal Strength (RSS) [10][11]. The latter can locate the position without information of the distance, and the angle of location node and location tag. Its accuracy is less than that of Range-base, but it does not need to pay extra costs for equipments to obtain the distance and angle data. Thus, it is more suitable for low-cost and low-power WSN applications. Its related methods include DV-Hop Localization Algorithm [12], and Signal Patent Matching [13].

TOA, AOA, and TDOA positioning methods require more complex calculations and additional hardware to perform the time synchronization mechanism; whereas the RSS positioning method has advantages of low-cost, simple operation, and without the need to modify hardware. So we adopt RSSI for indoor positioning.

3 Experimental Methods and Results

In the experiment, a laboratory with a rectangular space of 10 x 4.5 square meters is used, in which a space of 6 x 3 square meters is planned into 11 x 6 = 66 reference points with an adjacent distance of 60 cm, as shown in Fig. 1.

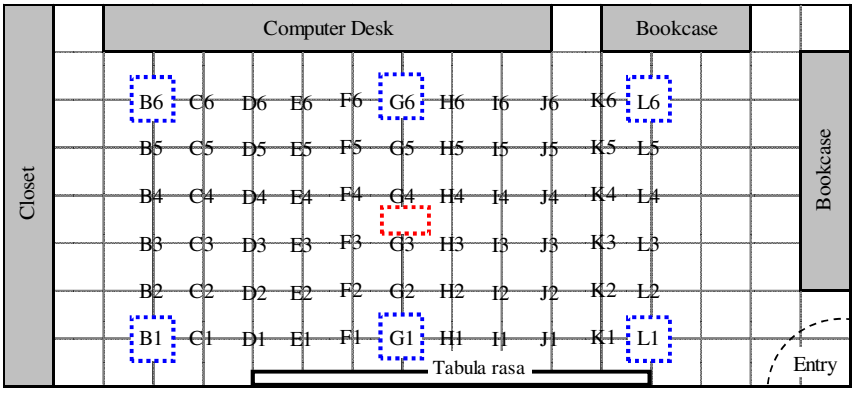


Fig. 1. Experimental environment

As listed in Table 1, four sets of experiments are performed using different location tag sampling methods with different rotation angle, in which the control group uses the fixed location tag sampling method. For Case 4, the object interference comes from a person or moving object.

Table 1. Experimental sets

	Case 1	Case 2	Case 3	Case 4
Sampling Methods	Fixed without rotation	180°	360°	360° with object interference

Figure 2 shows the experimental results. In view of accuracy of positioning, Case 3 is the best with an average location error about 65 cm, while Case 1 and Case 2 have close average location errors of 84 cm and 88 cm, respectively. Case 4 is the worst case with an average location error of 94 cm due to interference of moving objects.

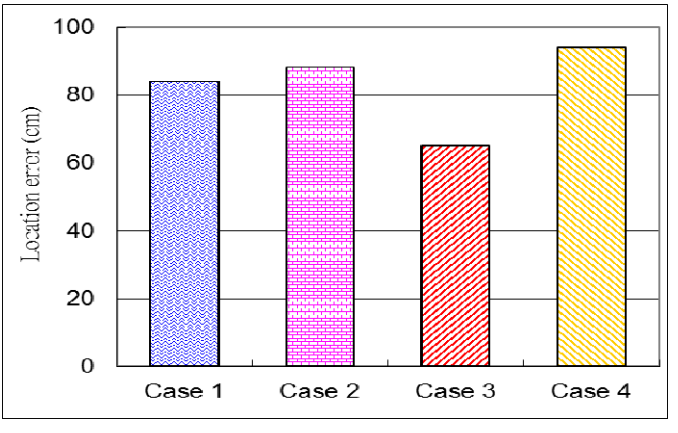


Fig. 2. Average location errors with different positioning sampling methods

Figure 3 shows distribution of location errors for the best three sampling methods in positioning. For Case 3, about 52% localization errors are under 60 cm, 41% localization errors between 61 cm and 120 cm; and especially, no location errors greater than 180 cm are measured, whereas in Case 1 and Case 2, about 9% location errors exceeding 180 cm.

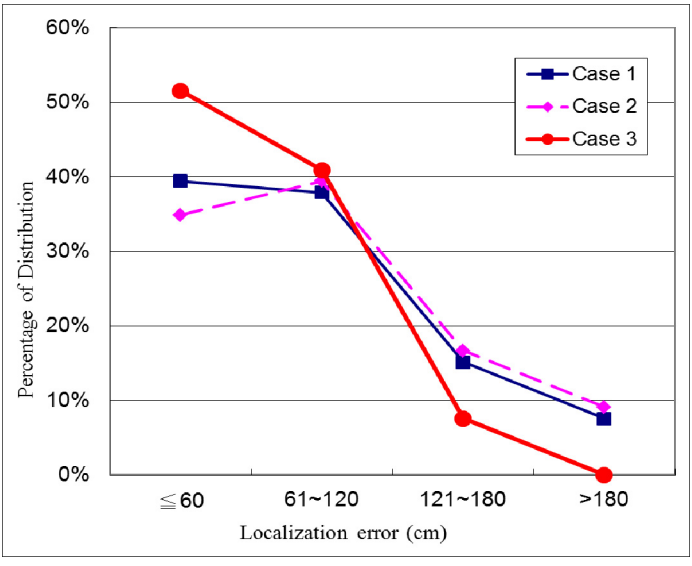


Fig. 3. Distribution of location errors with different sampling methods

4 Conclusion

The experiment results show that Case 3 measurement setting with a rotation angle of 360° performs best, in which about 22% enhancement of average location accuracy is achieved compared to Case 1 with the fixed and non-rotating sampling method; and particularly, more than half, that is, 52%, of localization errors can be controlled under 60 cm, so we infer that the Case 3 setting is best for indoor positioning in a small space. Moreover, based on average location errors for Case 3 and Case 4 settings, we observe that in wireless environments if the location tag is suffered from interference of moving objects, the average accuracy of positioning can be reduced by as much as 44% and the location error is much aggravated.

Acknowledgement. This work was supported in part by Taiwan National Science Council under grant: NSC 99-2221-E-029-039 -MY3.

References

1. Ladd, A.M.: Using Wireless Ethernet for Localization. In: IEEE International Conference on Intelligent Robots and Systems, Lausanne, Switzerland, September 2002, pp. 402–408 (2002)
2. Heidari, M.: Identification and Modeling of the Dynamic Behavior of the Direct Path Component in TOA-based Indoor Localization Systems. In: Electrical & Computer Engineering. Worcester Polytechnic Institute (May 2008)

3. Venkatraman, S., Caffery, J., You, H.R.: A Novel TOA Location Algorithm Using Los Range Estimation for NLos Environments. IEEE Transaction on Vehicular Technology 53(5), 1515–1524 (2004)
4. Cong, L., Zhuang, W.: Hybrid TDOA/AOA Mobile User Location for Wideband CDMA Cellular Systems. IEEE Transactions on Wireless Communications 1(3), 439–447 (2002)
5. Bocquet, M., Loyez, C., Delaï, A.B.: Using Enhanced-TDOA Measurement for Indoor Positioning. IEEE Microwave and Wireless Components Letters 15(10), 612–614 (2005)
6. Patwari, N., Ash, J.N., Kyperountas, S., Hero, A.O., Moses, R.L., Correal, N.S.: Locating the Nodes: Cooperative Localization in Wireless Sensor Networks. IEEE Signal Processing Magazine 22(4), 54–69 (2005)
7. Jiang, J.-R., Hsu, Y.-J.: Localization with Rotatable Antennas for Wireless Sensor Networks. Department of Computer Science and Information Engineering at National Central University (2010)
8. Peng, R., Sichitiu, M.L.: Angle of Arrival Localization for Wireless Sensor Networks. In: Sensor and Ad Hoc Communications and Networks, pp. 374–382 (2006)
9. Niculescu, D., Nath, B.: Ad Hoc Positioning System Using AOA. Computer and Communications 3, 1734–1743 (2003)
10. Lihan, M., Sawano, H., Tsuchiya, T., Koyanagi, K.: Comparison of Orientation-aware RSS-based Localization Algorithms. In: Mobility Conference, vol. (100) (2008)
11. Papamanthou, C., Preparata, F.P., Tamassia, R.: Algorithms for Location Estimation Based on RSSI Sampling. In: Algorithmic Aspects of Wireless Sensor Networks, July 2008, pp. 72–86 (2008)
12. Tseng, L.-M., Lee, K.-F.: Economical Traveling Scheme for Localizing Sensor Network with Mobile Anchor. National Digital Library of Theses and Dissertations in Taiwan (July 2006)
13. Bahl, P., Padmanabhan, V.N.: RADAR: An in-building RF-based User Location and Tracking System. IEEE Computer and Communications Societies 2, 775–784 (2000)

Research on the ZigBee-Based Indoor Location Estimation Technology

Chu-Hsing Lin[1], Jung-Chun Liu[2], Sheng-Hsing Tsai[3], and Hung-Yan Lin[4]

Department of Computer Science, Tunghai University, 407 Taichung, Taiwan
{chlin[1],jcliu[2],g96357002[3],g99350006[4]}@thu.edu.tw

Abstract. In recent years, as rapid advances in wireless and mobile communications and continuing decreases in hardware costs, applications of the wireless sensor network are widespread; whereas location estimations in the wireless sensor technology are crucial for their use. Among them, Location-based services have been developed rapidly and find their applications in areas such as medical care, warehouse management, and mobile guide systems in public spaces. In this paper, a fingerprint based location estimation technology in the ZigBee networks is investigated, in which the collection method to build the signal strength database and the configuration of sensor nodes are examined. Furthermore, the k-nearest neighbor algorithm is used to increase accuracy of location estimations for compliance with relevant applications.

Keywords: wireless sensor network, indoor localization, fingerprint, k-nearest neighbor algorithm.

1 Introduction

As the wireless sensor technology advances, its applications in environmental, military, and daily home life are numerous. The location based technology is very important for applications such as home care, automated warehouse management, personnel management, and museum tour guide. For home care, the location based technology is used to provide real-time information of the patients and avoid losing traces of them; and with the aide of control systems, it can automatically issue warning messages to signal needed assistance whenever it senses that the elderly stays in the stairway or bathroom too long. For storage management, the location based technology can be used to find the relevant items and automatically send them to their destinations with automatic control systems [1][2][3]. Applications of the location technology are very wide, for more examples, it can be used to seek one's friends or family members in vast venues or public spaces such as supermarkets, museums, and libraries.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 82–86, 2011.

2 Research Methods

In our study, we use the fingerprint approach together with the weight method, and then adjust the weight by the k-nearest neighbor algorithm. These methods are described in the following [4~9].

2.1 Fingerprint Method

In this paper we adopt the fingerprint method reported by Bahl et al. in [10] as the primary location method for indoor environments. The fingerprint method uses the signal intensity as the feature of the reference point. In order to determine the location of a tag, the strength of the received signal from sensor nodes is compared with the signal strength database to estimate similarity of its feature.

2.2 Weighted Scaling for Sensor Nodes

Thus, in our location estimation algorithm higher weights are assigned to sensor nodes with stronger signal strength. In the experiment, the weight is set as signal strength / 100. For applications need to estimate locations for tags in a wider range of place, the weight can be assigned with larger values. Equation (1) is used to calculate the weighted difference value:

$$diff\_val_N = \sum_{i=1}^{K} \left(\frac{RSS_i'}{100} \right) \times \left(RSS_i' - RSS_i \right)^2 \quad (1)$$

3 Experimental Settings and Results

The experimental settings and results are given and discussed in the following.

3.1 Effect of Sensor Nodes Configuration

As expected, the number of sensor nodes and configurations of their location affect accuracy of location estimations. In the experiment, four or six sensor nodes in different configurations are investigated as shown in Figure 1.

1. In Configuration 1, four sensor nodes are placed in B, C, D, and E; in this way, a half of the experimental environment is not surrounded by these sensor nodes.
2. In Configuration 2, four sensor nodes are placed in the four corners of the laboratory, A, C, D, and F.
3. In Configuration 3, six sensor nodes are placed in A, B, C, D, E, and F.

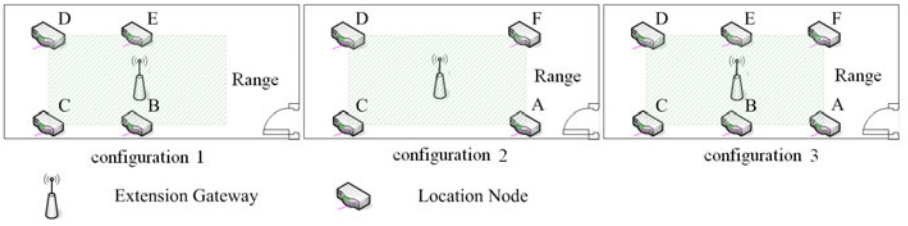


Fig. 1. Three configurations of sensor nodes

Figure 2 shows experimental results of the best, average, and worst location errors for three configurations of sensor nodes. We find that location errors are smallest with Configuration 3, in which the average location error is 86.6 cm; and largest with Configuration 1, in which the average location error is 128.1cm. In the experiment, the location errors are found to be largest at the left and right to the center locations of the laboratory. To solve this problem, two sensor nodes are added to the top and bottom locations in the center of the laboratory in Configuration 3. As shown in Figure 2, the average location error in Configuration 2 is 111.8 cm, and the average location error is diminished to 86.6 cm in Configuration 3. We also observe that the location errors are most uniformly distributed in Configuration 3. We conclude that by placing two sensor nodes at the top and bottom in the center, the location error is effectively reduced.

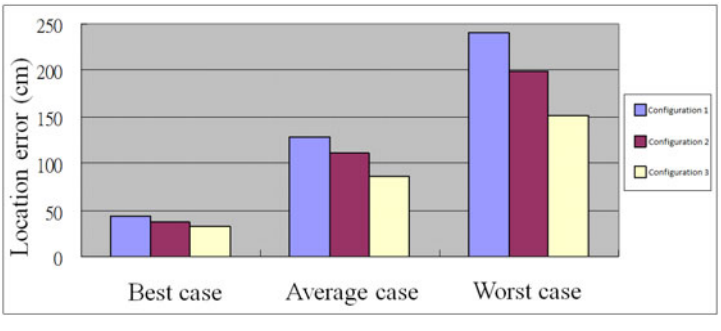


Fig. 2. Location errors for three sensor configurations

3.2 Weighting Effect

Figure 3 shows the best, average, and worst location errors with or without considering weights formulated in Equation (1). We find that the average location error drop from 86.6 cm to 84.8 cm, and in the best case, from 31.9 cm to 28.0 cm, an improvement of 12%.

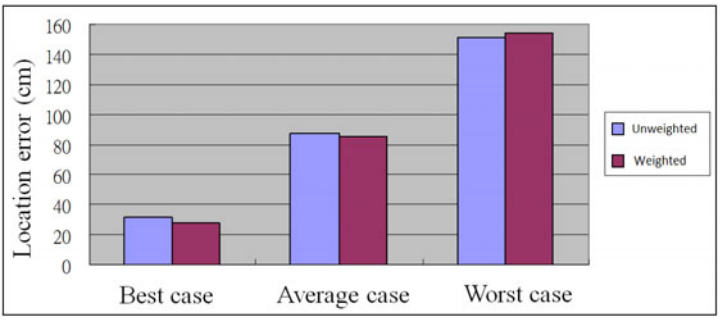


Fig. 3. Location errors with or without considering weights

3.3 Effect of Weight Assignments

Four sets of weight assignments are investigated to find their effects on the location estimation.

In Setting 1, the k-nearest neighbor method is not used, and the location is chosen according to the minimum difference value; whereas in Setting 2 to Setting 4, the k-nearest neighbor method is considered. In Setting 2, both the weights in the condensed and dispersed patterns are set as 1. In Setting 3, both the weights in the condensed and dispersed patterns are set according to their differences. In Setting 4, the weight in the dispersed pattern is set as 1, while the weight in the condensed pattern is set according to the square of the difference.

Figure 4 shows the location errors for the four weight assignments. We find that with the k-nearest neighbor method in Setting 2, the average location error is reduced by 10.1cm, while the average location error is further reduced by 2.8 cm by adjusting weights in Setting 4. Especially, in the best-case scenario the location error is reduced by 5.7cm. For applications that can exclude the worst-case scenario, the improvement on location estimations is even more effective.

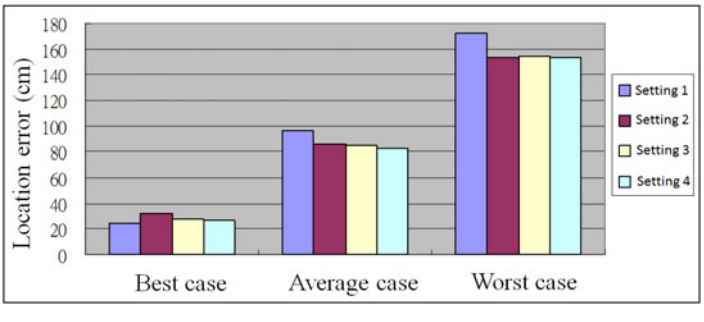


Fig. 4. Location errors with different weight settings

4 Conclusion

We investigate the fingerprint technique with weighted scaling for location estimations in indoor environments. The signal strength database is built with data collected with an all directions transmission approach by a turntable setting to evenly send out signals to reduce directional effect of wireless transmission and effectively reduce data collection time. A variety of sensor nodes configurations are tested to improve accuracy of location estimations. The experimental results show that the average location error is reduced from 128.1 cm for Configuration 1 to 82 cm for Configuration 3 and weighted scaling in Setting 4. For k-nearest neighbor algorithm, various weight settings are examined to reduce location errors. The experimental results show that the location error is reduced from 32.2 cm to 26.5 cm, an 18% improvement. For applications in which events of the worst case can be eliminated, the benefit of weighted scaling will be more significant.

Acknowledgments. This work was supported in part by Taiwan National Science Council under grant: NSC 99-2221-E-029-039 -MY3.

References

[1] Enge, P., Misra, P.: Special Issue on GPS: The Global positioning System. Proc. of the IEEE, 3–172 (1999)
[2] Roberts, R.: IEEE P802.15 Wireless Personal Area Networks: Ranging Subcommittee Final Report, Ranging Subcommittee of TG4a, Harris Corporation (2004)
[3] Heidari, M.: Identification and Modeling of the Dynamic Behavior of the Direct Path Component in ToA-Based Indoor Localization. EURASIP Journal on Advances in Signal Processing (2008)
[4] Seidl, T., Kriegel, H.P.: Optimal Multi-Step k-Nearest Neighbor Search. In: Proc. ACM SIGMOD, Harris Corporation (1998)
[5] Khoury, H.M., Kamat, V.R.: Evaluation of position tracking technologies for user localization in door construction environments. Automation in Construction 18(4), 444–457 (2009)
[6] Lihan, M., Tsuchiya, T., Koyanagi, K.: Orientation-Aware Indoor Localization Path Loss Prediction Model for Wireless Sensor Networks. In: Takizawa, M., Barolli, L., Enokido, T. (eds.) NBiS 2008. LNCS, vol. 5186, pp. 169–178. Springer, Heidelberg (2008)
[7] Dawes, B., Chin, K.-W.: A comparison of deterministic and probabilistic methods for indoor localization. Journal of System and Software 84(3), 442–451 (2011)
[8] Wessels, A., Wang, X., Laur, R., Lang, W.: Dynamic indoor localization using multilateration with RSSI in wireless sensor networks for transport logistics. Procedia Engineering 5, 220–223 (2010)
[9] Patwari, N., Ash, J.N., Kyperountas, S., Hero, A.O., Moses, R.L., Correal, N.S.: Locating the Nodes: cooperative localization in wireless sensor networks. IEEE Signal Processing Magazine 22(4), 54–69 (2005)
[10] Bahl, P., Padmanabhan, V.N.: RADAR: An in-Building RF-Based User Location and Tracking System. In: Proceedings of IEEE INFOCOM Conference, vol. 2, pp. 775–784 (2000)

Visible Watermarking Based on Multi-parameters Adjustable Gamma Correction

Chu-Hsing Lin[1], Chen-Yu Lee[2], Tzu-Chien Yang[1], and Shin-Pin Lai[1]

[1] Department of Computer Science, Tunghai University,
Taichung 407, Taiwan
chlin@go.thu.edu.tw, smf@islab.cs.thu.edu.tw
[2] Department of Computer Science, National Chiao-Tung University
1001 Ta-Hsueh Road, HsinChu, 30050, Taiwan
chenyu@cs.nctu.edu.tw

Abstract. With the progress of information technology, digital data is a major trend of business communication today. Therefore, copyright protection of digital data has been an important research issue of information security. In the paper, we proposed an improved visible watermark scheme using multi-parameters adjustable Gamma Correction to reduce the distortion, which can increase color saturation. Our experiments showed that the PSNRs of the stego images with watermark embedded were up to 54.50 dB and 58.66 dB.

Keywords: visible watermark, copyright protection, Gamma Correction, PSNR.

1 Introduction

Digital watermarking technology is used to embed watermarks in digital media including image, audio, video, and etc. On the Internet application, we embed some logo, like personal image, trademark, in digital media to prevent it from illegal download or use. Even though the illegal users steal the digital media, it is easy to approve and determine the intellectual property by extracting watermarks of digital media and to avoid being invaded.

The watermark technology is divided into visible watermarking and invisible watermarking. The main advantage of visible watermarking [1-3] is that the watermark can be seen by eyes without extracting. However, its shortcoming is that watermark would destroy the presentation of media and it would be subtly removed by illegal user. As a result, visible watermarking is not appropriate for the modern digital application. In contrast, the invisible watermark [4-5] must be extracted by specific process, but it can reserve the original presentation of the cover image. Fig.1 shows examples of visible watermark invisible watermark, respectively.

The paper proposed a comprehensive visible watermarking based on multi-parameters adjustable gamma correction, namely MPA-CVW, the watermark would change its background color to fit the embedded area. Multi-parameters adjustable gamma correction algorithm increases the saturation of the image and makes the

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 87–92, 2011.

(a)

(b)

Fig. 1. The examples of (a) visible watermark, and (b) invisible watermark

watermark easier to be identified by human eyes. Our experiments showed that the PSNRs of the images embedded the watermark were up to 54.50 and 58.66 dB.

The rest of this paper is organized as follows. Section 2 describes the concept of Gamma Correction and its applications. Section 3 represents the proposed MPA-CVW scheme and the experiments results are shown in Section 4. Finally, we have a conclusion.

2 Gamma Correction and Its Application

2.1 Gamma Correction

Many monitors have the Power-law property. The same unit voltage cause different brightness in early CRT monitors. Its strength isn't linear. Beside, for properly display, it produces a curve deliberately. Finally, output display is linear. It calls Gamma Correction.

We summarized two characteristics in human vision systems:

1. People are more sensitive to the difference of grayness than the difference of hue.
2. People are more sensitive to the difference of low brightness than the difference of high brightness.

In other words, people are harder to distinguish the details in the dark than in the bright.

2.2 Advantage of Gamma Correction

Hidden message explosion of digital imaging technology is widely used in recent years. The technology has been applied to retrieve the blurred message from images on geographical environment, astronomy, natural phenomena fields. For an example on the images of lunar surface, it is hard to distinguish the details in Fig. 2(a), but it would be easy to identify after Gamma Correction in Fig. 2(b).

Normal Gamma correction makes watermark appear on image, but it often causes overexposure that is hard for human eyes to recognize the watermark. In this paper, we proposed a comprehensive visible watermarking scheme based on multi-parameters adjustable Gamma Correction (MPA-CVW) to improve the quality of embedded watermark image.

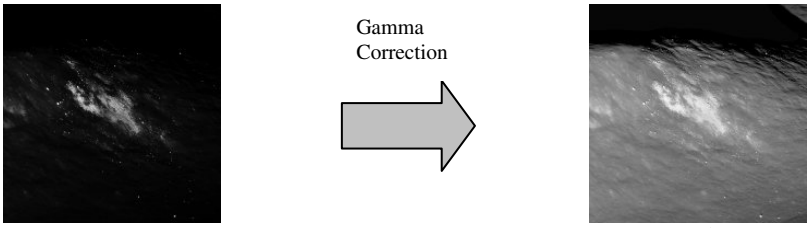


Fig. 2. An example of Gamma Correction application on lunar surface image. It is hard to distinguish the details in Fig. 2(a), but it would be easy to identify after Gamma Correction in Fig. 2(b).

3 Comprehensive Visible Watermarking Based on Multi-parameters Adjustable Gamma Correction

Comprehensive visible watermarking based on multi-parameters adjustable Gamma Correction (MPA-CVW) would change its background color to fit the embedded area. Multi-parameters adjustable gamma correction algorithm increases the saturation of the image and makes the watermark easier to be identified by human eyes. MPA-CVW is described as the following steps:

Step1. First, we do Gamma Correction on original image. Suppose that N_{input} and N_{output} are the strength of monitor input and output, respectively. Note that our experiments shown that the best constant r was 0.6 in Gamma Correction (Eq. 1) where c and r are constants.

$$N_{output} = c \times N_{input}{}^{r}. \tag{1}$$

Step2. We calculate the largest gradient i^* to select the area for embedding watermark.

$$i^* = \arg\max_{i} \nabla G(i) . \tag{2}$$

Step3. After Gamma Correction, we perform de-noising function D to obtain the image I'_j where δ_j is j times de-noising limited by δ_{man} and I_G is the image after Gamma Correction.

$$I'_j = D(I_G, \delta_j) , \ \delta_j \le \delta_{\max} . \tag{3}$$

Step4. We select the best location P for watermark embedding where has the maximum difference between each pixel and i^* on I'_j. The embedded region R is based on the location P and set the same block size with the watermark we will embed.

$$P = \arg\max_{x_0, y_0} \sum_{x=x_0}^{x_0+w_W-1} \sum_{y=y_0}^{y_0+h_W-1} | I'_j(x, y) - i^* | . \tag{4}$$

$$R = \{I'_j(x,y) \mid x_0 \le x \le x_0 + W_t - 1, y_0 \le y \le y_0 + W_w - 1\} \cdot \tag{5}$$

Step5. Suppose that T is a noise threshold. We continue performing de-noising again until the equation (Eq. 6) does not holds.

$$\begin{aligned} &\sum_{x=x_0}^{x_0+W_l-1} \sum_{y=y_0}^{y_0 W_w-1} C(I'_j(x,y)) \le W_l \times W_w \times T \text{ ,where} \\ &\sum_{x=x_0}^{x_0+W_l-1} \sum_{y=y_0}^{y_0 W_w-1} C(t) = \begin{cases} 1 & ,if\ t=0 \\ 0 & otherwise \end{cases} \end{aligned} \tag{6}$$

Step6. The maximum of *R*, *G*, and *B* value of each pixel in the selected region is set as index to adjust the saturation of the image. The two watermark strength α and β are adapted based on the index to be embedded on the cover image. However, if the added strength is over 255, subtraction is instead of addition in embedding operation.

$$\begin{aligned} &\tilde{I}(x,y) = \begin{cases} I'_j(x,y) - \omega & if\ (I'_j(x,y) - \omega) > 255 \\ I'_j(x,y) + \omega & otherwise \end{cases} \text{ ,where} \\ &\omega = \begin{cases} \alpha & if\ W(x - x_0 + 1, y - y_0 + 1) = 1 \\ \beta & otherwise \end{cases} ,and\ \alpha > \beta \end{aligned} \tag{7}$$

Step7. Convert the stego image $\tilde{I}$ from RGB color system into HSV system. The proposed multi-parameters adjustable gamma correction (MPA-GC) performs Gamma Correction based on hue γ_H, saturation γ_S, and value γ_v respectively. The range of value γ_H, and γ_S) are between 0.6 and 1 and γ_V is from 0.6 to 2. The stego image would be more saturated enough to make the watermark more integrated with the surrounding in a normal visual angle.

$$MPA - \tilde{I}_{output} = MPA - GC(\tilde{I}_{input}, \gamma_H, \gamma_S, \gamma_V) . \tag{8}$$

4 Experiment Results

This section describes the experiment results of MPA-CVW. Figures 3(a) and 3(b) were original image, and 3(c) was the binary watermark to be embedded. The Figures 3(d) and 3(e) were the stego images of 3(a) and 3(b) with the watermark embedded by using MPA-CVW, respectively. The embedded watermark was not binary because it changed its color to fit with the selected area. The PSNR is most commonly used as a measure of quality of watermark technology. The stego images embedded by MPA-CVW were high quality because the PSNR of 3(d) and 3(e) were 54.6 dB and 58.66 dB, repectively.

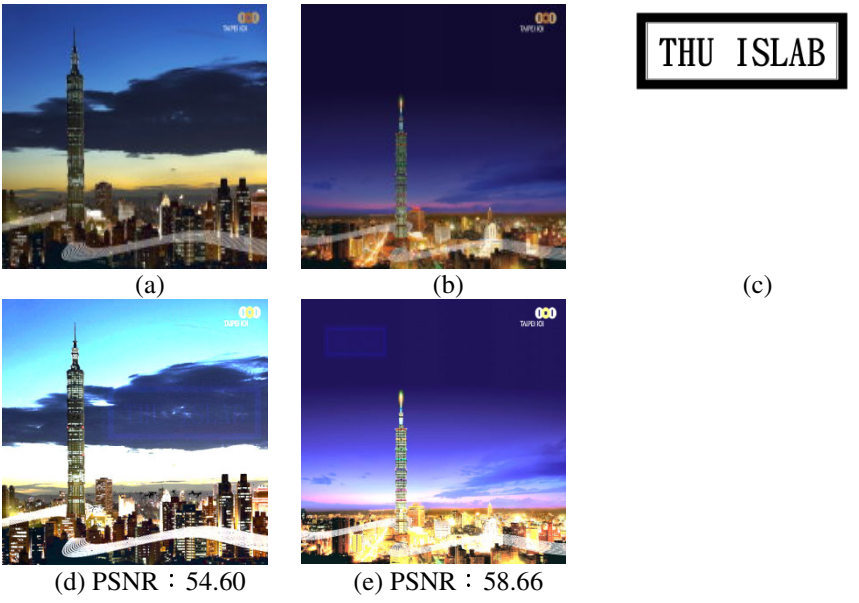


Fig. 3. (a) and (b) were original images with 400×400 pixels and 768×768 pixels. (c) was a binary watermark with 250×100 pixels. (d) and (e) were the stego images by MPA-CVW.

5 Conclusion

In the age of digital rights management (DRM), the applications of intellectual property right protections cover fields such as TV, broadcasting, publishing industries and etc. All the researchers in the fields tried to find a good technology to satisfy the both requirements: availability and security of contents. In practice, visible watermark is mainly applied to protect the image contents.

This paper proposed the an improved visible watermark scheme using multi-parameters adjustable gamma correction (MPA-CVW) to reduce the distortion, which can increase color saturation and is more integrated with the surrounding in a normal visual angle. Our experiments showed that the PSNRs of the stego images with watermark embedded were up to 54.50 dB and 58.66 dB.

Acknowledgments. This work was supported in part by Taiwan National Science Council under grant: NSC 99-2221-E-029-039 -MY3.

References

[1] Hu, Y., Jeon, B.: Reversible Visible Watermarking and Lossless Recovery of Original Images. IEEE Transactions on Circuits and Systems for Video Technology 16(11), 1423–1429 (2006)

[2] Tsai, H.-M., Chang, L.-W.: A High Secure Reversible Visible Watermarking Scheme. In: IEEE International Conference on Multimedia and Expo, pp. 2106–2109 (2007)
[3] Huang, H.-C., Chen, T.-W., Pan, J.-S., Ho, J.-H.: Copyright Protection and Annotation with Reversible Data Hiding and Adaptive Visible Watermarking. In: Second International Conference on Innovative Computing, Information and Control (ICICIC 2007), Japan, pp. 292–292 (2007)
[4] Mohanty, S.P., Guturu, P., Kougianos, E., Pati, N.: A Novel Invisible Color Image Watermarking Scheme Using Image Adaptive Watermark Creation and Robust Insertion-Extraction. In: Eighth IEEE International Symposium on Multimedia (ISM 2006), San Diego, pp. 153–160 (2006)
[5] Das, S., Bandyopadhyay, P., Paul, S., Ray, A.S., Banerjee, M.: A New Introduction Towards Invisible Image Watermarking on Color Image. In: IEEE International Advance Computing Conference (IACC 2009), Patiala, pp. 1224–1229 (2009)

A Tree Overlapping-Based Mesh Routing Protocol for Wireless Sensor Networks

Inwhee Joe, Yeonyi Choi, and Dongik Kim

Division of Computer Science and Engineering, Hanyang University
Seoul, 133-791 South Korea
iwjoe@hanyang.ac.kr

Abstract. In this paper, a mesh routing protocol is proposed to deliver sensing information effectively in a disaster environment. Many disaster application scenarios in the emergency network of wireless sensors require connectivity between nodes in order to transmit the collected data to a sink node. In case one critical connection is cut by accident, a tree routing structure has possibility of failure which results in delay or disconnection. In tree routing protocols, the packets follow the tree topology for forwarding the emergency data to the sink node, even if the node is located near to the source node. In the disaster site with the existing tree topology, the sensor network is not a very efficient network. In this paper, we present an enhancement of the tree routing protocol by implementing a mesh routing protocol based on tree overlapping. The concept of tree overlapping is to derive a mesh topology by overlapping all the tree structures in the sensor network where each tree structure is rooted at each sensor node. The simulation results show that the proposed algorithm provides shorter average end-to-end delay, reduces the number of hop counts, and also decreases the energy consumption compared to the conventional tree routing protocols.

Keywords: Tree overlapping, Mesh routing protocol, Topology control, Wireless sensor networks, Disaster environment.

1 Introduction

Recent advances in wireless communications and electronics for disasters prevention/detection have enabled improvements in low–cost networks through the use of various sensors. In a survey on sensor networks, Akyildiz et al. [1] reported results from a survey of protocols and algorithms, describing the design constraints and communication architecture of sensor networks. Willing [2] discussed the selection of promising and interesting research areas for protocols and systems for wireless industrial communications. Existing studies are difficult to apply to real-life situations because the existing networks for disaster and protection systems differ from those of research [2].

A wireless sensor network (WSN) is a system which can detect and manage environmental information, such as temperature, toxin concentration, and humidity, through a wireless sensor network. With developments in wireless communication

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 93–102, 2011.

technology, WSN systems are able to provide information on traffic, the climate, and the environment, all in real time. A WSN can change sensor network topology by inserting and eliminating sensor nodes [1] and can be ad hoc and self-organized [3,4]. In order to use this type of system for disaster protection in a disaster area, new network structures and routing protocols are required. Researchers in wireless sensor networking have suggested many multi-hop routing algorithms [5]. Existing routing protocols will be studied to discover which are best-suited for WSN. Related studies have been conducted on the design space of wireless sensor networks [6], but mesh routing on wireless sensor networks has many limitations regarding improvements in the reliability and robustness, power conservation, protocol complexity, and topology control. Routing protocol of wireless sensor network is generally the tree routing protocol of collection type, at the present. In general, tree topology is effective in terms of the fast information delivery to set up a network, but it is serious for the problem of reliability owing to the transmission failure at the disaster in the middle when there is a serious breakdown in communications and a loss in node because it is a lack of the flexibility of communication path. In addition, there is a problem in scalability and durability of network. Therefore in this paper, information is gathered in real time without delay in a disastrous situation, and based on it, we propose mesh routing protocol based on the tree overlapping to be able to react fast; data transmission, when a disaster happens with synchronousness to react promptly prompt, and with redundancy for being operated without a break for the prompt response in emergency situation. The purpose of our research is to develop the routing protocol that proposes mesh routing protocol based on tree overlapping and that monitor its information in real time in a disastrous situation, and that effectively copes with and prevents the situation. Ye et al. [7] reported results from minimum cost forwarding algorithm, is an algorithm for efficiently forwarding packets from sensor nodes to a base station with the useful property that it does not require nodes to maintain explicit path information or even unique node identifiers. In particular, Joe et al. [8] reported the results from an efficient bi-directional routing protocol, describing address notation and communication architecture of sensor networks. Researchers used the fixed address map 8 bytes in order to store the information of nodes in the routing table. Therefore the maximum number of nodes is 65535 in the routing table of the sensor network. And the size of the address map can be varied by 2 bytes if necessary.

2 Tree Overlapping-Based Mesh Routing Protocol

Tree routing algorithm realized on TinyOS has an advantage because it is simple and easy to calculate and manage, but it has also a problem to transmit data inefficiently via many nodes the message which can transmit by one hop because it transmits data considering only whole blood. This problem brings about severe loss of network, since transmission obstacle and its failure happen due to the destruction of sensor node in a disaster environment. We propose a tree overlapping-based mesh routing to solve this problem.

To solve the problem of path transmission failure according to data transmission and sensor node of destruction with ineffective transmission, we propose to decrease the transmission hop counts and end-to-end delay in this paper.

2.1 Mesh Network Node Types of Tree Overlapping

Tree overlapping Mesh Network is composed of three kinds of devices: mesh coordinator (MC), a Mesh Router Node (MRN), and a Mesh End Node (MEN). In IEEE 802.15.4-SPEAK, both MC and MRN are full-function devices, and an MEN is a reduced-function device. First of all, only one MC of the network components exists as the network coordinator. MC plays a role of initializing the network information and the role serves as the PAN coordinator in view of IEEE 802.15.4 standard. MC consisting of a FFD (Full Function Device) connects an external network through the gateway. The power supply into MC is easy, for it is connected with outside power. Second component of the MRN carries out the multi-hop routing within mesh network and several MRNs exist. MRN join the network both through mesh coordinator and through mesh routers already involved in the network. Function of MRN in the IEEE 802.15.4 standard serves as the role of the coordinator. The device type of mesh routers is composed of physical devices FFD. As the final component, MEN is an RFD (Reduced Function Device) device acting as the leaf of the network with limited functionality. It aims at working for the purpose of sensing data from the environment and transmitting to MRN, and is joined to the network through MRN. MEN can't accept any device to join the network and because of its limited energy it is changed to sleep mode to save its energy.

2.2 Mesh Routing Protocol

Cluster Setup

The first step forms a cluster tree structure based on the cost information of each sensor node, as shown in Figure 1. For the link cost, we evaluate the quality of communication path by using the concept of route cost, and consider, above all, low cost-optimal routing path [9]. Low cost means that small hop-count from root node and at the same time link quality are high. By using link quality indication (LQI) that we take when the algorithm to calculate link cost receives message frame of physical layer of IEEE 802.15.4 and MAC layer, we combine the distance of transmission path, the quality of it etc., and hop-count, and calculate synthetically them [10].

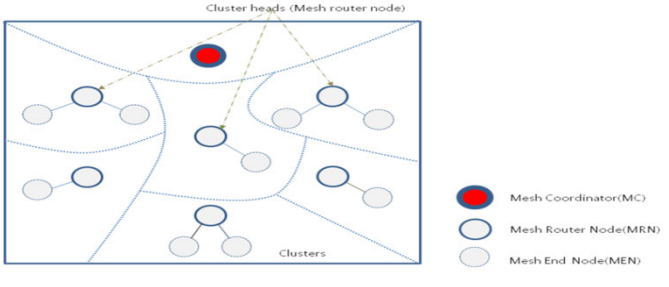


Fig. 1. Cluster tree formation

- Premise: After each MRN is initialized, and delayed random to escape crashes with other MRN nodes, beacon message is periodically broadcasted.
- Step 1: Child node, MEN connects on the network and receives beacon message to form a parent-child class system between MRN.
- Step 2: MEN receives beacon message and it checks link cost between MRN and itself. Link cost here, the main barometer of routing consists of EXT (Expected Transmission Count), LQI (Link Quality Indication), and RSSI(Receive Strength Signal Indication) [11]

$$\begin{aligned} &CC_P + C_L = CC_{LID} \\ &where \\ &CC_P : Parrent\ Current\ Cost \\ &C_L : Link\ Cost \\ &CC_{LID} : LocalID\ Current\ Cost \end{aligned} \tag{1}$$

- Step 3: The standard that MEN receiving beacon messages chooses parents the least cost value, namely, has a choice the best MRN as its parents in link condition considering the least calculated cost value,

Process of Tree Topology Structure

The formation of mesh network based on proposing tree overlapping is a process for making the communication between each MRN node possible and is progressed as following order. It forms tree topology of the central MRN each. It chooses parent-node and the process forming tree topology is as following. First, sink node, route MRN initializes cost value into zero, and broadcasts periodically to neighborhood MRN nodes. Neighborhood MRN receiving mesh beacon message, after calculating the total of cost value measured in the cost value received and 1-hop relationship. In this situation, chooses MRN node that has the least cost value and decides as parent-MRN. The tree topology formation process of each MRN centered node in whole network is as in the following, if optional MRN gets itself known as a root, MRN nodes are to form with route MRN as the center. Each MRN node at this moment calculates link cost to route set up. This process is repeated from root node to whole network. Through operating mode given above, if optional node in sensor network makes data, it transmits data to its parent-node. Through this process, MC plays a role in collecting data generated from all the nodes. If tree topology formation process centered each MRN node of whole network, gets it known, above all, each optional MRN in this network is its route, MRN nodes for routing.

Process of Tree Overlapping-Based Mesh Network Topology Structure

- Step 1: MRN transmit periodically mesh beacon message.
- Step 2: Neighbors MRN receive mesh beacon message from MRN which participate in network and put the least expense measured from MRN root and MRN address into beacon message, and transmit MRN list table.

- Step 3: Making each MRN including root number 0 as a root, cost value up to MRN node saves in mesh beacon message. Figure 2 shows tree topology as the center of MRN.
- Step 4: MC in the center overlaps tree topology which each MRN becomes root and consists, and completes tree overlapping-based mesh network topology as in Figure 3.

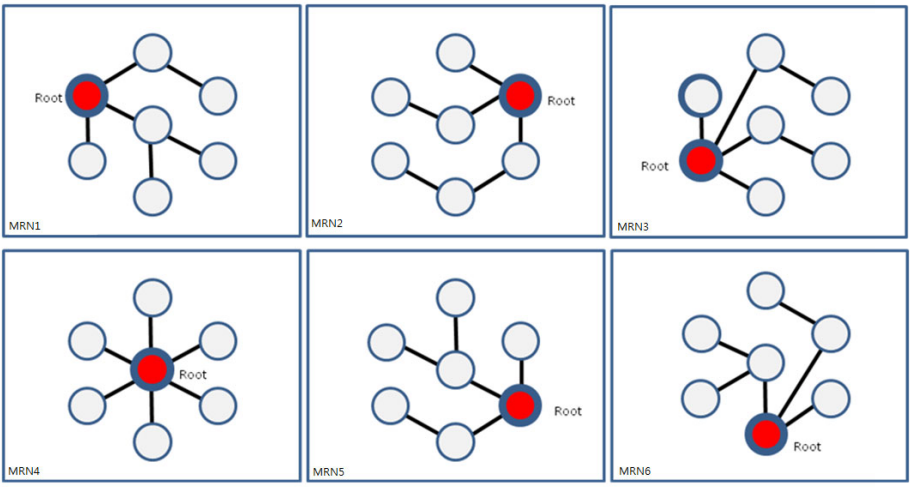
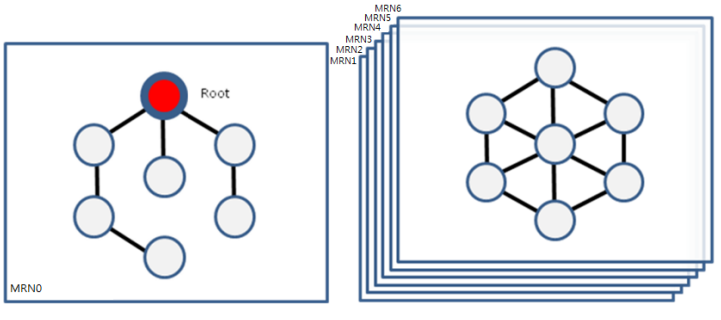


Fig. 2. MRN tree topology



(a) Tree topology (b) Mesh topology

Fig. 3. Root-based tree topology and tree-overlapped mesh topology

Mesh Routing Table

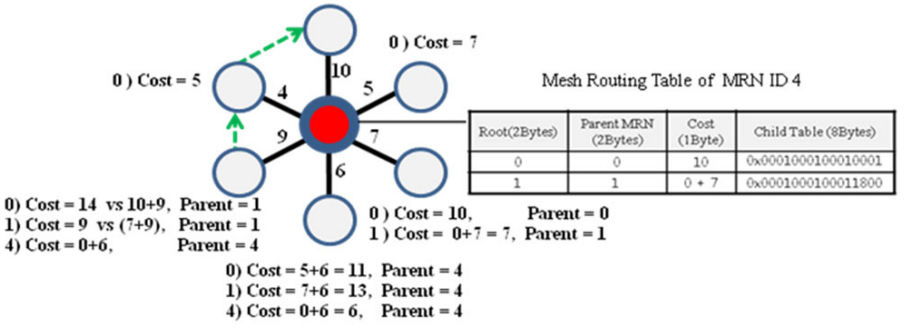
In the mesh routing table used in the proposed protocol, the routing information navigated to the present is stored. The mesh routing table format is shown in Table 1. It is to about the structure and field of routing table. Mesh router or coordinator exchanges routing information with other mesh routers (MRN) to generate routing path. When the routing information is exchanged, each MRN constructs mesh routing table as in Table 1 based on the routing information which is received from neighborhood MRN.

Table 1. Mesh routing table format

Filed Name	Data Length (Byte)	Description
MRN ID	1	Address of the source
Root	2	The node address of the origin of the packet.
Parent MRN	2	The address of the parent MRN node.
Hop Count	2	The number of network devices between the starting node and the destination node. Number of nodes traversed. Used for calculating route.
Cost	1	The measured cost of beacon messages received.
Child Table	8	The address of the child node list.

The management of mesh routing table is as following. Each router updates its own routing table and keeps up to date with received information being link state. For example, in the composition of mesh routing table, if MRN node number 0 broadcasts mesh beacon message, link cost with MRN number 0, root is measured at MRN node in 1 hop. MRN node in 1 hop that is, MRN number 1, 2, 4 receive mesh beacon message and measure cost value as in Figure 4. Then, MRN received beacon message, put root address and the least cost value up to root in beacon message, and send it. At this time, each MRN initializes cost value of root MRN into 0, and it constructs mesh routing table.

MRN number 4 transmit beacon message including mesh routing table information to neighboring node. Mesh beacon message broadcasts the lowest cost value in root number 0 MRN and in root number 1 MRN. At this moment, comparing its own mesh routing table value with the cost received, neighborhood nodes updates the lowest cost. The composition of mesh topology changes existing tree topology and turns every MRN participating in network into root, next makes tree topology, then decide parent MRN corresponding each root, finally, overlaps tree topology formed. If each MRN overlaps rooted tree topology, mesh topology is completed as in Figure 3(b).



Mesh Routing Table of MRN ID 4

Root(2Bytes)	Parent MRN (2Bytes)	Cost (1Byte)	Child Table (8Bytes)
0	0	10	0x0001000100010001
1	1	0 + 7	0x0001000100011800

Mesh Routing Table

MRN ID	Root (2Bytes)	Parent MRN (2Bytes)	Root Cost (1Byte)	Child Table (8Bytes)
1	0	0	0 + 5	0x0001000100010001
2	0	0	0 + 7	0x0001000100010001
4	0	0	0 + 10	0x0001000100010001

Fig. 4. Mesh beacon messages sent by MRN number 4

3 Performance Evaluation

In this work, we realize mesh routing protocol based on tree overlapping proposed in TinyOS [12]. We also compared and analyzed the performance about the average attrition rate of electric power, hop-count, and transit delay [13], comparing with existing tree routing protocol [14]. Test bed environment chooses TelosB affiliation as the platform of sensor network, and wireless module used CC2420 RF a transmitter-receiver and microprocessor used the sensor node loading MSP430 (8Mhz) at Texas Instruments (TI). Network size of test bed is 50m× 50m. Its component parts are also MC 1, MRN 6, and MEN nodes 9. We experimented 5 meters above ground with deck for smooth RF communication [15]. The protocol used in performance evaluation is used in tree routing based on the MulihopLQI protocol [16] realized at TinyOS 1.x and in mesh routing protocol based on tree overlapping at TinyOS 1.x.

In the first experiment we compared and analyzed performance evaluation whether how much the two routings has power consumption [15,17]. Figures 5 and 6 show the power consumption of proposed MEN. The method is to measure the serial multimeter and short-circuit with the sampling value of 50ms. As a result, as the average value of the power consumption 7.56mA was achieved. Compared with the figure 32.3mA of the previous tree routing protocol applied MultihopLQI, the proposed TOMesh has been improved to 425%.

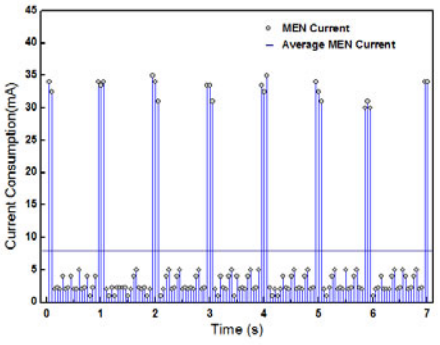


Fig. 5. Measured current consumption of the MEN

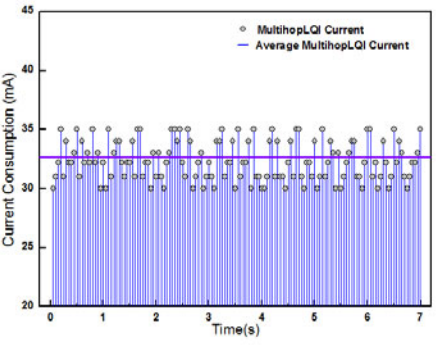


Fig. 6. Measured current consumption of the MultihopLQI

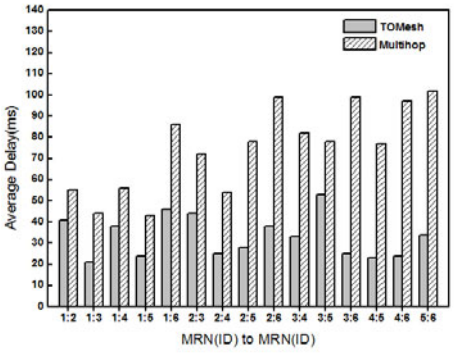


Fig. 7. The average delay between MRNs

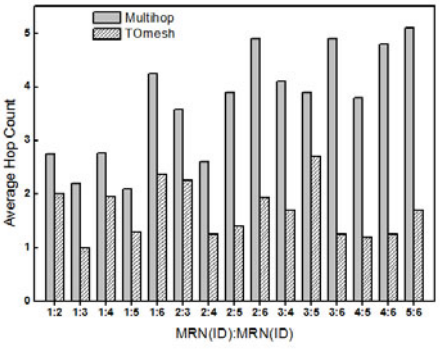


Fig. 8. The average hop-count between MRNs

In the second experiment we compared on the average transit delay and hop-count in communication between MRN. Figures 7 and 8 show that the farther are from sink node, more increased is the number of transmission-hop, so the average transmission delay was increased. As a result, if we use multi-hop routing protocol in this experiment, the farther it is in sink node, the more increased is the number of transmission-hop. As shown in Figures 7 and 8, the performance evaluation results of the proposed mesh topology (TOMesh) protocol are compared with those of the multi-hop tree (Multihop) protocol in terms of the average delay and average hop count. Our proposed mesh (TOMesh) protocol reduces average delay and average hop count. The enhanced algorithm is able to reduce the delay since the source MRN node by checking the most appropriate path to choose in transmitting the data packets to the destination node. In contrary multi-hop tree (Multihop) protocol sends all packets to the MC (mesh coordinator) then coordinator will transmit to destination node by following the long path which will increase the delay for transmitting the packets.

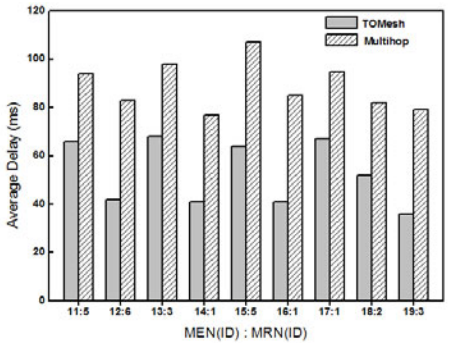


Fig. 9. The average delay between MEN and MRN

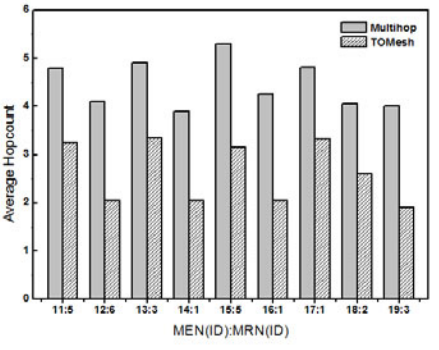


Fig. 10. The average hop-count between MEN and MRN

In the third experiment, we experimented data transmission between MEN and MRN and compared and analyzed the average transmission delay in Figure 9, the average hop-count in Figure 10 This experiment transmits data with MEN without routing function, sends with multi-hop more than average 2 hop than data transmission between MRN, and is about 20ms delay a hop. In addition, we found out that routing cost in TOMseh protocol in comparing tree routing protocol (Multihop) was remarkably low. Moreover, we found out that routing cost in TOMseh protocol was noticeably low as compared with tree routing protocol (Multihop). Protocol proposed compares the number of average hop per each node up to reaching destination. As a result, routing cost up to the destination node transmit the packet to the lowest neighborhood node. So routing cost was reduced. As a result, tree routing protocol (Mulihop) transmitted data via sink node, that is, takes not the shortest route but route and sent data to ineffective route via many nodes, and the average hop-count value and transmission delay was increased.

4 Conclusions

In this paper, we collected information in real time without delay in a disaster situation and proposed a mesh routing protocol based on tree overlapping to satisfy the ‘synchronousness’ that gives the immediate action based on the information and the ‘redundancy’ for operation without a break. In this paper, each MRN became root, formed tree topology, overlapped the topology, and constructed mesh network. Moreover, as the proposed mesh routing protocol transmits data through the best route, it is efficient and effective. The results of performance evaluation, the proposed routing protocol in end to end communication transmitted data more effective route than existing method, so it had effect that the number of average transmission hop decrease and PDT (Packet Delivery Time) was reduced. Also, paper proposed saved energy of each sensor node, and had 425% efficiency than the existing protocol in terms of the consumption of an average energy.

Acknowledgements. This work was supported by the ITRC Support Program (NIPA-2011-C1090-1121-0009) and the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (2011-0015681).

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, S.: A survey on sensor networks. IEEE Communications Magazine 40(8), 102–114 (2002)
2. Willig, A.: Recent and Emerging Topics in Wireless Industrial Communications: A Selection. IEEE Transactions on Industrial Informatics 4(2), 102–123 (2002)
3. Sohrabi, K., et al.: Protocols for self-organization of a wireless sensor network. IEEE Personal Communication 7(5), 16–27 (2000)
4. Mills, K.L.: A brief survey of self-organization in wireless sensor networks. Wireless Communications and Mobile Computing 7(7), 823–834 (2007)
5. Royer, E., Toh, C.: A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks. IEEE Personal Communications 6, 46–55 (1999)
6. Römer, K., Mattern, F.: The Design Space of Wireless Sensor Networks. IEEE Wireless Communications 11(6), 54–61 (2004)
7. Ye, F., Chen, A., Lu, S., Zhang, L.: A scalable solution to minimum cost forwarding in large sensor networks. In: Proc. IEEE 10th Int. Conf. Computer Communications and Networks, Scottsdale, Arizona, pp. 304–309 (October 2001)
8. Joe, I., Ahn, T.: An Efficient Bi-directional Routing Protocol for Wireless Sensor Networks. MobiQuitous, 1–4 (2007)
9. Kim, K.-H., Shin, K.: On Accurate Measurement of Link Quality in Multi-hop Wireless Mesh Networks. In: Proc. of the ACM MobiCom Conf., Los Angeles, CA, pp. 38–49 (September 2006)
10. Dyer, M., Beutel, J., Thiele, L.: S-XTC: A Signal-Strength Based Topology Control Algorithm for Sensor Networks. In: 2nd Int. Workshop on Ad Hoc, Sensor, and P2P Networks (March 2007)
11. Fonseca, R., Gnawali, O., Jamieson, K., Levis, P.: Four-Bit Wireless Link Estimation, `http://www.tinyos.net`
12. Tinyos 2 tep 123: The collection tree protocol, `http://www.tinyos.net/tinyos-2.x/doc/txt/tep123.txt`
13. Zhao, J., Govindan, R.: Understanding Packet Delivery Performance in Dense Wireless Sensor Networks. In: Proc. of the ACM Sen-Sys Conf., Los Angeles, CA, pp. 1–13 (November 2003)
14. Bose, P., Morin, P., Stojmenovic, I., Urrutia, J.: Routing With Guaranteed Delivery in Ad Hoc Wireless Networks. In: DIAL-M, pp. 48–55 (1999)
15. Texas Instruments: 2.4 GHz IEEE 802.15.4/ Zigbee-ready RF Transceiver (2007), `http://focus.ti.com/lit/ds/swrs041b/swrs041b.pdf`
16. MultiHopLQI (2004), `http://www.tinyos.net/tinyos-1.x/tos/lib/MultiHopLQI`
17. Zuniga, M., Krishnamachari, B.: An Analysis of Unreliability and Asymmetry in Low-Power Wireless Links. Transactions on Sensor Networks 3(2) (2007)

Performance Comparison among MIMO Techniques at Different Interference Levels for LTE

Mohammad T. Kawser, Md.K. Syfullah, Nawshad U.A. Chowdhury, and Md.T. Hoq

Department of Electrical and Electronic Engineering,
Islamic University of Technology
Dhaka, Bangladesh
mkawser@hotmail.com, khaled1107@gmail.com,
nawshad_ahmad@yahoo.com, tanbhir.hoq@gmail.com

Abstract. MIMO technique (multi antenna transmission and reception) is a key feature of long term evolution (LTE) which ensures higher spectral efficiency and better multiuser flexibility making a cellular system spectrally efficient but inherently densely populated and interference-limited. Performance of spatial multiplexing MIMO techniques degrades with the introduction of interference in the system. To choose optimum MIMO mode it is worthwhile to analyze the UE (user equipment) throughput under different interference scenarios before physical implementation of a system. This paper contains a comparative study of MIMO technique performances in terms of UE throughput under two different UE speeds and interference scenarios.

Keywords: SISO, Transmit Diversity, OLSM, CLSM, Round Robin, Interference, Mobility.

1 Introduction

The long term evolution (LTE) standard defined in 3rd generation partnership project (3GPP) is a revolutionary step towards next generation communication technology. LTE includes features like multicarrier channel-dependent resource scheduling, fractional frequency reuse, adaptive modulation and coding, advanced MIMO techniques and support for both FDD and TDD. LTE ensures higher spectral efficiency, lower delay and better multiuser flexibility compared to currently deployed networks [5].

LTE is advantageous in terms of efficient transmission and improved cell edge performance too. For efficient implementation of LTE, performance analyses of different parameters are worth investigating. In this paper, two different interference levels have been considered by varying the number of rings of eNodeBs around the serving eNodeB. Two different UE speeds have been considered; 1.25 m/s to represent a pedestrian and 25 m/s to represent a high speed vehicle. Section 2 explains different transmission modes. Section 3 contains simulation parameters. Section 4 compares the performance of transmission modes based on simulation results.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 103–110, 2011.

2 Transmission Modes

During dynamic resource scheduling, suitable transmission mode can be adapted semi-statically according to various channel conditions. The PDSCH channel employs different transmission modes utilizing multiple antennas in both transmitting and receiving sides. Till now nine transmission modes have been released but only first four have been implemented [2]. The nine transmission modes are:

1. Single antenna; port 0,
2. Transmit diversity,
3. Open loop spatial multiplexing,
4. Closed loop spatial multiplexing,
5. MU-MIMO,
6. Closed loop rank=1 precoding,
7. Single antenna; port 5,
8. Dual layer transmission; port 7 and 8 and
9. Up to 8 layer transmission; port 7-14.

2.1 SISO

SISO is used in transmission mode 1. It uses single antenna at the eNodeB. The data rate is the lowest compared to other transmission modes.

2.2 Transmit Diversity

Transmit Diversity (TxD) is used in transmission mode 2. Transmit diversity increases the signal to noise ratio at the receiver instead of directly increasing the data rate. Each transmit antenna transmits essentially the same stream of data and so the receiver gets replicas of the same signal. It improves the cell edge user data rate and coverage range. An additional antenna-specific coding is applied to the signals before transmission to increase the diversity effect. The transmit diversity is an open-loop scheme and feedback from the UE is not required. Transmit diversity is only defined for 2 and 4 transmit antennas and one data stream. The number of layers is equal to the number of antenna ports. Transmit diversity uses only one code word. The code word is mapped to two or four layers when there are 2 or 4 transmit antennas, respectively. To maximize diversity gain, the antennas typically need to be uncorrelated. So, the antennas need to be well separated relative to the wavelength or they need to have different polarization.

2.3 Spatial Multiplexing

Spatial multiplexing allows multiple antennas to transmit multiple independent streams. So it is sometimes referred to as the true MIMO technique. Instead of

increasing diversity, multiple antennas are here used to increase the data rate or the capacity of the system. Assuming a rich multipath environment, the capacity of the system can be increased linearly with the number of antennas when performing spatial multiplexing. In practice, the multiplexing gain is not that high because the SINR of the two parallel streams use little lower modulation and coding scheme.

Open-Loop Spatial Multiplexing (OLSM) is used in transmission mode 3. It makes use of the spatial dimension of the propagation channel and transmits multiple data streams on the same resource blocks. The feedback from the UE indicates only the rank of the channel using Rank Indication (RI) and not a preferred precoding matrix and hence, it is termed as open-loop. A channel-independent fixed precoding with large delay Cyclic Delay Diversity (CDD) is used. For high speed users, the PMI fails to indicate optimum use of precoding matrix and then the fixed precoding may achieve better performance.

Closed-Loop Spatial Multiplexing (CLSM) is used in transmission mode 4. In this case, the UE estimates the radio channel and selects the most desirable entry from a predefined codebook. Then the UE sends a feedback to the eNodeB and hence, it is termed as closed-loop. The UE sends the Precoding Matrix Indicator (PMI) to help the eNodeB select the appropriate precoding matrix from the codebook. The PMI can be optionally sub band based because the optimum precoding matrix can vary among resource blocks. The feedback can be optionally sub-band based because the optimum precoding matrix can vary among resource blocks. The preferred precoder is the matrix which would maximize the capacity based on the receiver capabilities. In interference free environment, the UE will typically indicate the precoder that would result in a transmission with an effective SNR by choosing most closely the largest singular values of its estimated channel matrix.

3 Simulation

Simulation was performed using LTE system level simulator v1.3_r427 [1]. A number of different UEs operating under a central serving eNodeB were used in the simulation. One or two rings of eNodeBs around the serving eNodeB were used for the purpose of creating two different levels of interference.

The same UE locations were used in all simulations and the UE locations are shown in the Fig. 1 and Fig. 2. The central eNodeB numbered 5 is the serving eNodeB when one ring of eNodeBs around the serving eNodeB was used. The central eNodeB numbered 11 is the serving eNodeB when two rings of eNodeBs around the serving eNodeB were used. The radiation pattern for one of the three cells from the eNodeB is shown in Fig. 3.

The simulation parameters are listed below:

Table 1. Simulation Parameters

Parameters	Assumptions
Frequency	2 GHz
Bandwidth	10 MHz
Transmission Mode	1, 2, 3 and 4
Number of TX antennas X Number of RX antennas	1×1 (Tx Mode 1) 2×2 (Other Tx Modes)
Simulation length	1000 TTI
Latency time scale	30 TTI
Inter- eNodeB distance	500m
Minimum coupling loss	70 dB
Macroscopic path loss model	TS 36.942
Macroscopic path loss environment	Urban
Inter-site shadow fading correlation	0.5
Intra site shadow fading correlation	1
eNodeB TX power	46 dBm
eNodeB power allocation	Homogeneous
The number of rings of eNodeBs around the serving eNodeB	1 and 2
UE thermal noise density	-174 dBm/Hz
UE speed	1.25 m/s and 25 m/s
Channel model type	PedB and VehA
Feedback channel delay	3 TTI
Maximum antenna gain	15 dBi
Scheduler	Round robin ensuring the best fairness for all UEs.
Subcarrier averaging algorithm	MIESM
UE receiver noise figure	9 dB

3.1 Figures

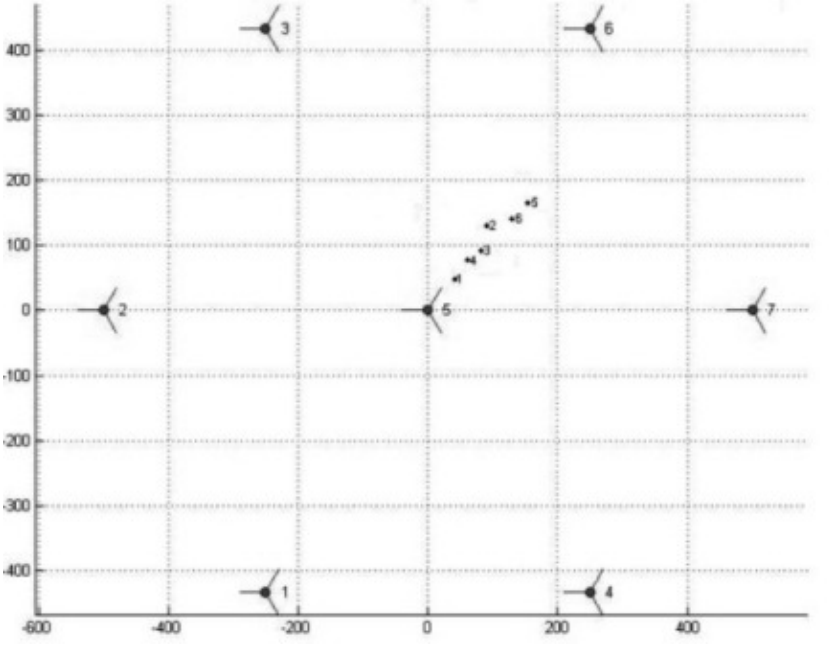


Fig. 1. UE position for one ring of eNodeBs around the serving eNodeB

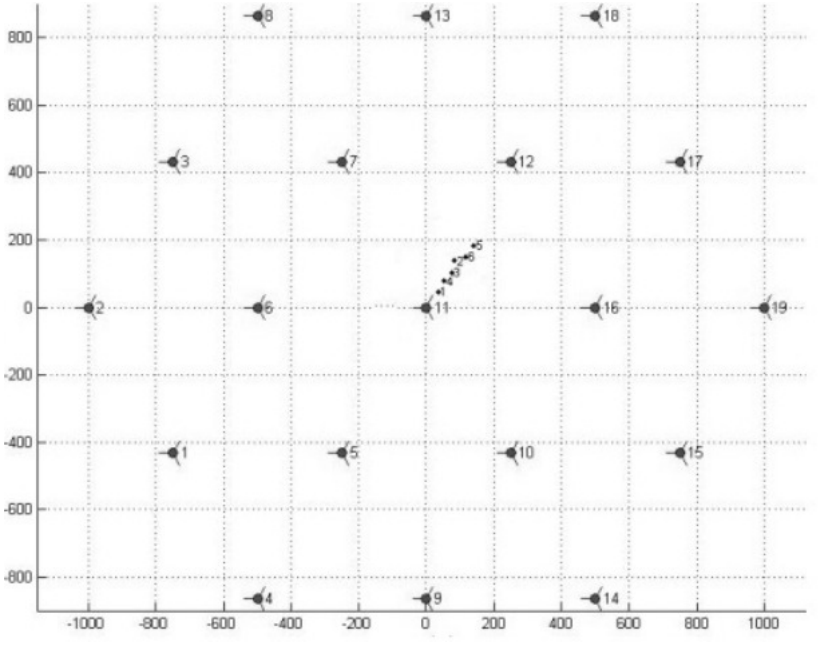


Fig. 2. UE position for two rings of eNodeBs around the serving eNodeB

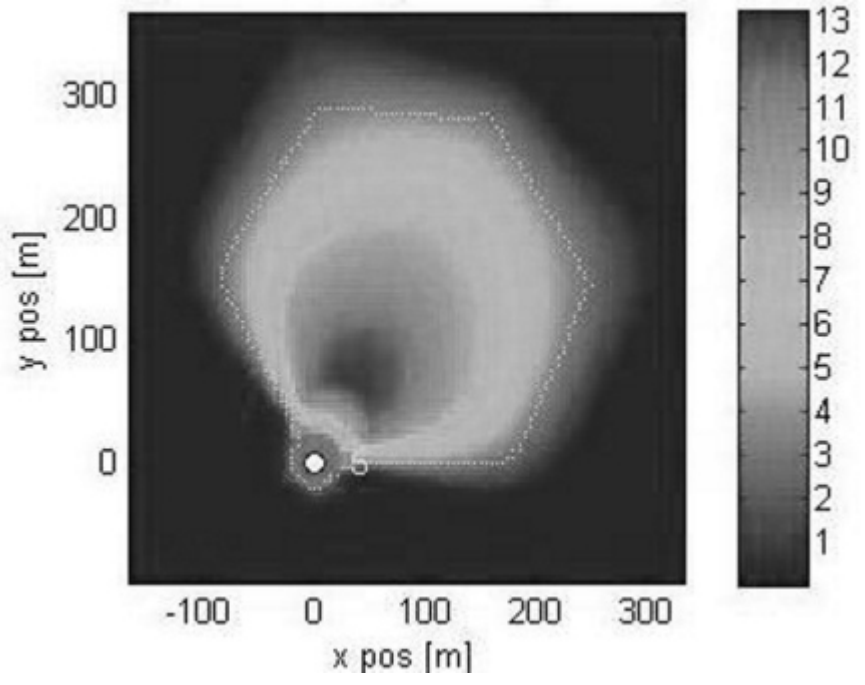


Fig. 3. Radiation pattern

4 Simulation Results

After execution of the simulations, the UE throughput is plotted against the distance of UEs form the serving eNodeB. The plot is shown in Fig. 4 for UE speed 1.25 m/s and in Fig. 5 for UE speed 25 m/s, respectively. The throughput falls off for greater distance of the UE because of increased path loss.

Fig. 4 and Fig. 5 demonstrate the amount of degradation in throughput for all transmission modes when interference from the second ring of eNodeBs around the serving eNodeB was introduced. SISO is generally found to have lower data rate compared to all MIMO techniques as expected. The transmit diversity achieves much better throughput compared to other transmission modes for UEs very far away from the serving eNodeB. CLSM performs better than OLSM in case of pedestrian but performs almost equally in case user in a high speed vehicle.

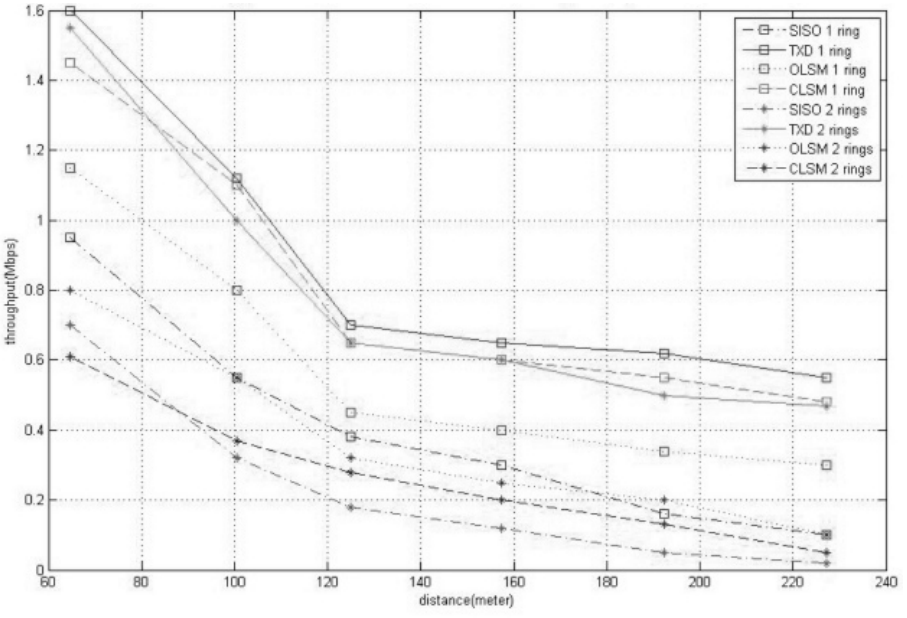


Fig. 4. Throughput for UE speed 1.25 m/s

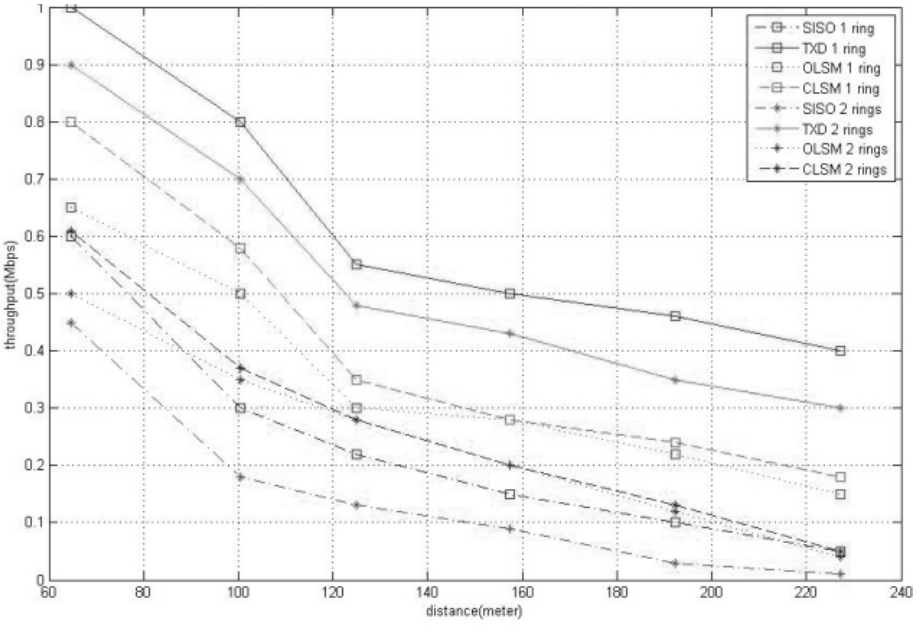


Fig. 5. Throughput for UE speed 25 m/s

5 Conclusion

In this paper, the performances of different MIMO Techniques under different interference levels and mobility have been analyzed. Transmit diversity achieves better performance for UEs far away because the signal combining technique in transmit diversity reduces fading variation and increases the signal-to-noise ratio at the receiver side. This provides for robustness of data transmission. When the UE speed was only 1.25 m/s, CLSM offered better throughput compared to OLSM. This is because the UE indicate the precoder that would result in a transmission with an effective SNR by choosing most closely the largest singular values of its estimated channel matrix in case of CLSM. However, CLSM had much more decline in throughput compared to OLSM when the UE speed was increased from 1.25 m/s to 25 m/s. This is because the Precoding Matrix Indicator (PMI) fails to indicate optimum use of precoding matrix for high-speed UEs and then the fixed precoding can achieve better performance. CLSM and OLSM were actually expected to provide better throughput than the results obtained. This poorer performance was probably due to the fact that system level simulator had not used very rich multipath environment.

References

1. LTE System Level Simulator v1.3_r427 by Institute of Communication and Radio Frequency Engineering, Vienna University of Technology, Vienna
2. Holma, H., Toskala, A.: LTE for UMTS - OFDMA and SC-FDMA Based Radio Access. John Wiley & Sons, Ltd. (2009) ISBN: 978-0-470-99401-6
3. 3GPP Technical Specification 36.213, Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures, http://www.3gpp.org
4. Ikuno, J.C., Wrulich, M., Rupp, M.: System level simulation of LTE networks. In: Proc. 2010 IEEE 71st Vehicular Technology Conference, Taipei, Taiwan (May 2010)
5. Sesia, S., Toufik, I., Baker, M.: LTE- The UMTS Long Term Evolution: From Theory to Practice. John Wiley & Sons, Ltd. (2009) ISBN: 978-0-470-69716-0
6. 3GPP Technical Specification 36.942, Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios, http://www.3gpp.org
7. Sanghyeon, L., Kyongkuk, C., Kwanghun, K., Dongweon, Y., Takkyu, K.: Performance Analysis of LTE-advanced System in the Downlink Spatial Channel Model. In: Proceedings of IC-NIDC 2009 (2009)

Handover Method Considering Power Consumption and Video Quality Satisfaction at the Mobile Node

Hyun Jong Kim and Seong Gon Choi*

College of Electrical and Computer Engineering, Chungbuk National University,
410 Seongbong-ro, Heungdeok-gu, Cheongju Chungbuk, Korea
{hjkim78,sgchoi}@cbnu.ac.kr

Abstract. In this paper, we propose a QoE-based handover method to consider mobile terminal power consumption and maintain experienced quality for a video streaming service like mobile IPTV. As a mobile terminal has dual mode interfaces with different power consumption and quality, a user can select an access network providing the satisfied quality and power efficiency. So, a mobile terminal needs the dynamic handover scheme between WLAN and alternative access network(3G) for better conditions. Through our simple experiment, existing RSSI-based handover schemes are not suitable to maintain the experienced quality of streaming multimedia service. Therefore we propose the QoE-based handover method through analysis of the relation between video experienced quality (MSSIM; Mean Structural Similarity Index).

Keywords: QoE, Handover, Dual mode, 3G, WLAN, power efficiency.

1 Introduction

Next generation wireless networks (NGWN) will be based on an all-IP based infrastructure with the support of heterogeneous access technologies. Thus, mobile users hope to handover across these IP-based heterogeneous wireless networks with no quality deterioration of multimedia services. There have been a lot of researches focusing on interworking issues between WLAN and cellular networks. However, there are not much researches available identifying the seamless handover solution on the mobile device by considering QoE(Quality of Experience) of multimedia services[1-2].

In these challenging scenarios, it is critical to guarantee an appropriate QoE (Quality of Experience) for the end user, according to the application to be developed. As various multimedia services are provided in the integrated wireless network environment, QoS(Quality of Service) and QoE concept is introduced in the IP network to describe satisfactions about end user's quality requirements. QoE can be defined as the overall performance of a system, from the user perspective. Many factors can affect the QoE, depending on the application and users expectations. Video perceptual quality is one of the most important aspects to consider in the user QoE.

* Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 111–119, 2011.

Existing RSSI threshold for handover is not suitable for the video streaming service because existing RSSI threshold is allocated for voice handover. New RSSI threshold for video streaming service and handover method is needed because video quality is influenced by packet re-ordering rate and packet loss rate unlike voice quality. Although service subscribers prefer to access WLAN, they want to have a video service of good quality using 3G(WCDMA) without the burden within the scope of arrangement fee. Therefore, we propose the QoE-based handover scheme for mobile video streaming service between WLAN and WCDMA.

The remainder of the paper is organized as follows. Section 2 outlines the QoS-based handover scheme on multi-RAT mobile terminal. Section 3 introduces the proposed QoE-based handover scheme on dual mode mobile terminal in integrated wireless network and section 4 describes implementation and results analysis of the proposed scheme. Finally, section 5 presents the conclusion and further study.

2 Related Works

This section covers QoS-based handover schemes between heterogeneous access networks and objective video quality measurement models.

2.1 QoS-Based Handover Schemes between Heterogeneous Access Networks

There have been a lot of researches focusing on interworking issues among cellular networks, WLAN and WiMAX(Worldwide Interoperability for Microwave Access). Owing to fast technology developments of wireless communication and mobile computing, the integration of 3G, WLAN and WiMAX is one of the most important issues[1]. In [2], they presented a novel mobile client architecture called HDEL(Handover Decision and Execution Layer) that provides seamless mobility while maintaining connectivity across heterogeneous wireless networks and provides better quality of service(QoS) support. To ensure QoS support during seamless handover, the proposed HDEL incorporates the procedures for activating a QoS session and ensures proper translation for network-specific QoS classifications.

Also, in [3] they proposed the handover scheme and algorithm that guarantee to simultaneously meet the three key QoS parameters, that is, the minimum data rate, the maximum data block delay, and the maximum bit error rate, for the arbitrary number of downlink and uplink multiservice connections. They defined the IEEE 802.11b network as the priority network for the mobile station. Thus, they maximize the time during which the mobility station is served by the IEEE 802.11b network while satisfying the QoS requirements in networks, as well as the maximum call-dropping probability and the maximum average number of ping-pong event constraints.

Another of Handover schemes between heterogeneous access networks have been proposed in [4, 5]. In [4], they proposed a movement-aware vertical handover algorithm between WLAN and Mobile WiMAX for seamless ubiquitous access for avoiding unnecessary handovers. And, in [5], they proposed soft handover mechanism according to the packet loss rates, based on the IPTV server and the consumer electronics devices. However, they consider only handover cost or packet loss rates to determine vertical handover, and do not consider QoS and QoE correlation of video streaming services.

Existing QoS-based handover schemes do not reflect changes in perceived video quality of service users because they only consider QoS parameters in network layer. Therefore, we propose the QoE-based handover scheme between heterogeneous access networks considering objective video perceptual quality.

2.2 Objective Video Quality Measurements

Initially Peak-Signal-to-Noise-Ratio(PSNR) was used as video quality metric because of its simplicity. However, it is well-know that PSNR does not necessarily accurately model perceptual quality. We compare video quality measurement approaches proposed in the recent years to select perceptual video quality measurement model for analysis of the proposed QoE-based handover scheme. The video quality metrics in this comparison include PSNR[6], VQM(Video Quality Metric)[7] and SSIM(Structural Similarity Index)[8]. This survey concludes with a comparison of these metrics in terms of computational complexity, correlation with subjective video quality measurement, and accessibility.

To numerically evaluate prediction of the objective metrics, we calculated Pearson's correlation coefficient between objective marks (after applying the fitting function) and subjective ones. Correlation coefficient belongs to the segment from -1 to 1 and reflects degree of dependency between values as shown in table 1. SSIM is the best objective metric as shown in table 1. We use SSIM metric to objectively compare and measure video quality, and verify the effectiveness of the proposed QoE-based handover scheme.

Table 1. Correlation between objective video quality metrics and MOS(Mean Opinion Score) subjective impression[9]

Metric	Correlation to MOS
PSNR	0.802
VQM	0.729
SSIM	0.937

3 Proposed QoE-Based Handover Scheme for Video Streaming Service

In the integrated wireless(WLAN and WCDMA) access network environment, although each of APs of WLAN does not provide sufficient RSSI service user and/or mobile terminal may prefer to use WLAN access network due to the cost and bandwidth. However, the weak RSSI of WLAN induces jerkiness and block distortion of video service. To achieve this, we propose the QoE based handover scheme in the integrated wireless network.

Fig. 1 shows the function block for QoE-based handover in a mobile terminal. Mobility policy function block carries out QoE-based handover between WLAN and WCDMA. Here, we may consider IEEE 802.21(MIH: Media Independent Handover) protocol for interworking, but this is out of scope in this paper.

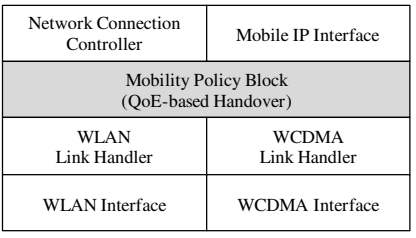
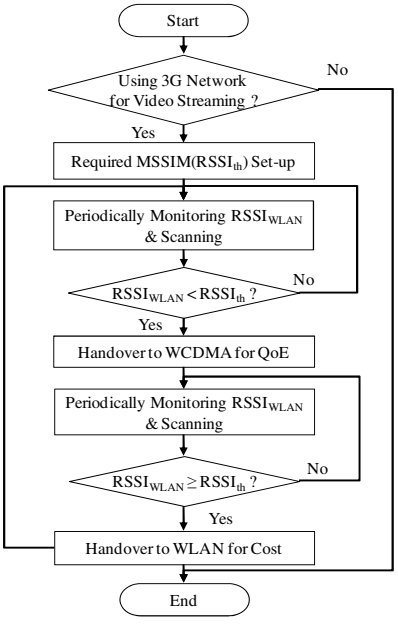


Fig. 1. Function block for QoE-based handover in mobile terminal

WLAN and WCDMA link handler periodically monitor RSSI through each interface, and delivery its results to mobility policy block. Mobility policy block compares measured RSSI with RSSIth which initially set up in a mobile terminal, and decides handover. If handover process is started, network connection controller deliveries new IP to media server and provides service continuity.



(a) The algorithm of the proposed QoE-based handover scheme

```
< Media Server>
begin:
ReceiveMsg()
  if MsgType == ServiceRequestMessage Then
    AllocateServiceID(ContentID)
    MediaStreaming(IP, Port, ServiceID, ContentID, TimeStamp)
  end
  else if MsgType == HandoverRequestMessage Then
    LookUpTable(UserID, ServiceID)
    Media_TransCoding(ContentID, TimeStamp)
    MediaStreaming(new_IP, Port, ContentID, TimeStamp)
  end

<Client>
begin:
Procedure Service Request
ServiceRequestMessage(UserID, Service ID, IP, Port)
Procedure Handover
while ServiceID > 0
  MultimediaStreaming()
  if RSSI_wlan < RSSIth Then
    WCDMA_flag == 1
    HandoverRequestMessage(UserID, new_IP, ServicePort,
ServiceID, ContentID, TimeStamp)
  end
  else if RSSI_wlan >= RSSIth Then
    WLAN_flag == 1
    HandoverRequestMessage(UserID, new_IP, ServicePort,
ServiceID, ContentID, TimeStamp)
  end
end
```

(b) The operation pseudo code of the proposed scheme

Fig. 2. The proposed QoE-based handover scheme for video streaming service between WLAN and WCDMA

Fig. 2-(a) shows the algorithm of proposed handover scheme. First, users setup required service quality level(MSSIM or $RSSI_{th}$) after they decide whether to use a 3G network or not. A mobile terminal periodically scans access network and monitors

RSSI of an access network. Mobile terminal carries out handover to WCDMA when measured RSSI of WLAN is lower than $RSSI_{th}$. Mobile terminal periodically measures RSSI of a WLAN after handover to WCDMA, and carries out re-handover to WLAN considering cost when RSSI provided by WLAN is higher than required quality level(or $RSSI_{th}$). Through this handover procedure, the improved multimedia service quality can be provided to users.

Fig. 2-(b) shows the media server and client operation pseudo code of the proposed QoE-based handover scheme for video streaming service between WLAN and WCDMA. Media server allocates content ID to distinguish serving service from the other services, and provides streaming service using received IP when client sends service request message with user ID, IP and port number. If RSSI of WLAN is smaller than RSSIth, client sends handover request message including user ID, new IP, service ID and content ID to media server. Media server can distinguish the using service from other services using service ID and content ID in handover request message, and provide continuous service using time stamp information and new IP.

4 Implementation and Result Analysis

We use network emulation package called NISTNET[12] to analyze the relation video quality and network quality parameters such as delay, jitter and packet loss. NISTNET is network emulation software for Linix® that allows a Linux server running as a router to emulate a variety of network conditions, such as congestion loss, packet reordering, or asymmetric bandwidth conditions. We configure test network as Fig. 3 to verify justification of the proposed scheme, and measure video quality using MSU video quality measurement tool[10]. We choose SSIM (structural similarity) score to accurately measure video experienced quality because SSIM is quite similar to MOS(Mean Opinion Score).

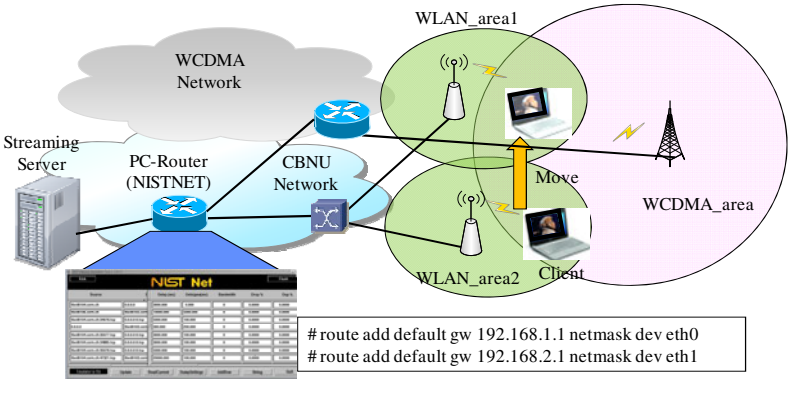


Fig. 3. Test network environment to verify justification of the proposed scheme

Since SSIM method can derive experienced video quality through measurement of similarity between original still images and tested still images, we can predict satisfaction of video quality passed through the experimental network using SSIM score. We make up media streaming server and client[11] for video streaming and set up de-jitter buffer size of client to 100ms. We experiment with video stream delivery in test network set to a certain delay, jitter and loss rate. We compare 2,700 original frames and 2,700 test frames passed through various network environments using full-reference method. In the test environment as shown in Fig. 3, mobile terminal receives and stores video streams to evaluate video quality while moving between WLAN_area1 and WLAN_area2. Here, WCDMA service area includes these two WLAN service areas. According to the results of experiments, we know that the mobile terminal is connected to WCDMA for approximately 15 seconds when using the proposed handover method.

Table 2 shows the video sources such as frame rate, resolution, codec and data rate for experiment. We use same frame rate, resolution and codec of the video sources used for experiment to analyze relation between RSSI of WLAN and video perceptual quality. However, we encode the video source using different data rate(783Kbps, 914Kbps and 1502Kbps) to analyze the impact of data rate of encoder.

Table 2. Video sources for experiment

No.	Frame rate	Resolution	Codec	Data rate
1	30fps	720*480	XVID	783Kbps
2	30fps	720*480	XVID	914Kbps
3	30fps	720*480	XVID	1502Kpbs

First, we experiment and analyze relation between RSSI of radio signal and video quality to allocate threshold(RSSIth) for QoE based handover method. Fig. 4 shows relation between RSSI of a WLAN and mean video quality(MSSIM: Mean Structural Similarity). As see in the Fig. 4, data rate of video steams does not influence video quality when bandwidth is sufficiently provided. However, we show that video quality sharply decline when RSSI of a WLAN is lower than -80dBm. New handover method and threshold for handover between wireless access points is needed to maintain video quality.

According to the test results, video quality(MSSIM) deteriorates to 0.3 by weak RSSI when the proposed HO method is not used. However, a mobile terminal can stably receive video streaming using the proposed QoE based HO method. Then, we know that video quality(MSSIM) is improved from 0.903 to 0.967. Like this, service users can use video multimedia service with improved quality when QoE based HO scheme is implemented on a mobile terminal in the integrated wireless access network.

Table 3 shows MSSIM(mean SSIM) score comparison per energy consumption of the test scenario. We know that enhanced perceptual video quality with lower power consumption can be achieved if we use the proposed QoE-based handover scheme when WLAN RSSI is weak. Available to the proposed method, we can get enhanced video QoE within limited energy of a mobile terminal.

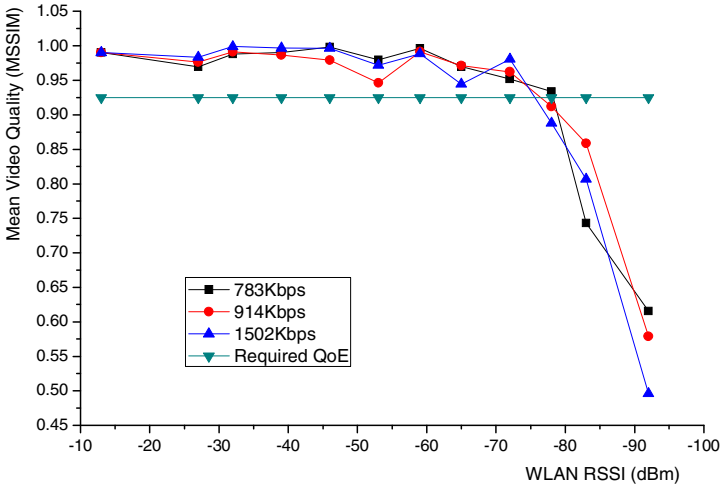


Fig. 4. Relation between RSSI of WLAN and mean video streaming quality(MSSIM)

Table 3. Mean SSIM per energy consumption of mobile terminal comparisons

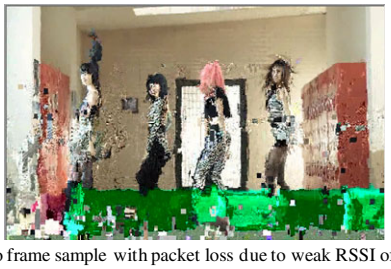
Test Scenario (2,700 frames)	MSSIM	Energy Consumption [KJ] [13]	Expected QoE against Energy Consumption
Only use WLAN	0.903	2.57	0.351
Only use WCMDA	0.977	4.22	0.232
Use the proposed scheme	0.967	2.84	0.341

Fig. 5 shows original video frame sample with data rate 1502Kbps, transmitted video frame sample with lower data rate 783Kbps and video frame sample with packet loss due to weak RSSI. Packet loss by weak RSSI causes block distortion and/or jerkiness of image. However, if the proposed QoE-based handover scheme is used between WLAN and WCDMA, the serious degradation of video quality can be prevented although image resolution is low.

Video streaming with 1502Kbps data rate has high packet loss rate in WLAN network, so serious degradation occurs as shown Fig. 5-(c) because bandwidth of the WLAN is greater than the WCDMA bandwidth. If video stream with lower 783Kbps data rate is transmitted in WCDMA network, the quality of video stream service can be improved by preventing packet loss although image resolution is decreased to 0.962 of SSIM score as shown Fig. 5-(b). Therefore, we can improve the video service quality by adaptively selecting stream data rate when the proposed QoE-based handover scheme is applied at mobile node and media server.

(a) Original video frame sample with data rate 1502Kbps

(b) The video frame sample with lower data rate 783Kbps



(c) Video frame sample with packet loss due to weak RSSI of WLAN

Fig. 5. Video frame samples according to streaming data encoding rate

5 Conclusion

We propose the QoE based HO method for video multimedia service in the integrated wireless access network. We know that the proposed method can provide enhanced QoE per power consumption between heterogeneous wireless networks. Various multimedia services such as video, voice and gaming service can be provided by mobile IPTV, and these services require QoE-based seamless handover for service quality satisfaction. At this time, video quality degradation like video jerkiness, block distortion and blurring is caused when RSSI of WLAN is weak on handover. Through the proposed HO method implementation and experiment, we show that service user can use multimedia service like mobile IPTV service by keeping the required quality level and considering cost and QoE. In future, we will analyze the correlation between various QoS parameters and video quality in wireless systems.

Acknowledgments. "This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0026214)".

References

1. Yang, S.-J., Chen, S.-U.: QoS-Based Fast Handover Scheme for Improving Service Continuity in MIPv6. In: Wireless Communications, Networking and Information Security (WCNIS) 2010, pp. 403–408 (June 2010)

2. Vijay Anand, S.: QoS based Handover layer for a Multi-RAT Mobile Terminal in UMTS and Wi-MAX Networks. In: COMSWARE 2008, pp. 472–479 (January 2008)
3. Garmonov, A.V., Cheon, S.H., Yim, D.H., Han, K.T., Park, Y.S., Savinkov, A.Y., Filiin, S.A., Moiseev, S.N., Kondakov, M.S.: QoS-Oriented Intersystem Handover Between IEEE 802.11b and Overlay Networks. IEEE Trans. on Vehicular Technology 57(2), 1142–1154 (2008)
4. Lee, W., Kim, E., Kim, J., Lee, I., Lee, C.: Movement-Aware Vertical Handoff of WLAN and Mobile WiMAX for Seamless Ubiquitous Access. IEEE Trans. on Consumer Electronics 53(4), 1268–1275 (2007)
5. Qi, Q., Cao, Y., Li, T., Zhu, X., Wang, J.: Soft Handover Mechanism Based on RTP Parallel Transmission for Mobile IPTV Services. IEEE Trans. on Consumer Electronics 56(4), 2276–2281 (2010)
6. ITU-T: Objective perceptual video quality measurement techniques for digital cable. ITU-T Recommendation J. 144, Geneva, Switzerland (March 2003)
7. Pinson, M., Wolf, S.: A New Standardized Method for Objectively Measuring Video Quality. IEEE Trans. on Broadcasting 50(3), 312–322 (2004)
8. Wang, Z., Lu, L., Bovic, A.C.: Video quality assessment using structural distortion measurement. Signal Processing: Image Communication, Special Issue on "Objective Video Quality Metrics" 19(2), 121–132 (2004)
9. Vatolin, D., Parshin, A., Petrov, O., Titarenko, A.: MSU Subjective Comparison of Modern Video Codecs. CS MSU Graphics & Media Lab Video Group (January 2006)
10. CS MSU Graphics & Media Lab Video Group, MSU Video Quality Measurement Tool Documentation (PRO version 2.7.1) (May 2010)
11. Fallon, H., Lattre, A.D., Bilien, J., Daoud, A., Gautier, M., Stenac, C.: VLC user guide. VideoLAN Project (2003)
12. `http://www.antd.nist.gov/tools/nistnet/index.html`
13. Balasubramania, N., Balasubramania, A., Venkataramani, A.: Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications. In: Proceedings of IMC 2009, pp. 280–293 (November 2009)

Shape Retrieval Combining Interior and Contour Descriptors

Solima Khanam[1], Seok-Woo Jang[2], and Woojin Paik[1,*]

[1] Dept. of Computer Science, Konkuk University,
322, Danwol-Dong, Chungju-Si, and Chungcheongbuk-do 380-701, Korea
{solima,wjpaik}@kku.ac.kr
[2] Dept. of Digital Media, Anyang University,
708-113, Anyang 5-dong, Manan-gu, Anyang-si, Gyeonggi-do, 430-714, Korea
swjang@anyang.ac.kr

Abstract. To represent and retrieve a shape, the interior as well as the contour features are very efficient. Fourier descriptor is one of the important global contour- based features to describe contour-based shapes. However, some images depend on complex interior details. In that case, only using contour information fails to give correct retrieval of a shape. On the other hand, interior-based descriptors are not applicable for all kinds of images. Therefore, in this paper, we represent our images in terms of two descriptors. Firstly, we describe centroid distance-based Fourier descriptor to represent the contour of a shape, and then, we describe the medial axis as a shape interior. Shape features are extracted locally by the medial axis while the Fourier descriptor extracts features globally. Using MPEG-7 database we have shown that proposed Shock graph-Fourier descriptor (SGFD) method is effective in increasing retrieval accuracy without increasing computational cost.

Keywords: Shape retrieval, Shock graph, Fourier descriptor, City block distance, Edit distance.

1 Introduction

In context-based image retrieval (CBIR), images are retrieved and classified based on their color, texture, and shape content. Where there is no information about color and texture; shape is an important cue. Some factors make the shape representation and retrieval task difficult. In the shape representation technique, 3-D real object is projected onto a 2-D plane which results in loss of one dimension information. Also, visual transformation of a shape is a challenging issue for shape matching. Therefore, for the purpose of shape representation and similarity measure, a shape retrieval can be generally classified into two class of methods: contour-based methods and interior-based methods. Accordingly, shape retrieval is based on whether shape features are extracted from the contour only or are extracted from the whole shape interior. Under each class, the different methods are further divided into local approaches and global

* Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 120–128, 2011.

approaches. This sub-class is based on whether the shape is represented as a whole or represented by segments [1].

Contour-based approaches are more popular than interior-based approaches in literature. This is because, human beings are thought to discriminate shapes mainly by their contour feature and in many applications, the shape contour is the only choices. In that case, global contour shape techniques take the whole shape contour as the shape representation. However, there are several limitations with contour-based methods. Firstly, contour shape descriptors are generally sensitive to noise and variations as they only use a small part of shape information. Secondly, in many cases, the shape contour is not available. Thirdly, in some applications, shape content is more important than the contour features. Considering those situations, some promising methods such as the moment methods, Fourier descriptor (FD), generic- Fourier descriptors (GFD) and wavelet-Fourier descriptors (WFD) techniques [1, 2, 3 and 4] are applied so far. On the other hand, the interior-based approaches show better performance than contour-based approaches in handling instability existing in an image database where partial matching is needed. Particularly, skeleton-based interior approaches show superiority to the contour-based approaches by providing topological and geometrical information as well as showing robustness against the visual transformations [5]. Main drawbacks of interior-based local approaches are failure to capture global features of a shape and computational complexity in similarity measurement. If the interior-based features connect to global shape features, then an interior-based method can cope well with shape defection which is a common problem for contour-based shape representation techniques. However, interior-based methods, like skeleton-based feature matching is more complex than contour-based methods. An extensive research has been performed [6, 7] to handle the complexity of similarity measure for skeletal-based method. Our recent work [8] also proposed some methods to reduce complexity of interior- based skeletal method.

Both interior and contour-based local or global methods have limited applications. Only using contour or interior information fails to give correct retrieval of a shape in some cases. Therefore, combination of contour and interior features is now-a-days a good choice as shape descriptors [9, 10]. In [9], it is indicated that shape-based trademarks need to be interpreted using different algorithms for global and local structure. With this strategy, the sample images can be compared separately to the local and global features of the query image. Curvature and centroid distance are used for describing the local feature while Zernike moments are employed to extract the global feature. This work [9] computes the first 4-order 15 Zernike moments to achieve the most effective measurement of the global shape. By utilizing the Euclidean distance to the range [0, 1] and setting up a threshold of 0.3, the similarity measurement is computed. However, in this paper we find some drawbacks. They did not consider the relation between the adjacent boundary points. Also, the presence of factorial calculations makes the computation of Zernike moment complex and an approximation. If an approximation method is used in solution, the retrieval accuracy cannot be guaranteed. Moreover, using a threshold value (without explanation) make the system database dependent. An effective solution of those drawbacks in [9] are proposed in the work of Heng et al. [10] by combining two shape features for representation and matching. In their work, contour-based descriptor includes the histogram of centroid distances and represents the relationship among two adjacent boundary points and the centroid. In the feature matching strategy, a statistics-based

method was proposed to compute the dissimilarity values between shape feature vectors of images. The combination of contour and interior-based descriptor may increase the computational time. However, in their study [10], the computational complexity is not analyzed. Also, comparison analysis with respect to other combined feature- based approaches is absent in [10].

Therefore, in our study, we have proposed an idea of combining shock graph-based interior feature with Fourier-based contour feature for shape retrieval. The proposed Shock graph- Fourier descriptor (SGFD) method does not affect the accuracy at a lower computational complexity. To the best of our knowledge, shock graph is a robust descriptor for shape retrieval. Although, shock graph was a computationally complex method, a less complex computation has been performed in our previous study [8]. Also, we do not experience any increment of complexity because of combining Fourier descriptor. Because, FD (Fourier descriptor) is one of the simplest ways for contour-based feature extraction [2, 3 and 4]. Moreover, the advantages of different FDs methods over many other contour-based shape descriptors are: simple to compute and normalization; each descriptor has specific physical meaning, capture both global and local features. The rest of the paper is organized as follows. Section 2 contains fundamental concepts of Fourier and shock graph descriptors. Method implementation and experimental results are discussed in Section 3. Finally, we draw some conclusions in Section 4.

2 Methods and Materials

Our method consists of two steps of feature extraction. Firstly, we will extract contour feature of a shape boundary and in the second step, shock graph-based interior feature will be extracted.

2.1 Fourier Descriptor as Contour Feature

Most FD-based works are dedicated to character recognition and object classification. The complex coordinates and the cumulative angle function are dominantly used to derive FD [3, 4]. However, Zhang and Lu [3] have found that for general shapes, the *centroid distance* function is the most desirable shape signature to derive FD. They have also found that 10 FD features are sufficient to represent shape and this is a significant reduction in the dimensions of FD where 60 FD features are usually used. Therefore, in our study, we will apply centroid distance-based Fourier descriptors (FDs) as shape contour features.

Fourier Descriptor. Fourier descriptors are derived from a *shape signature*. In general, a *shape signature,* f(t) is any 1-D function representing 2-D areas or boundaries. For a given shape, a shape signature can be defined as a closed curve *C* which in turn is represented by a one dimensional function f (t). The f (t) is complex and periodic at every time t. As f(t) is complex for period T, we have f(t+nT)=f(t), where 0<t<T. Different shape signatures have been used to derive FD [2]. In our experiment, we will use centroid distance-based (Fig.1 (A)) shape signature. To find the centroid distance-based shape signature, firstly, we need to extract the boundary

points of a shape contour. We assume that, the shape boundary coordinates (x(t), y(t)), t = 0, 1, …, N-1, have been extracted where, t usually means arc length. In our implementation, the shape boundary points are extracted through a 8-connectivity contour tracing technique [3]. Now, the centroid distance function is expressed by the distance of the boundary points to the centroid (x_c, y_c) of the shape and can be written by

$$f(t)=([x(t)-x_c]^2+[y(t)-y_c]^2)^{1/2}. \quad (1)$$

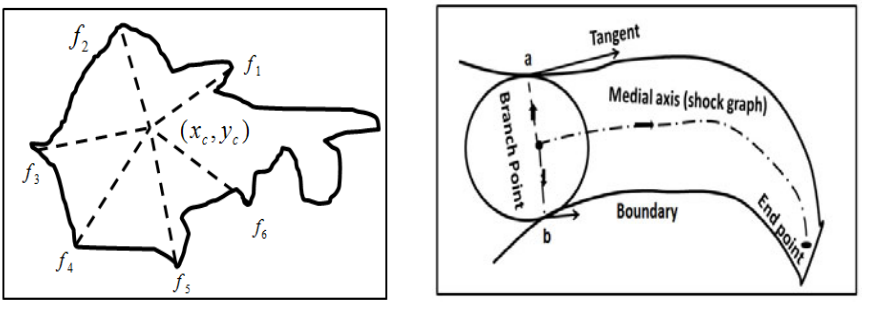
Consequently, using the shape signature the discrete Fourier transform can be defined by

$$F_n = \frac{1}{N}\sum_{t=0}^{N-1} f(t)\exp(-j2\pi nt / N), n \in Z. \quad (2)$$

The coefficients F_n, n=0, 1…, N-1, are used to derive Fourier descriptor of the shape [2].

2.2 Shock-Graph Descriptor

Medial axis is one of the important skeleton-based interior features which shows superiority to the contour-based approaches for shape retrieval [5]. Medial axis can be defined as the locus of the centers (called singularities or shock points) of maximal circles which touches the boundary at least at two points, as in Fig. 1 (B). The points which touches the boundary are referred to as characteristic points (*a* and *b* in Fig. 1(B)). Shock graph is an idea which arises from the concept of medial axis augmented with some additional dynamic properties. According to the type of tangency and the number of touching points on the boundary, a shock point can be of first, second, third or fourth order. The loci of all the shock points in Fig. 1(B) give the Blum's medial axis and also the idea of the whole shock graph [5]. The second and fourth order shocks are the generic cases of shock orders which are involved in occurring instability in shape retrieval. The second order shocks are the sources of flow while the fourth order shocks are termination points of flow, which represent branch and end points, respectively. These end and branch points will be used as interior shape descriptors in our proposed approach to represent and retrieve an image from a database.



(A) (B)

Fig. 1. Shape features; (A) centroid distance, (x_c, y_c) and (B) Segmented shape showing bitangent circle, medial axis, branch point, end point and characteriristic points (*a, b*)

At the starting of feature extraction, we adaptively select the nodes (shock points) of the images corresponding to the given query [8]. The time complexity of the selection algorithm is $O(x)$, where x is the number of nodes. For matching, we use shock point matching and edit operations in a discrete way. In this part, matching cost between shock graphs is measured by discretizing the graph into branch or end nodes. Let two segments of shock graphs be discretized at sample nodes. These nodes can be considered as elements of a matrix (m by n). Let $C(i, j)$ and $d([k, j],[l, j])$ be the matching cost of graphs and segments, respectively. Therefore, the cost of shock point matching $C(i, j)$ will be

$$C(i,j) = \min_{k,l}[C(i-k, j-l) + d([i-k,i],[j-l,j])]. \tag{3}$$

Here, the number of sub-problems solved by a dynamic programming algorithm is ($x \times y$), where x and y are the number of shock points (nodes) between graphs from the query and data image. Therefore, computing the matching cost, $C(i, j)$ needs complexity of order $O(xy)$. To deal with a visual transformation, we will consider deform cost resulting from the edit operation like contract and. splice cost [5]. The total cost will be the sum of the shock point matching cost and deform cost. The total matching cost, M_q can be written as

$$M_q = C(i,j) + D_q. \tag{4}$$

where $C(i,j)$, D_q and q represent matching cost, deform cost and the number of images, respectively.

3 Test Setup and Experimental Results

To verify the performance of the proposed combined shape retrieving algorithm, we tested it on the MPEG-7 CE Shape-1 Part-B dataset. We used this dataset without any kind of change to get the performance results in the presence of instability in the dataset. This image database contains 1400 images of 70 classes and 20 images are contained in each class. Images within the same class are considered to be similar. In our implementation, the shape boundary points are extracted through a 8-connectivity contour tracing technique [2]. It has been found from [3] that the increase of number of FDs over 60 does not significantly improve the retrieval performance. The actual retrieval performance does not changed significantly when the number of FDs is reduced to 10.

This means that for efficient indexing and retrieval, 10 FDs are sufficient for shape representation. So, in our implementation we have used 10 FDs to represent and retrieval of a shape. We use shock graphs as interior-based shape descriptor by applying the method described in Section 2.2. Adaptive algorithm [8] will be used to reduce the number of shock points. Figure 2 shows a shape of an apple image which is represented by medial axis with its boundary. Feature matching is conducted by measuring the distance between the feature vectors of the query image and the database images. Firstly, using city block distance [2] we get the similarity measurement for the images with respect to query images. In the second step, interior-based distance measurement is obtained by using shock graph-based edit distance [8].

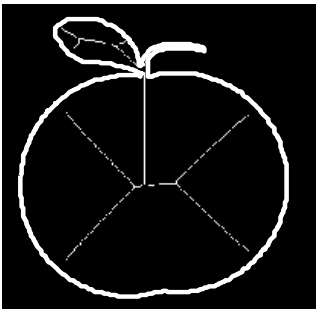


Fig. 2. An apple image with boundary and medial axis information in Matlab

Table 1. Shape retrieval using shock graph and Fourier descriptor method using MPEG-7 [11] shape images

	Query :	Apple	
Images	FD	SG	FD-SG
Apple	**27.9136**	**0**	**27.913598**
Apple	**55.8272**	**17**	**72.827195**
Apple	**83.74079**	**14**	**97.740793**
Apple	**111.6544**	**17**	**128.65439**
Apple	**139.568**	**23**	**162.56799**
Bat	167.4816	26	193.48159
Bat	195.3952	45	240.39518
Bat	223.3088	40	263.30878
Bat	251.2224	49	300.22238
Bat	279.136	41	320.13598
Beetle	307.0496	89	396.04957
Beetle	334.9632	90	424.96317
Beetle	362.8768	97	459.87677
Beetle	390.7904	113	503.79037
Beetle	418.704	124	542.70396

In Table 1, we see that the shock-graph distance gives good retrieval accuracy up to the 5th retrieval with respect to the query, an apple. Also, the contour-based Fourier descriptor gives an excellent retrieval performance for the apple query. Both of the two methods can effectively distinguish an apple images from the beetles or the bat images. Moreover, when we combine the two descriptors, the accuracy is not affected at all and all the apple images are retrieved successfully by FD-SG method.

However, there are some images which cannot be retrieved only by using a contour or interior-based features. In Table 2, we see that shock graph shows poor performance to retrieve bat images. Using shock graph, 3 bats are retrieved out of 5 while all the bats are retrieved by FD method. On the other hand, shock graph method gives good retrieval performance in retrieving beetles. However, FD method fails to distinguish between a bat and a beetle and no beetles are retrieved at all. In fact, shape images may have the same contour with different interior details. A standard database, like MPEG-7 contains all sorts of images and that is why only using interior or contour details may give erroneous retrieval performance. Therefore, during the feature matching stage, interior features and contour features should be compared separately in order to enhance the retrieval performance. It is a good choice to use the combination of both local interior and global contour feature simultaneously. The proposed FD-SG method shows an improved performance (Table 2), where only one descriptor fails to give correct retrieval.

Table 2. Shape retrieval by proposed Shock graph-Fourier descriptor (SGFD) Method using MPEG-7 database

	Query-Bat			Query -Beetle		
Images	FD	SG	FD-SG	FD	SG	FD-SG
Bat	**21.327509**	**0**	**21.327509**	**10.672622**	290	**300.67262**
Bat	**42.65502**	**30**	**72.65502**	**32.000132**	232	**264.00013**
Bat	**63.98253**	**44**	**107.98253**	**53.327642**	290	343.32764
Bat	**85.31004**	50	**135.31004**	**74.655152**	253	**327.65515**
Bat	**106.63755**	70	**176.63755**	**95.982662**	278	373.98266
Beetle	127.96506	290	417.96506	117.31017	**0**	**117.31017**
Beetle	149.29257	153	302.29257	138.63768	**166**	**304.63768**
Beetle	170.62008	82	252.62008	159.96519	**176**	335.96519
Beetle	191.94759	149	340.94759	181.2927	**189**	370.2927
Beetle	213.2751	64	277.2751	202.62021	**182**	384.62021
Bird	234.60261	**31**	265.60261	223.94772	223	446.94772
Bird	255.93012	**28**	283.93012	245.27523	207	452.27523
Bird	277.25763	137	414.25763	266.60274	193	459.60274
Bird	298.58514	135	433.58514	287.93025	207	494.93025
Bird	319.91265	55	374.91265	309.25776	236	545.25776

Table 2 shows that the retrieval accuracy increased for bat image applying the combined method (FD-SG). Moreover, this combination makes the distance larger between images of different groups. We understand that proposed method gives the robustness of retrieval and also, a course to fine matching which can further be used for image indexing.

In our implementation, we considered all the boundary points and for saving computation we applied fast Fourier transform [2, 3 and 4]. If the boundary points are sampled to the power of two, for example, 32 points or 64 points, then the sampling may result in a loss of boundary features. Therefore, all the points on the shape boundary are applied in our implementation. The computation of FD is $O(N)$, where N is the number of the boundary points. Also, using discrete shock graph for interior-based similarity measure [8] the complexity is $O(N^2)$. Therefore, for improving retrieval accuracy, we can take the benefit of combining contour-based Fourier descriptor (of low computational complexity) with shock graph-based interior descriptor. This combination gives better performance (Table 2) without increasing computational complexity. Therefore, combination of shock graph and Fourier descriptor (FD-SG) can be used as a general descriptor for an optimum retrieval performance.

4 Conclusions

To represent and retrieve a shape the interior as well as the contour feature is important. Because, some images depend on complex interior details while some other on contours. In this study, we have proposed an idea of combining shock graph-based local interior feature with Fourier-based global contour feature for shape retrieval. To the best of our knowledge, shock graph is a robust descriptor for shape retrieval. Also, Fourier descriptor is one of the important global contour-based descriptor to describe contour-based shapes perfectly. The proposed Shock graph-Fourier descriptor (SGFD) method does not affect the accuracy at a lower computational complexity. Using MPEG-7, we have shown that our method is effective in improving retrieval accuracy without increasing computational cost. A detailed comparative study of our approach is under investigation.

References

1. Zhang, D., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition 37, 1–19 (2004)
2. Zhang, D., Lu, G.: Study and Evaluation of Different Fourier Methods for Image Retrieval. Image and Vision Computing 23, 33–49 (2005)
3. Zhang, D., Lu, G.: A Comparative Study of Fourier Descriptors for Shape Representation and Retrieval. In: 5th Asian Conference on Computer Vision, Melbourne, Australia, pp. 23–25 (2002)
4. Yadav, R.B., Nishchal, N.K., Gupta, A.K., Rastogi, V.K.: Retrieval and Classification of Shape-based Objects Using Fourier, Generic Fourier, and Wavelet Fourier Descriptors Technique: A comparative study. Optics and Lasers in Engineering 45, 695–708 (2007)
5. Sebastian, T.B., Klein, P.N., Kimia, B.B.: Recognition of Shapes by Editing Their Shock Graphs. IEEE Trans. Pattern Anal. Mach. Intell. 26(5), 550–571 (2004)
6. Goh, W.-B.: Strategies for shape matching using skeletons. Comput. Vis. Image Underst. 110(3), 326–345 (2008)
7. Bai, X., Latecki, L.J.: Path Similarity Skeleton Graph Matching. IEEE Trans. Pattern Anal. Mach. Intell. 30(7), 1282–1292 (2008)

8. Khanam, S., Jang, S.W., Paik, W.: Fast and Simple 2D Shape Retrieval Using Discrete Shock Graph. IEICE Trans. Inf. & Syst. 94(10), 2059–2062 (2011)
9. Chia-Hung, W., Yue, L., Wing-yin, C., Chang-Tsum, L.: Trademark Image Retrieval Using Synthetic Feature for Describing Global Shape and Interior Structure. Pattern Recognition 42, 386–394 (2009)
10. Heng, Q., Keqiu, L., Yanming, S., Wenyu, Q.: An Effective Solution for Trademark Image retrieval by combining Shape description and Feature Matching. Pattern Recognition 43, 2017–2027 (2010)
11. MPEG-7 shape image database, `http://www.imageprocessingplace.com/root_files_V3/image_databases.htm.Loudon`

Hardware Architecture of Bilateral Filter to Remove Haze*

Eun-Kyoung Kim[1], Jae-Dong Lee[1], Byungin Moon[2], and Yong-Hwan Lee[1,**]

[1] School of Electronic Engineering, Kumoh National Institute of Technology, Gumi, Korea
{roruca,yofule,yhlee}@kumoh.ac.kr
[2] Gyeongbuk National University, Daegu, Korea
bihmoon@knu.ac.kr

Abstract. In this paper, we propose the hardware architecture of bilateral filter that can smooth image while preserving edges for the real-time image processing. Bilateral filter is nonlinear combination of nearby image values. As the time complexity of bilateral filter is O (total pixel size $*$ filter kernel size ^{2}), the real-time processing is difficult. We reduce the time complexity of bilateral filter using parallel architecture to achieve real-time performance.

Keywords: bilateral filter, soft matting method, haze removal, dark channel.

1 Introduction

Image processing system is used to save an image or detect the object of the image. Generally, the system processes the image with the assumption that the input image is clear. However, in real environment, the quality of the image is worse due to the haze, fog, and smoke. Thus, to improve the quality of the image, the several techniques have been widely researched [1]-[5]. Moreover, it is difficult to achieve the real image of the objects due to the unpredictable factors such as light.

After the image is filtered, the noise is removed and the edge is emphasized but this process often involves the unwanted noise. In this paper, a bilateral filter is proposed to preserve the edges of the image and to reduce the noise.

The rest of the paper is organized as follows. We first briefly review the method of removing the haze in section 2. Section 3 explains the hardware design of bilateral filter. We describe the performance evaluation of haze removal system in section 4 and finally section 5 presents concluding remark.

2 Method of Removing the Haze

The halo effect occurs in image when the haze remove using dark channel. Therefore, filtering needs to reduce the halo effect. Soft matting method and bilateral filter method

* This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the CITRC(Convergence Information Technology Research Center) support program (NIPA-2011-C6150-1102-0011) supervised by the NIPA(National IT Industry Promotion Agency).

** Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 129–135, 2011.

are used as filtering algorithms. The advantage of soft matting method is less execution time and bilateral filter method uses less amount of memory. These methods are evaluated to decide which one is more suitable for real-time processing.

2.1 Dark Channel Prior

The dark channel prior is defined as follows. The dark channel prior is the smallest value in RGB value of a pixel. J is observed image, J^{dark} is dark channel of J. The Equation(1) is defined as follows.

$$J^{dark}(x) = min_{c\in\{r,g,b\}}(min_{y\in\Omega(x)}(J^c(y))) \tag{1}$$

Where, J^C is color channel of J, $\Omega(x)$ is the center of patch. The haze in image have the high brightness and it shows Fig.1. The transmission map of haze image obtains by calculated dark channel. Using the transmission map remove the haze.

Fig. 1. Dark channel prior of haze image

2.2 Soft Matting Method

Soft matting method can eliminate the block effect but the size of matting Laplacian matrix by using Equation(2) is proportional the squares of the number of pixels. The Equation(2) is proposed by Levin[1].

$$\sum_{k|(i,j)\in w_k}\left(\delta_{ij} - \frac{1}{|w_k|}\left(1 + (I_i - \mu_k)^T\left(\sum_k + \frac{\epsilon}{|w_k|}\right)^{-1}(I_j - \mu_k)\right)\right) \tag{2}$$

Where, I_i and I_j are the input image pixel values of the i^{th} and j^{th}, respectively.

δ_{ij} is the value of Kronecker delta. μ_k is the average of window w_k. $\sum_k$ is the covariance value of window w_k. $|w_k|$ is the number of pixels of window w_k. Thus, the size of the equation (3) is used to calculate the size of the matrix in consideration with the rare matrix.

$$\text{size(L)} = \left(\left\lceil \frac{n^2}{2} \right\rceil + \frac{n}{2}\right)(w-2) * (h-2) \tag{3}$$

Where, n is the number of pixels of window w_k in the Equation (2). w and h are the horizontal and vertical pixel number of the image, respectively. The size of the image is set to 35×10^6 for XGA(1024 x 768) and 14×10^6 for VGA(640 x 480). However, the required size of the memory for the calculating matting Laplacian matrix is three times of size in Equation(3) in order to store the index information is stored in the rare matrix.

2.3 Bilateral Filter Method

By Tomasi and Manduchi, the bilateral filter was proposed to preserve the edges of the image and to reduce the noise [2]. The equation (3) shows that the bilateral filter is the convolution of the function using the weights of Gaussian on the difference between the values of pixels in the image around.

$$\text{t(x)} = \frac{1}{w(x)} \sum_{y \in \Omega(x)} G_{\sigma_s}(\| x - y \|) G_{\sigma_r}(|I_x - I_y|) \hat{t}(x) \tag{4}$$

Where, G_{σ_s} and G_{σ_r} are Gaussian functions, Ω is the spatial domain, I is the input image. $|I_x - I_y|$ means the difference between pixels around the location at which filter applies. $\| x - y \|$ means at the location applied Filters of pixel values between distance**.** w(x) is the function for normalization. To achieve the property that the bilateral filter is a weighted mean, we assign a weight W = 1 to the input value and the sum of the weights is set to the pixel x.

$$\text{w(x)} = \sum_{y \in \Omega(x)} G_{\sigma_s}(\| x - y \|) G_{\sigma_r}(|I_x - I_y|) \tag{5}$$

3 Hardware Design of Bilateral Filter

3.1 Structure

In this paper, the size of designed bilateral filter is 15 x 15 and a bilateral filter formula uses Equation (3) and (4).

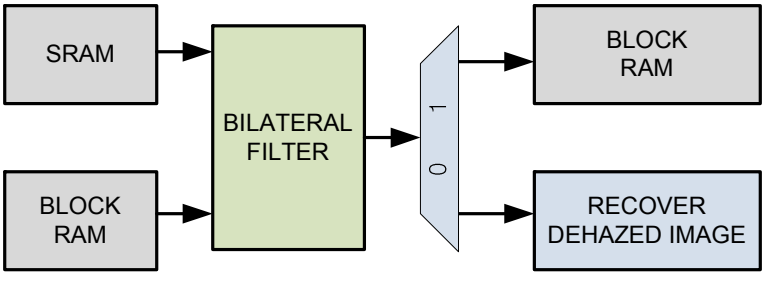


Fig. 2. Refined transmission map using bilateral filter

An original image and an estimated transmission map are read from SRAM and BLOCK RAM, respectively. It is refined from bilateral filter PE(process element) array. Refined transmission map are stored BLOCK RAM for repetitive refinement or If it has final stage of refinement, the pixel that is remove the effect of the final haze shows in the block to remove haze of image.

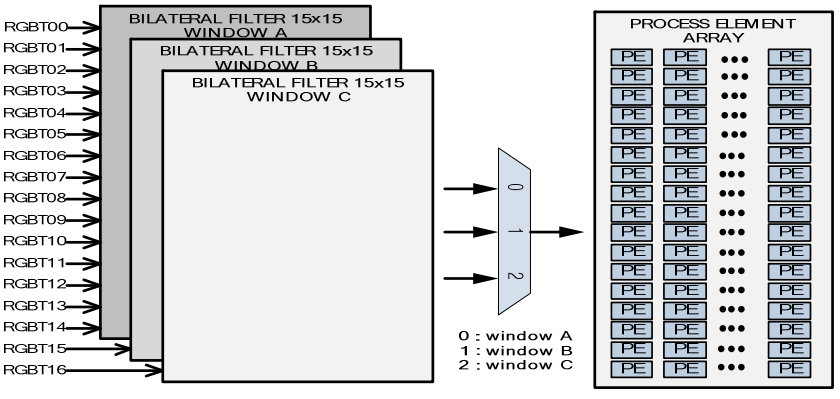


Fig. 3. Window form of bilateral filter

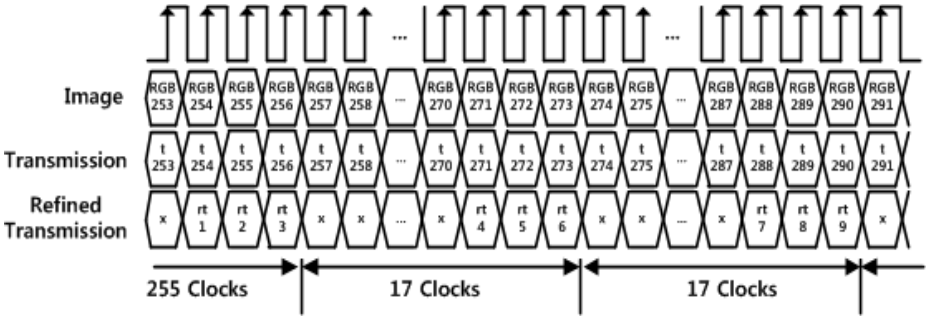


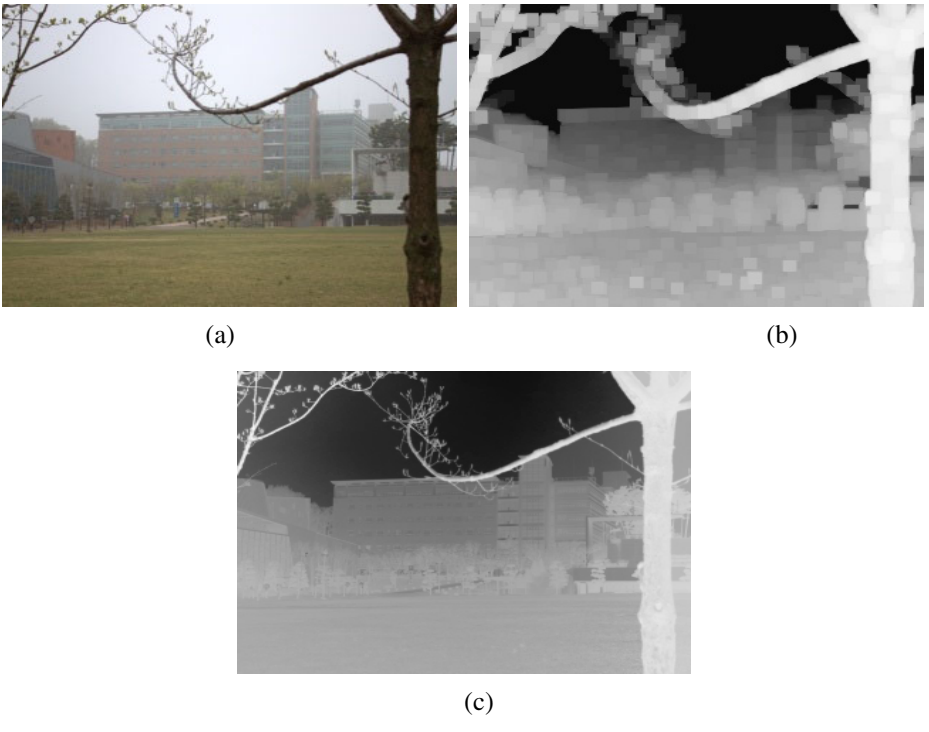
Fig. 4. I/O timing of bilateral filter PE array

To process the calculation on the bilateral filter in real time, the values are calculated by shifting the window whose size is 15 x 15. When boundary rules are ignored, I/O timing in Fig. 4 shows for the shift motion of bilateral filter window.

To output the first filtered data, window starts to move. At that moment, the 15 data should be sequentially inserted. After that, the calculated value is showed by window A. The window switching movement is changing from window A to window B after the insertion in 16^{th} data and the calculated value printed by bilateral filter window B. The data are inserted in 17^{th} print calculated value by window C in the same manner.

4 Performance Evaluation of Haze Removal System

A bilateral filter uses less memory than the soft matting method. However, the bilateral filter appears to perform once on the screen as shown Fig. 5(c). It does not perform as bilateral filter that is disappeared the halo effect than Fig. 5(b) but it is seen somewhat to remain. To eliminate the effect of halo, the bilateral filter operates several times to make the boundary line more clear. However, the operating time is continuously increased by hysteresis effect.



(a) (b)

(c)

Fig. 5. (a) Original image (b) Image after execution of dark channel prior Bilateral filtering of (b)

Table 1 shows the simulation results of the proposed hardware with previous work that uses soft matting method [2]. The proposed hardware was simulated by Matlab. The soft matting method needs more operation time compared to the one period of the bilateral filter but since the bilateral filter need to operate at least a couple of times to eliminate the halo effect. Therefore the soft matting method is more effective than the bilateral filter. However, the memory size of the soft matting method is 13 times as large as the bilateral filter. Thus, soft matting method is not suitable for the design of hardware.

Table 1. Transmission map method compare

	Soft Matting	bilateral filter
Resolution	320 x 240	320 x 240
execution time	13148ms	11290ms(1st refinement)
used of memory	89.5M	6.6M

The proposed design method which is based on parallel processing architecture has 3.7 times fast on the time of the transmission map and compared with the method that uses only software. Furthermore, the time of refining the transmission map can be reduced with bilateral filter by 1/14 compared to the software as shown Table 2.

Table 2. Computing speed compare

	Software implementation	Hardware implementation
Resolution	320 x 240	320 x 240
patch size of local minimum	9 x 9	9 x 9
patch size of bilateral filter	15 x 15	15 x 15
system requirements	Dual Core 2.67GHz RAM 4GB	Virtex4-LX60-10C operating frequency 40MHz
creation time of the first transmission map	37.7ms	10ms
refine time of one time transmission map	214.9ms	15ms
refine time of three times transmission map	658.5ms	45ms

5 Conclusion

Images of outdoor scenes are usually degraded by the turbid medium (e.g., particles, water-droplets) in the atmosphere. Haze, fog, and smoke are such phenomena due to atmospheric absorption and scattering. In this paper, The halo effect occurs in image when the haze remove using dark channel. The bilateral filter is used to improve the performance of removing haze. The soft matting method needs less execution time and the bilateral filter needs less used of memory. Therefore, parallel architecture of bilateral filter should use for the real-time image processing.

Acknowledgments. The authors would like to thank to IDEC for supporting EDA tools for this work.

References

1. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. In: CVPR, vol. 1, pp. 61–68 (2006)
2. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: VPR, pp. 1956–1963 (2009)
3. Sun, K., Wang, B., Zheng, Z.: Fast Single Image Dehazing Using Iterative Bilateral Filter. IEEE, 978-1-4244-7941-2 (2010)
4. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV, pp. 839–846 (1998)
5. Lee, J.-D., Kim, J.-S., Lee, J.-H., Kwon, S., Moon, B.-I., Lee, Y.-H.: Design of sun of Absolute Differences Based on Shifting Window 14(1), 825–827 (2010)

An Efficient Interworking Architecture of a Network Processor for Layer 7 Packet Processing

Kyeong-ryeol Bae[1], Seung-Ho Ok[1], Hyeon-Sik Son[1], Sang Yoon Oh[2], Yong-Hwan Lee[3], and Byungin Moon[4]

[1] School of Electrical Engineering & Computer Science, Kyungpook National University, Daegu, Korea
{puris1,wintiger,soc_shs1984}@ee.knu.ac.kr
[2] Omniflow Processor Research Team, ETRI, Daejeon, Korea
ohsy@etri.re.kr
[3] School of Electronic Engineering, Kumoh National Institute of Technology, Gumi, Korea
yhlee@kumoh.ac.kr
[4] School of Electronics Engineering, Kyungpook National University, Daegu, Korea
bihmoon@knu.ac.kr

Abstract. This paper presents a new interworking architecture for a network processor (NP) that is able to process packets from OSI layer 2 (L2) to layer 7 (L7) by combining a conventional NP with a general-purpose processor (GP). In general, most commercially available NPs could not afford to support a variety of network services. This is mainly because the conventional NPs are not able to process L7 packets. Thus, one of the most important requirements for the state-of-the-art NP is the ability to process packets of L2 to L7. To process L7 packets efficiently through both the conventional NP and GP, the proposed interworking architecture uses a deep packet inspector (DPI) and it controls the packet processing flow depending on the OSI layers of packets. Experimental results show that the proposed interworking architecture is not only able to process packets of L2 to L7 but also increase the throughput and load balance of the packet processing in the NP without large hardware overhead when compared with the conventional interworking architecture.

Keywords: Network processor, Interworking Architecture, DPI, Layer 7 packet processing.

1 Introduction

In recent years, there has been a tremendous growth in Internet traffic leading to increasing network bandwidth requirements and demanding high-speed packets processing capabilities on network devices such as routers and switches. Most commercial backbone network routers on OC-768 links (40Gbps) have to process up to 40 million packets per second [1]. In addition, the processing capabilities on the routers have increased due to the increasing demands of various networks applications

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 136–146, 2011.

such as encryption/decryption of data, packet filtering for firewalls, and virus detection.

In order to meet high-speed packets processing capabilities, conventional routers are mainly built with Application-Specific Integrated Circuits (ASIC) chips. However ASIC chips have a long development time and often cannot be adapted to various types of network protocols or standards. In addition, as the demands for the various network services have increased in the design of routers, needs for high-flexibility and short time-to-market requirements beyond ASIC chips have increased extremely. As a consequence, the NPs that combine high-speed packets processing capabilities of the ASIC chips and outstanding flexibilities of the GP have become a core element in design of routers [2].

Most commercial NPs today have to perform deep packet inspection for various network applications. Thus in order to provide various network services in a network device such as a router, an ability to process packets of L2 to L7 is essential to the NP of the network device [3, 4]. In addition, typical NPs today employ the multiple packet processing engines (PEs) to exploit packet-level parallelism and achieve high throughput in the packet processing. Major NP vendors have developed diversified parallel architectures to develop NPs that suitable for processing packets of L2 to L7. Commercial products for such NPs are the IBM PowerNP [5], EZchip NP-1 [6] and Intel IXP2800 [7]. IBM PowerNP consists of an embedded processors complex (EPC) with coprocessors [5]. EzChip developed NP-1 by using both paralleled and pipelined architectures. Intel employed a core processor with an array of multi-threaded packet processors.

However, most commercially available conventional ASIC-based NPs are not able to process L7 packets. Thus network devices that use conventional NPs are not able to provide a variety of network services. For this reason, to provide various network services, it is necessary to replace the conventional NP to the state-of-the-art NP or modify internal modules of the NP of the network device. However, these methods require high cost or a complicated process [8].

There have been several studies done to propose optimal architectures for the NP. Reference [9] describes an approach to explore the design space of the NP architectures on the system level. The highly paralleled architecture based on the paralleled processing-engine cluster (PPC) is introduced in [4]. This paper proposes a unified solution to deal with a load-balancing, intra-flow packet ordering, and memory contention problem in the NP. Reference [10] presents various NP architectural aspects and describes NP design characteristics. In addition, this paper also examines the internal organizational structure and functions of the NP building blocks including PEs and packet schedulers.

However, distinct from previous studies that focus on finding the optimized internal hardware architectures to develop the high-end NP, this paper presents a new interworking architecture for an NP that is able to process packets of L2 to L7 by combining a conventional ASIC-based NP with a GP. Specifically, our contributions in this paper are as follows:

- We describe a conventional interworking architecture of the NP employing a ASIC-based NP with a GP.

- We propose a new interworking architecture of the NP by applying a DPI module that classifies the packet processing flow by analyzing the packets [11] to the conventional interworking architecture.
- We show experimental results investigating the impact of our interworking architecture of the NP and compare it with the conventional interworking architecture.

Experimental results show that the proposed interworking architecture increases the utilization and throughput of the NP when compared with the conventional interworking architecture. The rest of this paper is organized as follows. In Section 2 and 3, we describe the conventional interworking architecture and the proposed interworking architecture. Section 4 presents and analyzes the experimental results. Finally, we summarize and conclude the paper in Section 5.

2 Conventional Interworking Architecture

Most commercially available conventional ASIC-based NPs are not able to process L7 packets. Thus, to design the NP that is able to process packets of L2 to L7 using conventional ASIC-based NPs, an interworking architecture with a GP is required. Usually, the GP as a co-processor is intended to be used as an accelerator with the conventional ASIC-based NP on line cards. It can process the L7 packets from the ASIC-based NP and return results back to the NP. Fig. 1 shows the conventional interworking architecture of the NP that consists of the conventional network processor (CNP) and the GP [12].

In general, the CNP is able to process only packets of L2 to layer 4 (L4) but the GP can process packets of L2 to L7. In this architecture, all the incoming packets are stored by the data memory controller (DMC) of the CNP, and packets of L2 to L7 are processed by the Processor Array (PA) in the GP. The brief description of each functions of the module in the conventional interworking architecture is as follows:

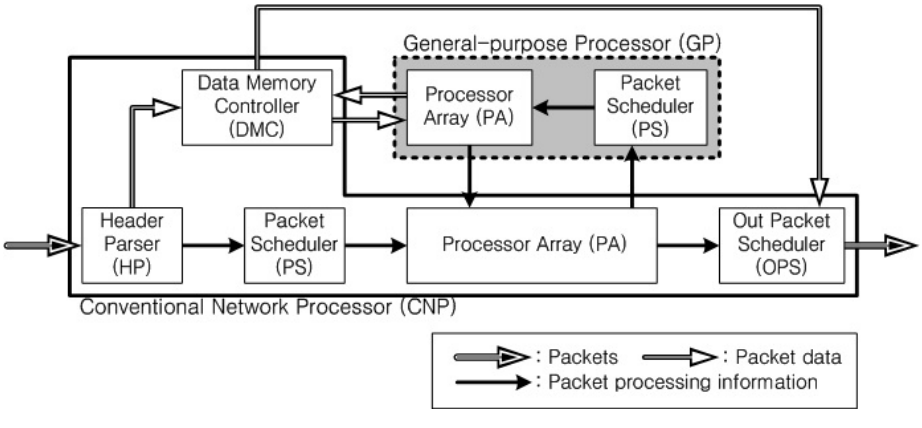


Fig. 1. Conventional interworking architecture for the L7 packet processing

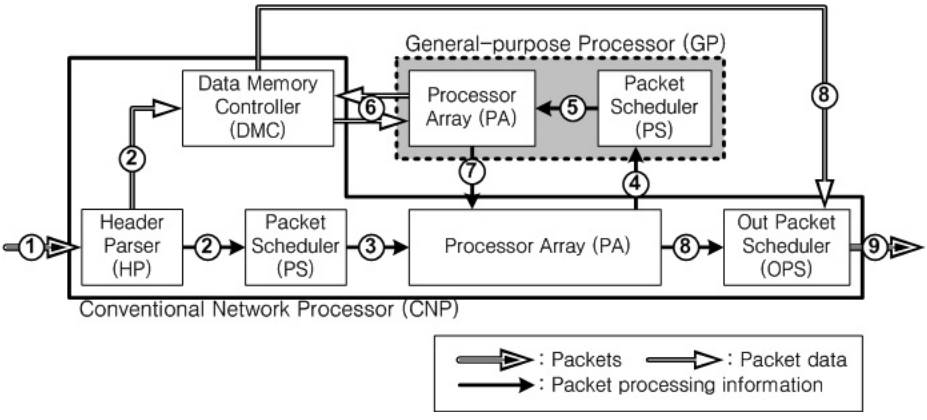


Fig. 2. Packet processing flow in the conventional interworking architecture

- Header Parser (HP): The HP analyzes all the incoming packets and then classifies the packets. The processing information of the packets is transferred to the packet scheduler (PS) and corresponding packet data are stored in a main memory by the data memory controller.
- Packet Scheduler: The PS receives the packet processing information from the HP and then the PS allocates it to the appropriate PE in the PA to process the packets.
- Processor Array: The PA in the CNP transfers all the packet processing information to the PA in the GP to process the packets of L2 to L7. This is because the PA in the CNP cannot determine whether the incoming packet should be processed by GP for the L7 packet processing or not.
- Out Packet Scheduler (OPS): The OPS plays a role in the management of processed packet output and regulation of the amount of the output traffic.

Fig. 2 shows the packet processing flow in the conventional interworking architecture. The packet processing flow of the conventional interworking architecture is as follows. First, for incoming packets, the HP analyzes the packet headers and then the processing information of the packets is transferred to the PA in the CNP and the packet data are stored by the DMC in the CNP. The PA in the CNP transfers the incoming packet processing information to the GP to process the packets of L2 to L7. The processed packets by the PA in the GP are transferred to the PA in the CNP again to output the packets to the OPS.

One of the most important bottlenecks for the conventional interworking architecture in general is the unbalanced distribution of packet processing. This is mainly because most of packets are processed by the PA in the GP. Although utilization of the PA in the CNP is decreased, this interworking architecture does not need additional time to determine whether the incoming packets should be processed by the GP for L7 packet processing or not. However, L2, L3, and L4 packets that do not need the L7 packet processing are also processed by the GP, thereby increasing the packet processing overhead of the PA.

3 Proposed Interworking Architecture

In this section, we propose a new interworking architecture of the NP that can increase load balance of the packet processing in the NP by employing the DPI. In the proposed interworking architecture, the DPI is required to classify packets to identify whether the incoming packet should be processed by GP for the L7 packet processing or not and it determines the packet processing flow, so that the GP performs only the L7 packet processing in contrast with the conventional interworking architecture. Also, for these reasons, the packets of L2 to L4 are processed only by the PA in the CNP. Thus, the packet processing overhead of the GP can be decreased compared with the conventional interworking architecture.

Fig. 3 shows the proposed interworking architecture. In the proposed interworking architecture, the following modules should be added or modified as described below:

- Deep Packet Inspector: For incoming packets, the DPI analyzes the header and payload of the packets and then classifies packets into OSI layers in which the packet processing should be performed.
- Data Memory Controller: In the proposed interworking architecture, two data memory controllers are required in contrast with the conventional interworking architecture. This is mainly because the packets are processed by either the GP or CNP, depending on the OSI layers in which the packets should be processed.

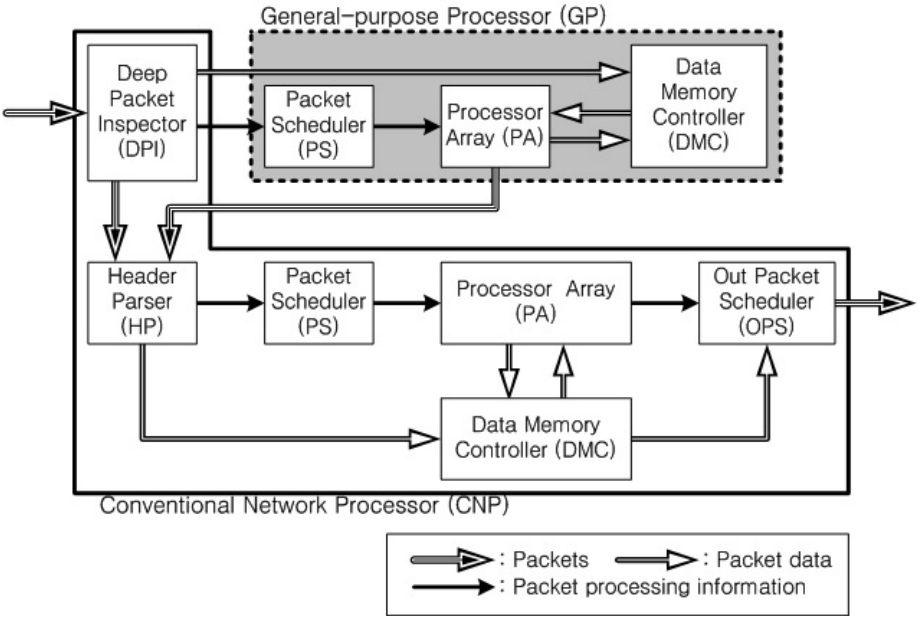


Fig. 3. Proposed interworking architecture for the L7 packet processing

The DPI not only has to inspect the headers of the packets, but also has to refer to the payloads of the packets, to determine whether it is necessary to perform L7 packet processing. Thus, the DPI needs a high memory usage and computational cost for accurate packet classification. In recent years, several hardware architectures and algorithms of the DPI have been studied [11]. In this paper, however, we do not deal with the hardware architecture and algorithm of the DPI.

In the proposed architecture, there are two packet processing flows as shown in Fig. 4 and Fig. 5. If the incoming packet requires the L7 packet processing, the packets are processed by the GP and then the processed packets are transferred to the CNP. However, if the packets do not require the L7 packet processing, the packets are directly transferred to the CNP for L2 to L4 packet processing. Thus the proposed interworking architecture can decrease the packet processing overhead of the PA in the GP by employing the DPI. However, the proposed interworking architecture requires additional processing time for the DPI and two data memory controllers are required for the separated packet processing flows in the NP.

Fig. 4 shows the packet processing flow for the packets of L2 to L4. The sequences of the packet processing are summarized as follows. First, the DPI classifies the packet processing flow by analyzing the packets. If the packets do not require the L7 packet processing, the packets are transferred to the HP in the CNP directly. The HP transfer the packet processing information to the PS and the packet data are stored by the DMC. The processed packets are transferred to the OPS to manage the processed packet output and regulate the amount of the output traffic.

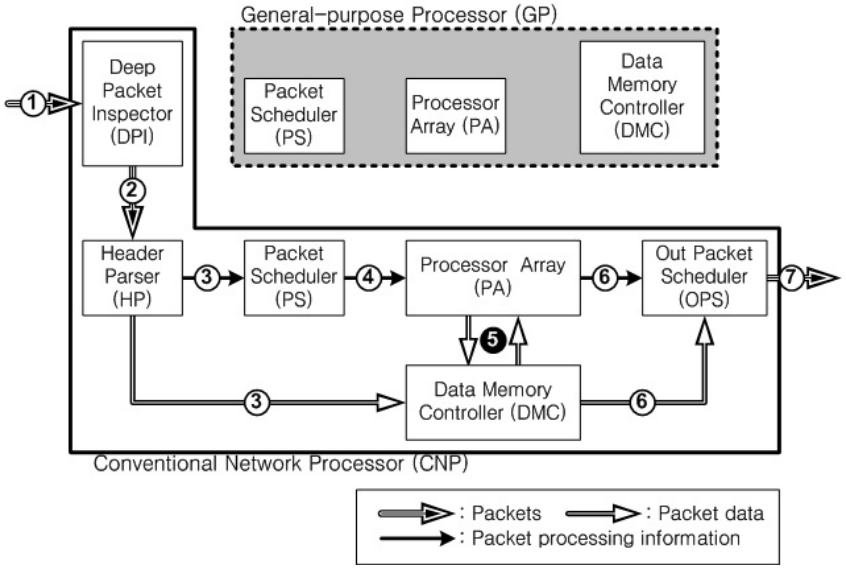


Fig. 4. Packet processing flow for the packets which do not require L7 packet processing in the proposed interworking architecture

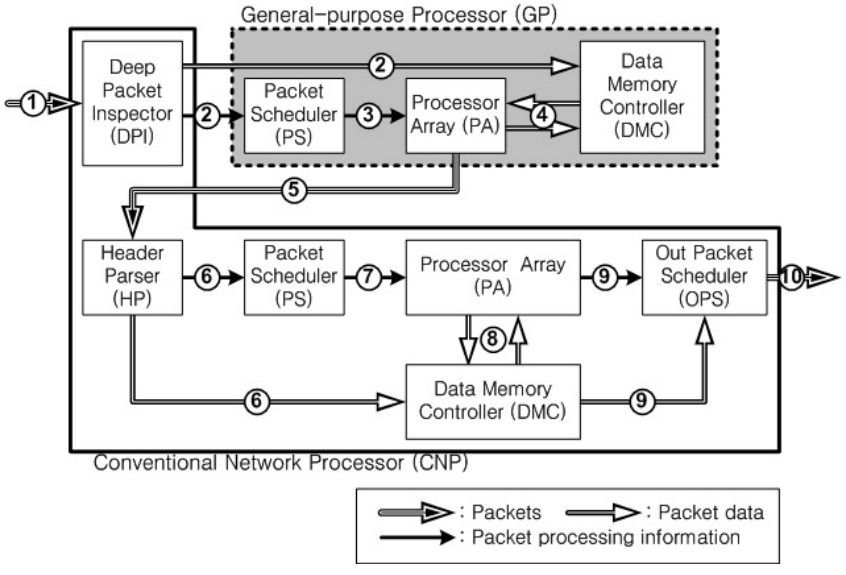


Fig. 5. Packet processing flow for the packets which require L7 packet processing in the proposed interworking architecture

Fig. 5 shows the packet processing flow for the packets which require the L7 packet processing. The sequence of the packet processing is summarized as follows. First, all the incoming packets are transferred to the DPI in the CNP and then DPI identifies the packets to determine the packet processing flow. In the case of the packet that is required to the L7 packet processing, the packet processing information are transferred to the PS directly and the packet data are stored by the DMC in the GP. The processed packets in the GP are transferred to the CNP for L2 to L4 packet processing.

In the proposed interworking architecture, the DPI is used to identify whether the incoming packet should be processed by GP for L7 packet processing or not and then transfers the packet to the appropriate processor as the OSI layers of the packets. Thus, the packet processing overhead of the GP can be decreased and the load balance of the packet processing in the NP can be increased as compared with the conventional interworking architecture.

4 Experimental Results and Analysis

4.1 Experimental Environment

To compare the conventional interworking architecture (arch. 1) with the proposed interworking architecture (arch. 2), these two architectures were modeled using C. For fair comparison, all the architectures consist of the sixteen PEs in the PA and the same PS that is used to schedule the incoming packets to each PA.

In the multiple-PE based NP, key design issue of the NP is to achieve high utilization of the PE, which is greatly affected by the PS and internal architecture of the NP. In this paper, we use per-flow packet scheduler in all the interworking architecture of the NPs to minimize the influences of the PS and to focus on the performance variations by the interworking architectures. In the PS, incoming packets of the same flow are assigned to the same per-flow queue in the PA, and then dispatched in order to the same PE in the PA [13]. As each PE processes the packets assigned to it and sends them to the OPS in the order that the packets come into the PE, the correct output sequence of packets of the same flow is easily maintained.

The other main design issue of the NP is to execute routing table lookup based on the destination IP addresses of packets at line rates. In this paper, we assume that miss ratio of the routing lookup is zero. Also, we assume that the DMC has sufficient wide bandwidth and storage space, thus removing the influence of the memory conflicts among the packet processing operations of the PEs.

In addition, the same traffic patterns were used in all the experiments and various traffic patterns were generated with a varying I_{L7}, the probability that L7 packet enters the NP consecutively (i.e. the frequency of burst-type L7 packet traffic patterns that affects the total amount of L7 packets in the traffic). And also, to analyze the packet processing time, 40,000 packets were generated with variable lengths, ranging from 50 to 1500 bytes, as most Ethernet LANs use maximum transfer unit (MTU) of 1500 bytes. While packet processing time depends on various factors, packet length is the key factor of packet processing time. Thus, in the experiments, it was assumed that the packet processing time is proportional to the packet length and clock speed with 400 MHz [14-16]. The DPI processing time is influenced by the traffic patterns and DPI matching algorithms [11]. Thus, we assume that the DPI processing time is constant and also lower than the L7 packet processing time in order to make the comparison of the experimental results easy.

4.2 Experimental Results and Analysis

Fig. 6 shows the average packet processing time of the proposed interworking architecture and the conventional interworking architecture as a function of I_{L7}. Fig. 6 (a) shows the results when the low-latency (LL) traffic patterns that generally consist of the packets from the routing protocols are used and Fig. 6 (b) shows the results when the high-bandwidth (HB) traffic patterns that generally consist of the stream of the packets from multimedia and video are used. The experimental results showed that when I_{L7} was increased, the packet processing time of the two interworking architectures are gradually increased, while the packet processing time of the proposed interworking architecture was faster than conventional interworking architecture. This is mainly because in the conventional interworking architecture, all the packets of L2 to L7 were processed by GP. Thus packet processing overhead of the PA is increased. However, in case of the proposed architecture, packets are processed by either the CNP or GP, depending on the OSI layers in which the packet processing should be required, so that the overhead of the packet processing of the PA in the GP is decreased and the packet processing time is decreased. In addition, as shown in Fig 6 (b), when the HB traffic patterns are used, the packet processing time is increased and the differences of the packet processing time between the

conventional interworking architecture and the proposed interworking architecture is decreased when compared with LL traffic patterns are used as shown in Fig 6 (a). This is mainly because the HB traffic patterns often require the L7 packet processing.

To investigate the load balances between the PA in the CNP and the PA in the GP, we compare the total number of the packets in the buffers of the PS in the CNP and GP. Fig. 7 (a) shows the maximum number of the packets in the buffer of the PS of the conventional interworking architecture when the LL traffic patterns are used. As shown in Fig. 7 (a), maximum number of packets in the PS in the CNP is constant. However, in case of the GP, when I_{L7} is increased, the maximum number of packets in the buffer also increased. This is because, in the conventional interworking architecture, all the incoming packets are transferred to the PA in the CNP directly and the PA in the GP have to process all the incoming packets regardless of whether the incoming packet should be processed by GP for the L7 packet processing or not. Fig. 7 (b) shows the maximum number of the packets in the buffer of the PS of the conventional interworking architecture when the HB traffic patterns are used. As shown in Fig. 7 (b), when the HB traffic patterns are inputted to the NP, the maximum number of the packets in the buffers of the PS is increased compared with the LL traffic patterns as shown in Fig. 7 (a).

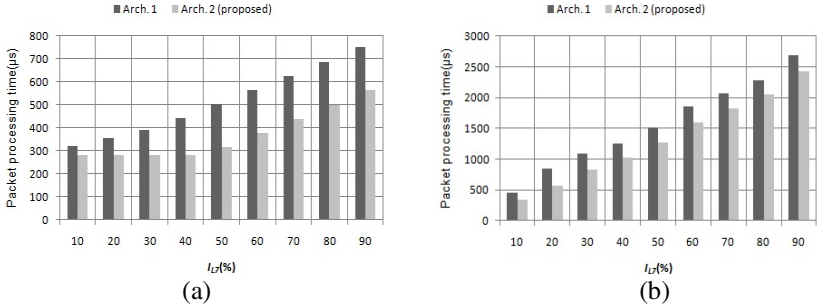


Fig. 6. Packet processing time in the (a) LL and (b) HB traffic patterns as a function of I_{L7}

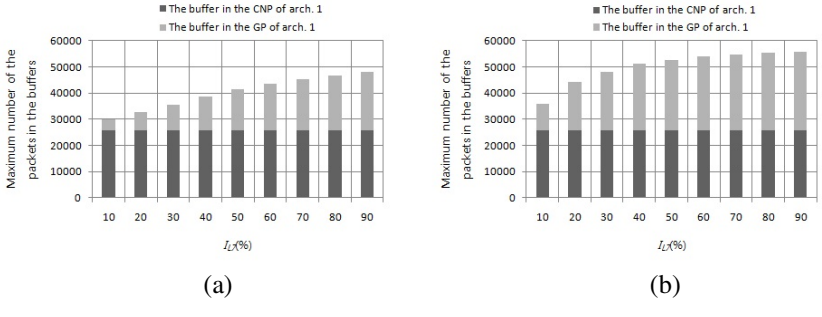
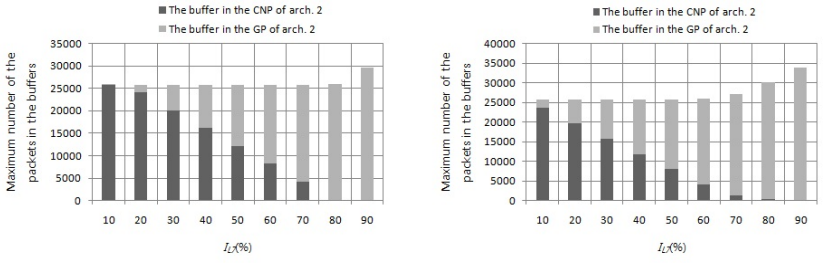


Fig. 7. The maximum number of the packets in the buffer of the PS of the conventional interworking architecture when the (a) LL and (b)HB traffic patterns are used



(a) (b)

Fig. 8. Maximum number of the packets in the buffer of the PS of the proposed interworking architecture when the (a) LL and (b) HB traffic patterns are used

Figs. 8 (a) and (b) show the experimental results when the LL and HB traffic patterns are used in the proposed interworking architecture, respectively. As shown in Fig. 8 (a), the total number of the packets in the buffers of the PS in the CNP is decreased while the total number of packets in the buffers of the PS in the GP is increased. This is mainly because in the proposed architecture, the sequence of the packet processing is divided by the DPI and only if when the L7 packet processing is required, the packets are transferred to the PA in the GP. As a result, the total number of the packets in the buffer of the PS in the GP is increased with I_{L7}. However, as shown in Fig. 8, varying I_{L7} had a minimal influence on the total amount of the packets in the buffers of the PS than the conventional interworking architecture as shown in Fig. 7 for the following reasons. First, in the proposed interworking architecture, only L7 packets are processed by the GP. Second, the packets of L2 to L4 are processed by the CNP. Thus, the packet processing overhead of the GP were decreased.

5 Conclusions

In this paper, we propose a new interworking architecture for an NP that is able to efficiently process packets of L2 to L7. To increase the throughput and utilization of the NP, proposed architecture determines whether the L7 packet processing of the current incoming packet is needed or not at the DPI and then assigns the packet to the appropriate processor to process the packet. Experimental results show that the proposed interworking architecture decreases the overhead of the PAs and increases the throughput of the NP when compared with the conventional interworking architecture. This performance improvement was also accomplished without high cost and complicated process by combining a conventional NP with a GP. Thus proposed interworking architecture is suitable for NPs of network devices for L7 packet processing.

Acknowledgments. This work was supported by the IT R&D program of MKE/KEIT. [KI002197, Development of Scalable Micro Flow Processing Technology].

References

1. Yusuf, S., Luk, W., Sloman, M., Dulay, N., Lupu, E.C., Brown, G.: Reconfigurable architecture for network flow analysis. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 16(1), 57–65 (2008)
2. Peyravian, M., Calvignac, J.: Fundamental architectural considerations for network processors. Computer Networks 41(5), 587–600 (2003)
3. Forouzan, B.A.: Data Communications and Networking. McGrawHill (2003)
4. Shi, L., Zhang, Y., Yu, J., Xu, B., Liu, B., Li, J.: On the extreme parallelism inside next-generation network processors. In: 26th IEEE International Conference on Computer Communications, Alaska, USA, pp. 1379–1387 (2007)
5. Allen, J., et al.: IBM PowerNP network processor: Hardware, software, and applications. IBM Journal of Research and Development 47(2.3), 177–193 (2003)
6. EZchip Technologies Ltd., NP-1, `http://www.ezchip.com/`
7. Adiletta, M., Rosenbluth, M., Bernstein, D., Wolrich, G., Wilkinson, H.: The next generation of Intel IXP network processors. Intel Technology Journal 6(3) (2002)
8. Coss, M., Sharp, R.: The Network Processor Decision. Bell Labs Technical Journal, 177–189 (2004)
9. Thiele, L., Chakraborty, S., Gries, M., Künzli, S.: Network Processor Design, Issues and Practices. In: Crowley, P., Franklin, M., Hadimioglu, H., Onufryk, P. (eds.) Design Space Exploration of Network Processor Architectures, vol. 1, pp. 55–89. Morgan Kaufmann Publishers (2002)
10. Haas, R., et al.: Creating advanced functions on network processors: experience and perspectives. IEEE Network 17(4) (2003)
11. Liu, T., Sun, Y., Guo, L.: Fast and Memory-Efficient Traffic Classification with Deep Packet Inspection in CMP Architecture. In: IEEE Fifth International Conference on Networking, Architecture and Storage, pp. 208–217 (2010)
12. Iyer, S., Kompella, R.R., Shelat, A.: ClassiPI^{TM}:An Architecture for Fast and Flexible Packet Classification. IEEE/ACM Trans. Networking 15, 33–41 (2001)
13. Qi, Y., Xu, B., He, F., Yang, B., Yu, J., Li, J.: Towards high-performance flow-level packet processing on multi-core network processors. In: Proc. 3rd ANCS, Orlando, USA, pp. 17–26 (2007)
14. Choi, J.Y., Han, J.Y., Cho, E.S., Kim, H.C., Kwon, T.K., Choi, Y.H.: Performance Comparison of Content-oriented Networking Alternatives: A Tree versus A Distributed Hash Table. In: IEEE 34th Conference on Local Computer Networks, pp. 253–256 (2009)
15. Zhang, J., Qian, Z., Shou, G., Hu, Y.: Online Automatic Traffic Classification Architecture in Access Network. In: ICEMI 2009, vol. 3, pp. 24–29 (2009)
16. Tian, X., Sun, Q., Huang, X., Ma, Y.: A Dynamic Online Traffic Classification Methodology based on Data Stream Mining. In: WRI World Congress on Computer Science and Information Engineering, vol. 1, pp. 298–302 (2009)

A Rectification Hardware Architecture for an Adaptive Multiple-Baseline Stereo Vision System

Hyeon-Sik Son[1], Kyeong-ryeol Bae[1], Seung-Ho Ok[1],
Yong-Hwan Lee[2], and Byungin Moon[3]

[1] School of Electrical Engineering and Computer Science, Kyungpook National University, Daegu, Korea
{soc_shs1984,puris1,wintiger}@ee.knu.ac.kr
[2] School of Electronic Engineering, Kumoh National Institute of Technology, Gumi, Korea
yhlee@kumoh.ac.kr
[3] School of Electronics Engineering, Kyungpook National University, Daegu, Korea
bihmoon@knu.ac.kr

Abstract. In this paper, we propose a new rectification hardware architecture that is suitable for an adaptive multiple-baseline stereo vision system. The adaptive multiple-baseline stereo vision system can obtain precise distance information compared with the fixed-baseline stereo vision system. To reduce the computation overhead of the stereo matching, the rectification is an essential step in the stereo vision system. However, the conventional rectification hardware architectures are designed for the fixed-baseline stereo vision system. Also, previously proposed the lookup table (LUT) based rectification hardware architectures are not suitable for the adaptive multiple-baseline stereo vision system. This is because these require large memory resources as the number of baseline increases. The experimental results show that the proposed hardware architecture does not need memory resources and also only few additional hardware resources are required as the number of the baseline increases as compared with the LUT-based rectification hardware.

Keywords: Rectification, Adaptive multiple-baseline, Stereo vision, Pipelined hardware architecture.

1 Introduction

Over recent years, stereo vision systems have been proved as the useful technique for obtaining three-dimensional information from two-dimensional images. Fig. 1 shows the baseline of the stereo vision system. The baseline (B) is the distance between the left and right camera. The Equation (1) presents the relationship between the disparity (d) and the distance (Z) where f is the focal length. As shown in (1), the baseline length is a magnificent factor in measuring d [3]. It means that the longer baseline is required to obtain the more precise disparity. Thus, the multiple-baseline stereo vision systems have been proposed [1, 2]. These systems vary the baselines of the stereo vision system as the surrounding circumstance for obtaining more precise distance information.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 147–155, 2011.

$$d = f \cdot B / Z \tag{1}$$

In obtaining distance information of the object in image, searching the corresponding pair of the pixel between two images needs high computational overhead. To reduce the computational overhead, the rectification is an essential step in stereo vision systems. Fig. 2 shows the overall flow of the stereo matching. The calibration step determines the extrinsic and intrinsic parameters of the left and right cameras. The rectification is the process of re-aligning epipolar lines to parallel with the x axis by re-mapping the image pixels from the input image to rectified image as shown in Fig. 5. Thus, the searching space can be reduced from two dimensions to one dimension [3, 4].

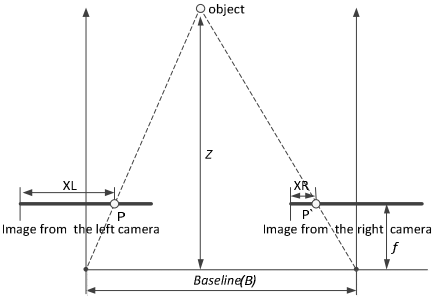


Fig. 1. The baseline of the stereo vision system

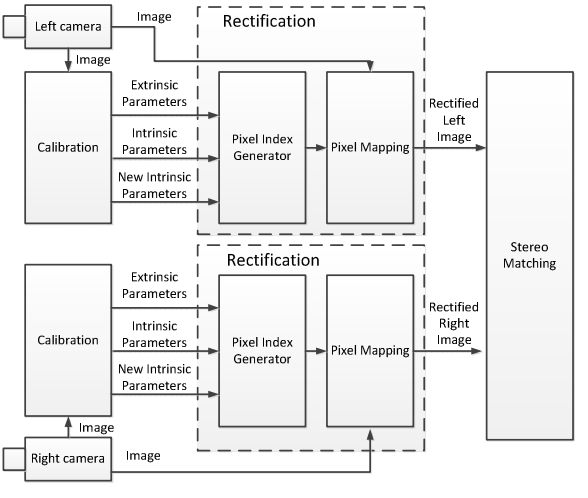


Fig. 2. The overall flow of the stereo matching processing

The rectification algorithm has been studied broadly in recent years. For the rectification procedure, two kinds of hardware architectures have been generally used. The first is the LUT-based rectification hardware architecture [5, 6] and the second is

the computational logic (CL) based rectification hardware architecture [7]. In general, for the real time rectification, the LUT-based hardware architectures are broadly used. However, as the resolution of the image increases, the LUT occupies very large memory space [5]. Also, in order to support the multiple-baseline stereo vision system, these architectures require large memory space for the LUT as the number of baseline increases. Thus, this method is not suitable for the adaptive multiple-baseline stereo vision system. The CL-based rectification architecture does not need large memory space unlike the LUT-based rectification architecture. However, the CL-based rectification architecture requires higher computational overhead than the LUT-based architecture.

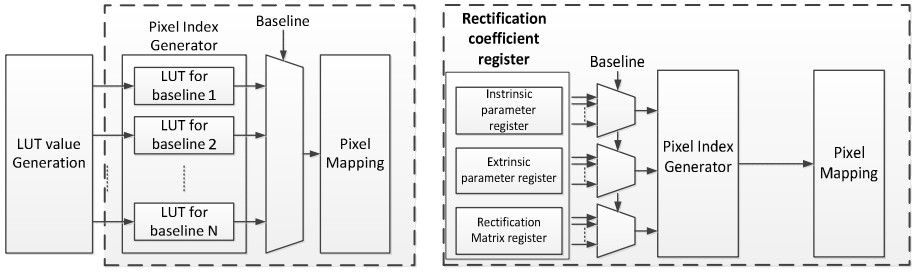
The previously proposed LUT-based rectification hardware architectures and the CL-based rectification hardware architecture are designed for the fix-baseline stereo vision system. Thus, the study of the rectification hardware architecture for the adaptive multiple-baseline stereo vision system has been strongly required as the importance of the multiple-baseline stereo vision system increases. For the multiple-baseline stereo vision system, the LUT-based rectification hardware architecture is not suitable because of the large usage of memory. Thus, we propose the fully pipelined, CL-based rectification hardware architecture that aims at the adaptive multiple-baseline stereo vision system.

The rest of the paper is organized as follows. In section 2, we describe the proposed rectification hardware architecture with the equations of the rectification. Section 3 presents the experimental environments then analyzes the results of the experiments. Finally, we summarize and conclude the paper in Section 4.

2 Proposed Rectification Hardware Architecture

Fig. 3-(a) shows the block diagram of the LUT-based rectification hardware. In the LUT-based rectification, the mapping indexes of every pixel are pre-computed and then stored in the LUT. Thus, this method can perform the rectification procedure easily. However, this architecture is not suitable for the multiple-baseline stereo vision system. This is mainly because the LUT is proportionally increased as the number of the baseline increases. The block diagram of the proposed rectification hardware architecture is shown in Fig. 3-(b). The proposed architecture computes the mapping indexes of every pixel using the Rectification Coefficient Register.

The LUT-based architecture needs large memory resources to store the mapping indexes of every pixel for the rectification than the Rectification Coefficient Register. For this reason, as the number of the baseline increases, the LUT needs very large memory resources. On the contrary, because the proposed rectification hardware architecture computes the mapping indexes of every pixel, it can, unlike the LUT-based rectification hardware architecture, perform rectification for the multiple-baseline stereo vision system without additional hardware resources. Thus, the proposed rectification hardware architecture is more suitable for the multiple-baseline stereo vision system than the LUT-based rectification hardware architecture. The proposed rectification hardware architecture is designed based on the equations in Table 1 [8].



(a) LUT-based hardware architecture (b) Proposed CL-based hardware architecture

Fig. 3. The block diagram of the rectification hardware architecture

Table 1. The equations of the rectification

Stage	Equations	Descriptions
0	$\begin{bmatrix} Gx \\ Gy \\ Gz \end{bmatrix} = R' \times KK\_new^{-1} \begin{bmatrix} Xn \\ Yn \\ 1 \end{bmatrix}$	R : Rotation matrix for rectification KK_new : Camera matrix for rectification (Xn, Yn) : Position index of rectified image
1	$Xd = \frac{Gx}{Gz},\ Yd = \frac{Gy}{Gz}$	
2	$\begin{bmatrix} X \\ Y \end{bmatrix} = KK \times \begin{bmatrix} Xd \\ Yd \end{bmatrix}$	KK : Old camera matrix
3	$alphaX = X(fraction)$ $alphaY = Y(fraction)$ $a1 = (1 - alphaX) \times (1 - alphaY)$ $a2 = alphaX \times (1 - alphaY)$ $a3 = (1 - alphaX) \times alphaY$ $a4 = alphaX \times alphaY$	
4	$R1 = a1 \times I(X_{\text{int}}, Y_{\text{int}})$ $R2 = a2 \times I(X_{\text{int}} + 1, Y_{\text{int}})$ $R3 = a3 \times I(X_{\text{int}}, Y_{\text{int}} + 1)$ $R4 = a4 \times I(X_{\text{int}} + 1, Y_{\text{int}} + 1)$	X_int : Integer of X Y_int : Integer of Y I(X, Y): Input image data
5	$rect_{image} = R1 + R2 + R3 + R4$	

The proposed pipelined rectification hardware architecture is as shown in Fig. 5. The proposed rectification hardware architecture consists of the Rectification Coefficient Register, the Pixel Index Generator and the Interpolator. The Rectification

Coefficient Register stores camera parameters and the rectification matrix coefficients as a function of the number of baseline. The Pixel Index Generator computes the mapping index of every pixel. The Interpolator computes the pixel value of the rectified image.

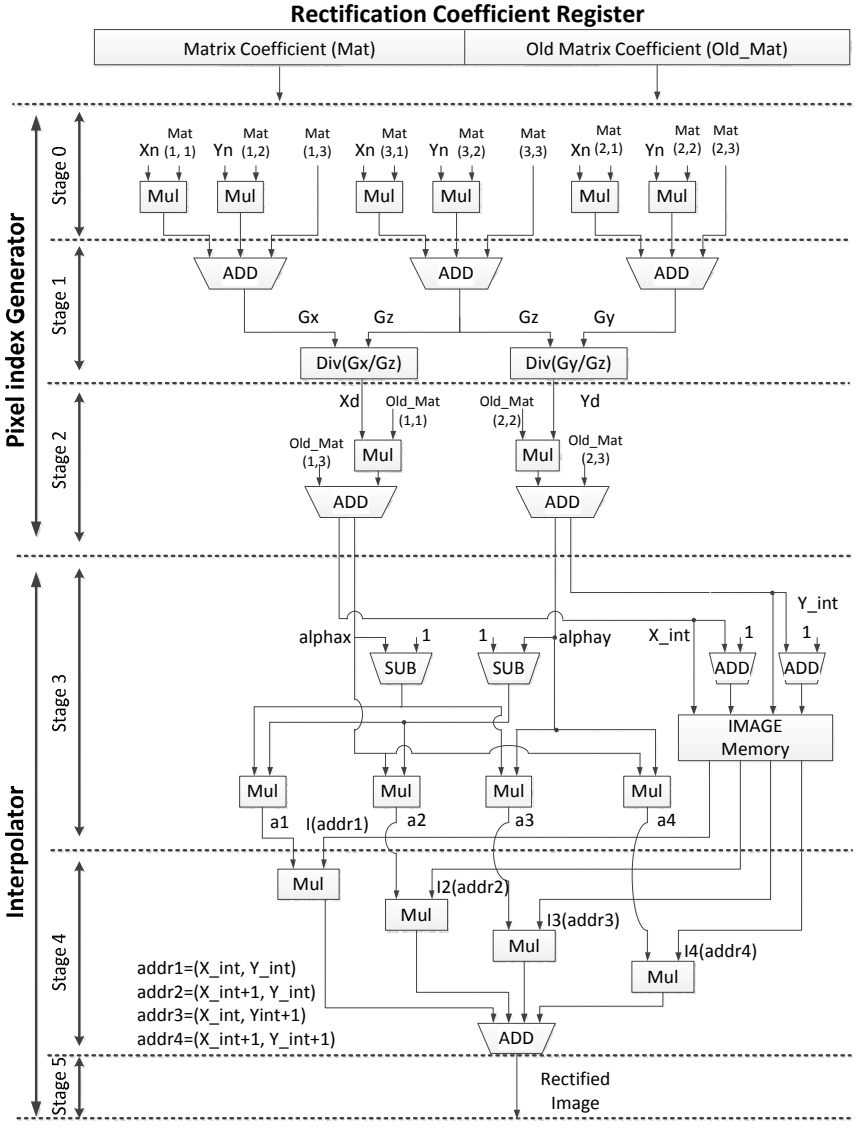
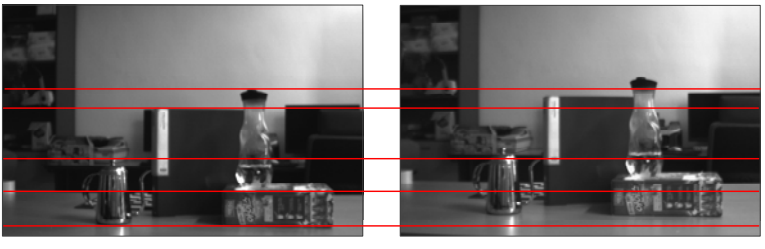


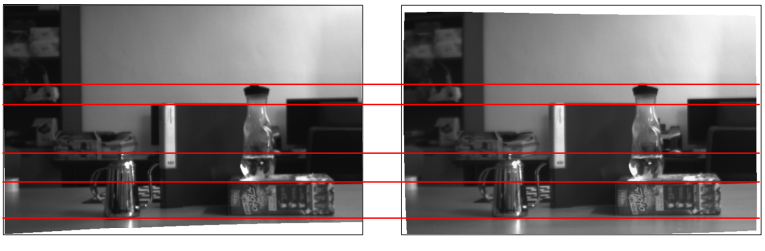
Fig. 4. The proposed pipelined rectification hardware architecture

3 Experimental Results

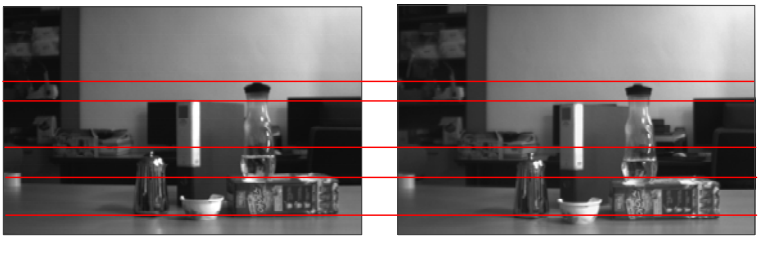
The proposed rectification hardware architecture is designed using Verilog HDL and implemented in Xilinx Virtex4 LX60 FPGA. Fig. 5 shows that the results of image rectification as a function of the baseline. As shown in Fig. 5, the epipolar lines are re-aligned. The proposed rectification hardware architecture is compared with the LUT-based rectification hardware architecture [6]. Table 2 shows FPGA implementation results. As shown in Table 2, the LUT-based hardware architecture not only needs large memory resources but also requires additional memory resources for the multiple-baseline stereo vision system. However, the proposed architecture does not use any memory resources. In addition, it requires only few additional hardware resources for the multiple-baseline stereo vision system compared with [6]. Also, as shown in Fig. 6, the memory usage of [6] is proportionally increased as the number of the baseline increases. In contrast, the proposed architecture does not use the memory resources and only few additional hardware resources are required as the number of the baseline increases.



(a) Original left and right images (baseline = 5 cm, image size = 752 x 480)

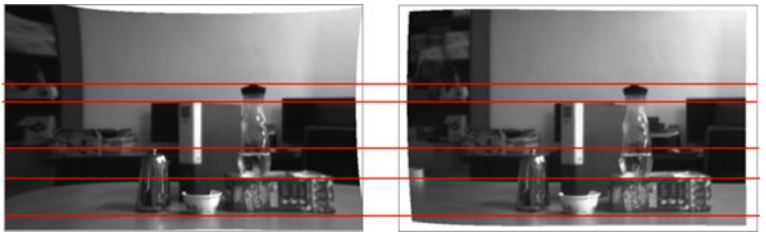


(b) Rectified left and right images (baseline = 5 cm, image size = 752 x 480)

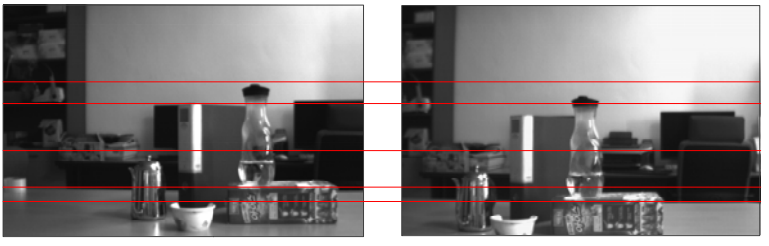


(c) Original left and right images (baseline = 10 cm, image size = 752 x 480)

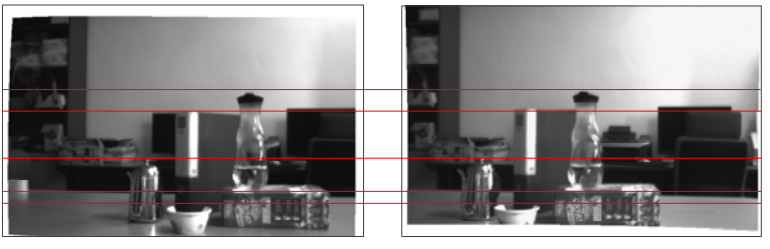
Fig. 5. Rectification results as a function of the baseline



(d) Rectified left and right images (baseline = 10 cm, image size = 752 x 480)

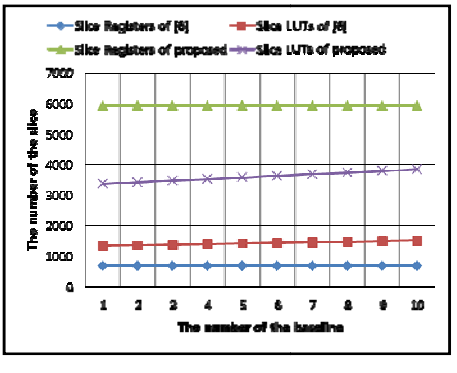


(e) Original left and right images (baseline = 15 cm, image size = 752 x 480)

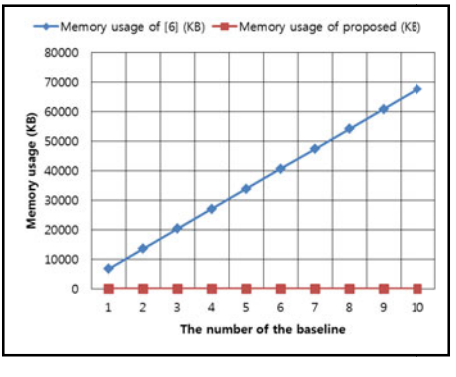


(f) Rectified left and right images (baseline = 15 cm, image size = 752 x 480)

Fig. 5. (*continued*)



(a) The number of the slice



(b) Memory usage

Fig. 6. Hardware resources as a function of the number of the baseline

Table 2. FPGA implementation results

	Fixed-baseline		Multiple-baseline (the number of the baseline = 2)	
[6]	Slice Registers	677	Slice Registers	677
	Slice LUTs	1,335	Slice LUTs	1,354
	Memory used (KB)	6,768	Memory used (KB)	13,536
Proposed	Slice Registers	5,932	Slice Registers	5,932
	Slice LUTs	3,365	Slice LUTs	3,418
	Memory used (KB)	0	Memory used (KB)	0

4 Conclusion

In this paper, to support adaptive multiple-baseline stereo vision systems, a new fully pipelined, rectification hardware architecture is proposed. The proposed rectification architecture consists of six pipeline stages. The proposed rectification hardware architecture does not require memory resources unlike the previous LUT-based rectification hardware architectures, and it requires only few additional hardware resources for the multiple-baseline stereo vision system. Thus, the proposed rectification hardware architecture is suitable for the adaptive multiple-baseline stereo vision system.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (2011-0013948).

References

1. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(4), 353–363 (1993)
2. Klarquist, W., Bovik, A.: Adaptive variable baseline stereo for vergence control. Proceedings of the 1997 IEEE International Conference on Robotics and Automation 3, 1952–1959 (1997)
3. Faugeras, O.: Three-Dimensional Computer Vision: A Geometric Viewpoint. MIT Press (1993)
4. Hartley, R., Andrew, Z.: Multiple view geometry in computer vision. Cambridge University Press (2003)
5. Park, D.H., Ko, H.S., Kim, J.G., Cho, J.D.: Real time rectification using differentially encoded lookup table. In: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication (2011)

6. Vancea, C., Nedevschi, S.: LUT-based Image Rectification Module Implemented in FPGA. In: IEEE International Conference on Intelligent Computer Communication and Processing, pp. 147–154 (2007)
7. Jung, H.S.: Real-time stereo image matching system. M.S. Thesis, Department of Electronics Graduate School, Kyungpook national University (2010)
8. Camera Calibration Toolbox for Matlab, http://www.vision.caltech.edu/bouguetj/calib_doc/htmls/parameters.html

Building Self-organizing Autonomic Agent Based on a Mobile Cell

Kiwon Yeom

Human Systems Integration Division, NASA Ames Research Center
San Jose State University Research Foundation
Moffett Field, CA 94035, USA
{kiwon.yeom}@nasa.gov

Abstract. This paper proposes and evaluates a biologically inspired autonomous system that makes networked agents or applications to be autonomous, scalable, adaptive. With the proposed system, a network application consisting of an associate of agents is designed by decentralized agents, which are analogous to mobile cells that have ability to migrate in biological systems. Each agent has a unique functionality for network systems, and implements biological behaviors such as migration, replication, reproduction, and death. The proposed system allows agents to autonomously sense its surrounding environments to evaluate whether they adapts well to the sensed environment conditions. Empirical measurement results show that the agent adapts well to the dynamic environments.

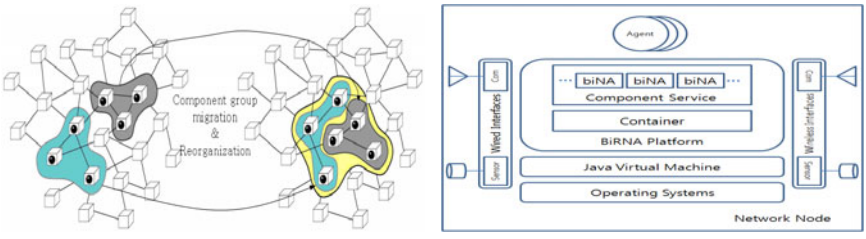
Keywords: self-organization, federation of agents, modular agents.

1 Introduction

In the near future, the ubiquitous computing calls for the deployment of a wide variety of smart devices throughout our working and living spaces [1]. Furthermore, many devices will be mobile. They are expected to dynamically discover other devices at a given location and continue to function even if they are disconnected. The overall goal is to provide users with universal and immediate access to information and to transparently support them in their tasks [2]. It is difficult to adapt to changes in user requests or changes in execution environments during runtime such as the spontaneous addition and removal of components and changing network topology. These factors will be more critical in both ubiquitous computing environments and large scale distributed systems such as grid computing environments [3], [4].

This paper addresses autonomous adaptability of agents (i.e. network application) in the biRNA (bio-inspired reconfigurable networking architecture) framework. Autonomous adaptability is an ability of network applications to allow adapting to dynamic changes in the network intelligently without any administrative intervention from and to human users. In this paper, the author describes a self-organizing system framework that allows for highly distributed and dynamic network applications using autonomous adaptive agents. This paper introduces several key features of the autonomous adaptive agents in our architecture and depicts functional requirements for our agent-based framework (see Fig. 1(b)).

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 156–165, 2011.



(a) (b)

Fig. 1. (a) Group migration in distributed systems. (b) The biRNA framework.

We describe the design and implementation of the framework, showing how the proposed framework satisfies the expected requirements of future networked applications. In order to explore the scalability and efficiency of the framework, we present some results from our empirical computations.

2 Related Works

Most existing agent platforms presuppose the existence of a host controller or of centralized agents. For instance, Hive deals with decentralized agents [5], but its current implementation is based on a central host and deployment of the remote message by Java RMI. In a few attempts, the architecture for real distributed systems have been presented [6], [7], [8], [9]. For example, the Anthill project established a bio-inspired middleware for peer-to-peer systems, which consists of a aggregation of interconnected dens [10]. Autonomous agents, called ants, can move around the network to satisfy user requests. The Co-Field project suggested the concept of general coordinator for the movements of agents and organization of a group of agents, which consist of mobile devices and mobile robots. However, it is only available within limited simulation environments. M. Murata and F. Dressler suggested biologically-inspired networking for communications between sensor nodes under mobile wireless sensor network environments. Their researches have been focused on data transmission mechanism which can consume low power source [11], [12].

In contrast, the proposed framework allows agents to organize a virtual network among agents using their functionalities and relationships. They also carry out distributed discoveries through interactions with each other, which results in social networking. Our proposed framework is that agents are independent objects that can accomplish their tasks and are flexible components that can be combined with and segregated from each other through the proposed framework. Table 1 describes the main differences between other agent based platforms and biRNA [13].

3 Association of Agents

The proposed framework presupposes that each application is composed of one or more components as shown in Fig. 1(a). Each component has a mobile single-cellular

Table 1. Comparison with other agent platforms

	AgentSpace	SOMA	Anthill	Pole	Co-Field	biRNA
Migration	Yes	Yes	Yes	Yes	Yes	Yes
Discovery	No	No	Partial	Yes	Yes	Yes
Policy	No	No	No	No	No	Yes
Asynchronous	Yes	Yes	Yes	Yes	Yes	Yes
ORB support	No	No	No	No	No	Yes
Messaging	No	No	Yes	Yes	No	Yes
Host Comm.	No	No	No	No	No	Yes

structure (modular agent or motile cell), since it is self-comprised and self-roving. A collection of components is akin to pseudo-deformation in biology because such a combination can modify the structure of a group of components.

The framework provides interaction measures for migration of components. It controls the migration of components to other computers to accomplish particular requirements or tasks. The framework should be used to develop an unicellular application as a set of agent-based components for components or application, which is able to migrate to other computing devices and reorganize components while the application is running [14]. The movement of one component may affect other components. For example, two components, which have different latencies in communication and velocities in movement and reorganization, are requested to simultaneously combine by a nearby computing device. One is for the control of the keyboard and another is to display content on the screen. Since each component has its own unique properties and control strategy, synchronizing a collection of components tends to be impossible over a distributed network environment. Therefore, components cannot efficiently coordinate with each other.

Therefore, the framework ought to enable each component to explicitly specify its own restraints for migration and reorganization. A component has to migrate to one of the most suitable computing devices under their own controls and constraints to satisfy their requirements and tasks. However, it is not always possible to decide exactly which destination is the most suitable. For this reason, the proposed framework allows a component to simultaneously distribute its clones to multiple computing devices. The framework then selects the most suitable clone while the others naturally vanish.

Our framework should be used to allow network protocols to manage the migration of mobile agents within a distributed network environment and as a middleware for self-organizing adaptive systems in distributed computing systems. We can easily imagine that there may be many different approaches for component management in a distributed system because most applications have their own specific management strategy. Therefore, the framework should be independent of any component deployment and organization approaches and of any biologically inspired processes [18].

The proposed framework can be employed as a general middleware architecture, which enables components to be exchanged between computers and allows applications to be implemented utilizing such components. That is, each component can have its own control and constraints for specifying spatial relationships between its location and other components' locations at neighboring computing devices. As a result, instead of any global policy, a collection of components should be managed through the association of the components' schemes.

4 Principal Attributes of biRNA

Agents in the biRNA framework are decentralized (*Decentralization*). There are no central entities to control and coordinate agents. Decentralization allows network applications to be scalable and simple by avoiding a single point of performance bottlenecks and failure, and any central coordination in deploying agents. Agents in the biRNA are autonomous (*Autonomy*). Agents sense their local surrounding environments and they autonomously behave without any human intervention or from/to other agents, platforms. Biological entities strive to seek and consume food for living (*Energy Management*). In biRNA, agents store and expend energy for living. Each agent gains energy in exchange for performing its service to other agents and expends energy to use network and computing resources. Agents in the biRNA framework are adaptive to dynamically changing environmental conditions (*Adaptability*). The adaptation is achieved from designing agent behavior policies to consider local environmental conditions. The abundance or scarcity of stored energy in agents affects their behaviors and triggers natural selection (*Natural selection*).

5 Agent Behavior Process

In this section, the algorithmic aspects of the behavior selection mechanisms are visualized with the UML (Unified Modeling Language) sequence diagram. The base class, migration schemes of components, and coordination strategy are described.

The biRNA framework provides the default implementation architecture which is automatically generated with basic class and implementation interfaces that can be observed in programming tools such as Visual C++, NetBeans, and IBM Architect [20], [21], [22]. Fig. 2(a) shows the basic class and behavior interfaces. In biRNA framework, an agent checks the energy level to determine the behavior. When the agent is generated at the first time, the agent has a certain amount of energy level. If the energy level of the agent becomes very low (i.e., below the death threshold), the agent is eliminated from the platform due to energy starvation (see Fig. 2(b)).

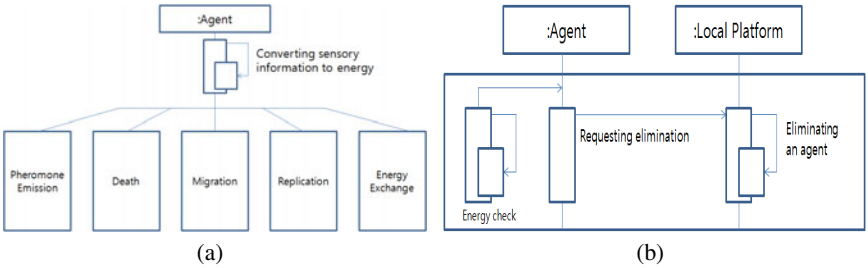


Fig. 2. (a) The basic class and behavior interfaces of the biRNA agent. (b) Agent's death behavior process.

An agent emits a pheromone (i.e., cytokine) if its energy level exceeds the threshold of pheromone emission. Agents continuously adjust their threshold value by CUSUM (cumulative sum control), which is a sequential analysis technique and typically used for monitoring change detection [23], of its energy level.

When a pheromone is emitted on a platform, all the agents on the platform can sense it. It may stimulate their replication or migration behavior. Each pheromone has its own concentration (i.e., 1000) and fade away by lapse of time. The pheromone completely vanish from the platform when the value becomes zero (see Fig. 3(a)).

$$S_m = \sum_{i=1}^{m} \bar{x}_i - \hat{\mu_0}, \qquad \text{cumulative sum} \tag{1}$$

$$S_{hi}(i) = max[0, S_{hi}(i-1) + x_i - \hat{\mu_0} - k], \qquad \text{for upper limit} \tag{2}$$

$$S_{lo}(i) = max(0, S_{lo}(i-1) + \hat{\mu_0} - k - x_i], \qquad \text{for lower limit} \tag{3}$$

where, m is the sample number, $\hat{\mu_0}$ is the estimate of the in-control mean.

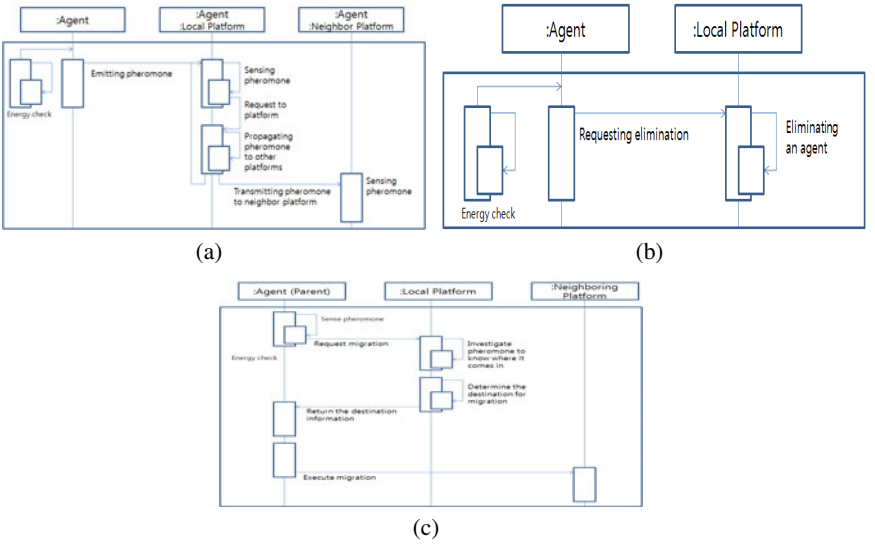


Fig. 3. (a) Pheromone emission behavior. (b) Replication behavior. (c) Migration procedure.

Replication occurs when the energy level of the agent exceeds its replication threshold, or pheromone stimulation threshold is bigger than pheromone concentration value. The agent keeps replicating itself until its energy level becomes less than its replication threshold value. When the agent replicates itself, pheromone stimulation threshold decreases at the same time. When an agent replicates itself, the agent gives its energy units to its child agent (replicated agent). The child agent gathers the sensory information from its parent agent that stimulated the parent agent to perform the replication behavior (see Fig. 3(b)).

Because the pheromone decays on a hop-by-hop basis and includes where it is flowed from, the agent can sense where the original platform exist approximately, and move toward the original platform by climbing pheromone gradients. This allows agents to move to neighbor platforms (see Fig. 3(c)). In order for agents to determine which neighboring platform they have to move to, platforms periodically relay pheromone to other platforms.

The component references are critical for tracking moving components and invoking their methods. This framework provides the APIs for invoking the methods of components on local or remote computers. It does not have to statically define any stub or skeleton interfaces through a precompiler approach because our target is a dynamic computing system (see the example invoking procedure). The component receives a reference at the current, which is given when the host is originally created. The reference is compared with the GID of the host at the new location. Then other components can invoke the methods of the newly attracted component. In addition, based on ORB's publish/subscribe concept, our framework allows users to implement the generic remote publish/subscribe mechanism that enables subscribers to express their requirements or tasks for an event. As a result, components may easily recognize whether or not they can accomplish their requirements at the current host. Table 2 is brief summary for Message protocol.

Table 2. The protocol for message exchange

Field	Value Set
Message Header	
deliveryMode	The same as that specified on the request
expiration	Derived from the request. Decide whether to degrade
priority	Copied from the request
correlationID	Copied from the request MessageID
destination	Copied from the ReplyTo property in the request
Message Properties	
requestURI	Copied from the requestURI property in the request message
bindingVersion	Copied from the bindingVersion property
contentType	Inferred from the Envelope and presence of attachments
Message Body	
body	Serialized message according to the media type

6 Experimental Results

In the empirical evaluation of the framework presented in this section, various measurements were obtained involving varying numbers of agents (1 to 1000 agents) and frameworks (1 to 16 platforms, two platforms per PC). Eight Windows PCs were used in the empirical evaluation, each running the Java 2 standard edition JVMs (version 1.4.2 from Sun Microsystems). These three PCs were divided into four groups with two multithreads in each group, according to their CPU speed and memory size (namely 4.2GHz with 512MB RAM and 4.3GHz with 1024MB RAM). These PCs were connected through 100 Mb/s Ethernet. In this measurement, 16 biRNA platforms are deployed on 8 PCs (i.e., two platforms per PC), and they are connected with each other based on a grid topology with equivalent distance (see Fig. 4).

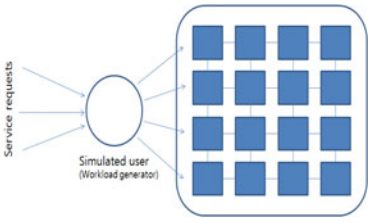


Fig. 4. Network topology for testbed

The bootstrap overhead measures the time for the framework to initialize each framework instance, and the bootstrap memory footprint measures the amount of memory space each framework instance consumes when it is initialized.

Table 3. Average bootstrap overhead and memory footprint

Framework component	Overhead (msec)	Footprint (KB)
Message transport	11.98	3.32
Agent manager	130.02	4.44
Class loader	6.72	1.58
Framework indicator	34.11	6.29
Resource sensing	32.28	22.01
Energy management	49.01	4.73
Migration	12.43	2.98
Cytokine/chemokine emission	15.79	3.53
Discovery	32.43	9.48

The measurement results demonstrate that the bootstrap overhead and memory footprint of each framework instance is very small. The footprint of the agent management service is relatively large because the service creates a thread pool, which consists of ten threads. The thread pool is used to improve the service response time for component requests. In lieu of a thread pool, if 2 components arrive at a host, one of the components must wait till the other component service is over. This situation may cause considerable latency in service and cause the system to stall. However, our framework is able to avoid such issues through the introduction of thread pooling. Incoming components are allocated to the most inactive thread. This ensures that all services will be served within a certain amount of time.

The total overhead of cytokine sensing is determined by the amount of time it takes for the cytokine sensing service to find the migrating agent's cytokine (by accessing a cytokine list maintained by the cytokine emission service), contact a representative of the platform that the cytokine specifies (i.e. the platform that the agent migrated to), and locate the migrated agent on the remote platform.

Fig. 5(a) shows how the overhead changes when an agent senses cytokines emitted on remote platforms, multiple hops away. It also illustrates that the overhead increases linearly as the hop count to remote platforms increases, which indicates that the cytokine sensing service is scalable. Fig. 5(b) shows the roundtrip time of a message between

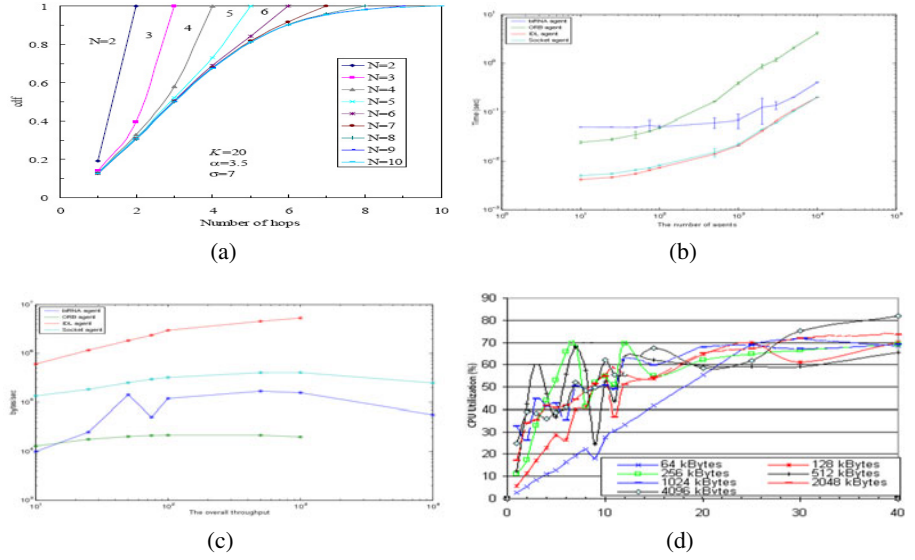


Fig. 5. (a) Overhead of environment sensing using cytokine emission services. (b) Roundtrip time of message between two components on different frameworks. (c) Comparing to other platform technologies. (d) CPU utilization.

two components running on different frameworks. In this measurement, a single component is deployed on a platform and a disparate set of components (from 10^1 to 10^4 receiver agents) is deployed on another platform. The sender randomly chooses one of the receivers and sends a component initialization message to the chosen receiver. Then the receiver sends back an acknowledge message to the sender. Fig.5(c) shows that the roundtrip time is comparable with well-known Java- based distributed object platforms (CORBA ORB and Java IDL), indicating that the networking interface for message transport is implemented efficiently. It also shows that the roundtrip time remains relatively constant as the number of receiver agents grows up to 10,000, which indicates scalability. Fig.5(d) shows CPU utilization of the web server and the proposed framework. The potential for high bandwidth has little value in practice if communication overheads leave no CPU power to process the data. CPU utilization is just as important as bandwidth since bandwidth will drop if application processing overwhelms the CPU.

Table 4. Migration overhead

Size	Data size (byte)			Execution time (seconds)			Migration time (seconds)		
level	1	4	8	1	4	8	1	4	8
1024	49316	49732	50321	0.534	0.548	0.567	0.018	0.018	0.018
2048	98168	99024	99675	1.725	1.757	1.778	0.041	0.041	0.047
4096	193872	197788	198376	4.799	4.821	4.997	0.096	0.103	0.104
8192	391480	394916	399684	11.021	11.534	11.891	0.179	0.192	0.198
16384	783696	787292	788120	24.185	24.478	24.729	0.327	0.386	0.394

Table 4 shows the overhead for an agent to migrate from one platform to another using the migration service. The migration overhead includes the transmission time over the network and the processing time at both the origin and destination platforms. As the size of mobile code grows, the overhead increases linearly, instead of exponentially, indicating that the migration service is scalable.

7 Conclusions and Future Work

This paper presented a middleware system comprised of dynamic collections of components on a distributed system. With biologically inspired principles and mechanisms, network applications were created based on the proposed architecture satisfying the key requirements of future network applications such as autonomy, scalability, and adaptability. We designed and implemented a prototype middleware system and demonstrated its effectiveness in several applications. Empirical evaluation shows that the platform is efficient, scalable, and reusable. There are still further issues that need to be resolved. Although the current implementation focuses on the deployment of components, we plan to extend the framework so that it can be used to modify the behavior of each component during execution. The final goal of this architecture is to provide a general test bed for various bio-inspired approaches for adaptive distributed systems. Also, since the current performance is not satisfactory, further measurements and optimizations are required.

Acknowledgments. This work has been partially supported from NASA Ames Research Center, Moffett Field, CA, USA. I would like to acknowledge and extend my heartfelt gratitude to my advisors, Dr. Stephen R. Ellis and Dr. Bernard D. Adelstein whose encouragement, guidance and support in Advanced Control and Display Group of Human Systems Integration Division. I offer my regards to Prof. Kevin P. Jordan in San Jose State University Research Foundation who supports me in any respect.

References

1. Esler, M., Hightower, J., Anderson, T., Borriello, G.: The future of computing. Scientfic American 281(2), 52–55 (1999)
2. Grimm, R., Davis, J., Lemar, E., MacBeth, A., et al.: Programming for pervasive computing environments. In: Proceedings of ACM Symposium on Operating Systems Principles
3. Roman, M., Ho, H., Campbell, R.H.: Application mobility in active spaces
4. Brumitt, B., Meyers, B., Krumm, J., Kern, A., Shafer, S.: EasyLiving: Technologies for Intelligent Environments. In: Thomas, P., Gellersen, H.-W. (eds.) HUC 2000. LNCS, vol. 1927, pp. 12–29. Springer, Heidelberg (2000)
5. Minar, N., Gray, M., Roup, O., Krikorian, R., Maes, P.: Hive - distributed agents for networking things. In: International Symposium on Agent Systems and Applications / International Symposium on Mobile Agents, pp. 569–571 (1999)
6. Satoh, I.: Building reusable mobile agents for network management. IEEE Transactions on Systems, Man and Cybernetics 33(3)
7. Yeom, K., Park, J.: Bio-inspired self-organization architecture for distributed components. In: Proceedings of International Conference on Ultra Modern Communications (2009)

8. Suzuki, J., Nakano, T., Fujii, K., Ikeda, N., Suda, T.: Reconfiguration of network applications and middleware systems in the bio-networking architecture. In: Proceedings of IEEE LARTES (2002)
9. Wang, M., Suda, T.: The bio-networking architecture: A biologically inspired approach to the design of scalable, adaptive, and survivable/available network applications. In: Proceedings of the 1st IEEE SAINT Conference (2001)
10. Overeinder, B., Osthumus, E., Brazier, F.: Integrating peer-to-peer networking and computing in agentspace framework. In: IEEE International Conference on Peer-to-Peer Computing (2002)
11. Murata, M.: Biologically inspired communication network control. In: International Workshop on Self-* Properties in Complex Information Systems (2004)
12. Dressler, F., Kruger, B., Fuchs, G., German, R.: Self-organization in sensor networks using bio-inspired mechanisms. In: ARCS 2005: Workshop Self-Organization and Emergence, pp. 139–144 (2005)
13. Yeom, K., Park, J.: Bio-Inspired Adaptive Framework Enabling Self-Organization and Management for Distributed Modular Agents. Journal of Internet Technology 11(5)
14. Montresor, A., Babaoglu, O.: Biology-inspired approaches to peer-to-peer computing in bison. In: International Conference on Intelligent System Design and Applications (2003)
15. MetaMicrobe, D. discoideum, http://www.metamicrobe.com/dicty
16. Davis, D.M.: Nature reviews immunology 7, http://www.nature.com/nri/journal/v7/n3/full/nri2020.html
17. Cooper, H.: The Cell: A Molecular Approach. ASM Press (2003)
18. Szyperski, C. (ed.): Component Software. Addison-Wesley (1998)
19. Yeom, K.: Bio-inspired self-organization for supporting dynamic reconfiguration of modular agents. Mathematical and Computer Modelling
20. Microsoft, Microsoft visual c++, http://www.microsoft.com
21. NetBeans, Netbeans 7.0, http://www.netbeans.org
22. IBM, http://www-01.ibm.com/software/awdtools/systemarchitect/
23. Basseville, M., Nikiforov, I.: Detection of abrupt changes: theory and application, Citeseer, vol. 10 (1993)

Modeling u-Healthcare Frameworks Using Mobile Devices

Haeng-Kon Kim

Dept. of Computer Information & Communication Enginerring,
Catholic University of Deagu, Korea
hangkon@cu.ac.kr

Abstract. It is widely anticipated that there will be a paradigm shift in the coming year from desktop-centric computing to mobile computing. This will have a number of significant ramifications. One issue of critical importance is the design, development and deployment of software solutions for mobile users. In this paper, we suggested a mobile collaboration framework based on distributed object group framework (DOGF). This paper focuses on the use of this framework to support mobile collaboration. Therefore, we improved the existing work and apply it to the construction of u-healthcare agent services. For supporting mobile collaboration, we divided into two agent types such as the stationary and the moving-typed agents according to the function of mobile devices. The data collected by sensors attached on arbitrary spaces can be shared by 2-typed agents or a home server, and exchanged with each other using the Push and Pull methods. For managing this information, the DOGF provides functions of object group management, storing information and security services to our mobile collaboration framework via defined application interfaces. We also showed via GUI the executablility of healthcare application supporting for medical work in hospitals on our mobile collaboration framework. Information Management using mobile phone in Wireless Sensor Network is a very common factor. This system will be as simple as possible providing all the facilities those are available in remote PC based applications. A wireless mesh network furnishes the WSN. Mobile phones are going to be the nodes or access points of this network. Some predefined activities like error messages or warning messages after sensing the system environment along with all types of responses to user queries can be achieved in this system. A simple user friendly GUI is implemented for user and there is a server to authenticate the logged in user, listen to the requests and replies them accordingly. The user can(i) get the reports regarding all/individual tags and or routes, emergency messages, sensors and also the location information and /or (ii) send actuation command to the server using which immediate decision about the system can be taken from server side. Thus any remote user can be assisted by this system and can communicate with WSN anywhere, anytime.

Keywords: u-health Agent node, Mobile collaboration frameworks, Agent UML modeling, Mobile development.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 166–178, 2011.

1 Introduction

It is widely anticipated that there will be a paradigm shift in the coming years from desktop-centric computing to mobile computing. This will have a number of significant ramifications. One issue of critical importance is the design, development and deployment of software solutions for mobile users. These users utilise such applications in a highly dynamic environment and such systems need to perceive and demonstrate responsiveness to environmental changes. While pre-existing methods, possibly enhanced and extended in various ways, will offer partial support, it is clear that new technologies of a design, implementation and methodological nature are demanded. The approach we advocate is the use of intelligent agents encapsulated in an agent oriented design methodology. The implementation apparatus is provided via Agent Factory [1], an environment for the rapid prototyping of agent-based systems, the design notation is that of agent UML (AUML) [2] and the methodology utilised is that of the agent factory development methodology. Up to now, the computational complexity of delivering an agent-based solution within the computational constraints of mobile devices has rendered such an approach futile at worst or at best impudent. However, recent developments in mobile computing hardware have made the deployment of agent-based solutions feasible.

Ubiquitous computing integrates computation into the environment, rather than having computers which are distinct objects. This means that individual physical environment is a share and exchange of information that are collected data from the devices, sensors and desktop computer over wireless network [1,2]. Recently, research in ubiquitous computing is towards the development of an application environment able to deal with the mobility and interactions of both users and devices [3,4,5]. In these researches, collaboration is a very important application of information technology, especially for a u-healthcare environment. With the fast development of mobile computing, wireless mobile networks and mobile devices are becoming widely used in both commercial and academic organizations. Collaboration turns out to be much more useful in a mobile network.

Therefore, this paper presents the mobile collaboration framework based on the distributed object group framework [6]. The DOGF provides functions of object group management, storing information and security services to our mobile collaboration framework via defined application interfaces. With the aim of analysing a complete understanding of the medical work practices, how they manage the medical information, and how they interact with each other, we considered a workplace study [7]. We decided to scenarios as a way of integrating our researching for u-healthcare framework. These scenarios that require for mobile collaboration include exchange information, sensing data process and group based distributed object collaboration by agent on mobile devices. There are also a lot of applications that can integrate collaboration into healthcare application services.

When considering our environment we found that there were only two mobile devices that would satisfy our demands: Personal Digital Assistants (mobile devices) and mobile phones. The mobile devices had a clear advantage over mobile phones when considering sensor data collection and mobility. It was therefore natural to focus

on the available mobile devices in a hospital space and technologies available such as mobile devices. According to the function, we divide into two classes; stationary and moving-typed agent. The functions of a stationary-typed agent on mobile devices or PC are it collecting related patient health data and environment information about the ward from the healthcare sensors or devices. While the moving-typed agent on mobile devices can support the nurse's business that provides the interaction with stationary-type agents and home server. For our implementation environment, we used TMO scheme and TMOSM for interactions between distributed components. Also, we used the Bluetooth technology for interactions between mobile devices. Finally, we showed via GUI the executablility of a healthcare application on our mobile collaboration framework.

The rest of the paper is organized as follows. Section 2 presents related research work. The next section 3 describes the architecture of mobile collaboration framework. Also, we explained the mobile collaboration type which is the type according to the function of mobile devices. Section 4 describes the healthcare application that implements the architecture and demonstrates its features and abilities based on mobile collaboration framework. The last section describes the conclusion and future works.

2 Related Works

2.1 The Mobile User

Developing applications for mobile users is difficult and problematic. A simple comparison between the resources available to the average desktop user and their mobile counterparts starkly illustrates this. Processing power, screen size, battery or power source life span and network capability are all significantly poorer on PDAs when contrasted with their desktop cousins. Indeed, it is probable that this resource gap will always exist. Likewise, methods of interacting with portable devices differ from those traditionally used. Unfortunately, the issue of solution deployment is further complicated when it is considered that portable devices can differ significantly between themselves.

Despite such limitations, it is the dynamics of both the mobile user's behavior and the environment within which the system is being used which form the more interesting challenges and, as such, prove of particular interest to us. A mobile user_s context is by definition dynamic as their needs and their environment are in a continual state of flux. This implies that their needs and expectations may change as they roam about the physical environment. In addition, the quality of what we might term the prevailing electronic infrastructure, for example networking ability, may vary considerably at any given time. How best to engineer applications that can operate under such varying circumstances while simultaneously meeting users' needs and expectations is still an open question.

2.2 Agents: A Possible Solution

So when should intelligent agents be considered as a basis for a possible software solution? There is no hard and fast answer to this question. However, agent practitioners consider those systems that can be logically subdivided into a series of dynamic interacting components as being potentially suitable for agents. Furthermore, applications that must operate within environments that are complex and dynamic are perceived as being particularly suitable, as traditional approaches do not always include mechanisms for effectively modeling the myriad of scenarios that such environments present. We need look no further than our immediate surroundings for an example of an environment that is inherently complex. A mobile user with varying information requirements operating within such an environment introduces a dynamic component that increases the difficulty in delivering services and applications further. Therefore we can consider that such a scenario offers a fertile area for the deployment of agent-based solutions. Such classes of problem are inherently suited to an embedded agent approach. The agents are embedded within the mobile device and constantly perceive both the environment and the user's behaviour. Subsequently, intentional agents may affect their environment through actuator activation. Agents typically would collaborate, each offering a restricted set of capabilities or responsibilities like user profiling or interface event listening or map/content presentation. Such collaborative behaviour would offer the requisite adaptivity to reflect the dynamics of such classes of system. We do not suggest however that agents are a panacea offering an all-encompassing solution to the difficulty of developing mobile applications. Rather, it is the nature of the domain in which the application is deployed, as well as the user_s expected behaviour and requirements, that will ultimately dictate as to whether an agent-based solution is desirable or otherwise.

2.3 Our Research

Our research enables the design of more flexible and suitable for ubiquitous healthcare environments. Currently, various researchers have been suggested in order to mobile collaboration. This section describes some of the related researches and projects for this paper. Mobile Collaboration services enable users utilizing mobile device to share information and communicate with each other. Related research has been conducted on various models that support collaboration. Pervasive Collaborative Computing Environment (PCCE) project[8] that offers an environment for supporting scientific collaborations. This environment houses various tools needed in collaborations such as a synchronous/asynchronous messaging, video conferencing, and file sharing/transfer. iClouds project is an architecture for supporting spontaneous exchange of information between users that come within each other's digital sphere[9]. The goal of iClouds is to provide a platform for spontaneous collaboration, taking advantage of the observation that people gathered at the same location often have common interests or goals. JXTA[10] based on J2SE project is java framework for P2Pnetworks that targeted on desktop computers. For solve this platform problem, Sun Microsystems suggests the JXME and JXTA for J2ME for

support on small mobile devices. PROEM[11] is an open computing platform that provides a complete solution for developing and deploying P2P applications for MANETs. The platform is based on experiences from developing a series of mobile applications. It is implemented as a platform independent framework using J2SE. Such a project constitutes a high level of collaboration between individuals via the use of computers and wireless technologies. But, several approaches have been proposed to provide aspects related to low-level collaboration. We have developed a distributed object group framework that is component-based framework. Therefore, we improved the existing work and apply it to the construction of u-healthcare agent services for mobile collaboration. For supporting mobile collaboration, we focus on high level collaboration supported by DOGF.

3 Modeling of MC Framework

This section presents an overview of our MC (Mobile Collaboration) framework architecture. This framework provides flexible support for defining and configuring the group through the service components of DOGF and behavior of the principal elements of a mobile collaboration environment. Also, we describe the type of mobile collaboration.

3.1 MC Architecture

The characteristics of agents, though the subject of much debate in the past, have coalesced around a number of core features. These include autonomy, reactivity, proactivity and co-ordination. If we consider a typical mobile user for example, an agent might operate on their PDA, monitor the user_s behavior in an autonomous fashion, react to any perceived changes in the user_s status, and proactively anticipate what the user_s future behaviour will be. Finally, in co-ordination with other agents, it might arrange access to those services that it thinks the user may need as their context changes. While such characteristics satisfy the criteria of agenthood for many people, some researchers consider them as being only the minimum required. Indeed, such people would augment these characteristics with other stronger features such as sophisticated reasoning capabilities. For the purposes of this article we subscribe to this latter view and utilise agents that have been realised through the BDI paradigm [3]. BDI agents are a classification of agents constructed around the concepts of Beliefs, Desires and Intentions. To summarise briefly: an agent is assigned a number of tasks to perform. Such tasks essentially form the agent_s raison d_eˆtre and, in the BDI scheme, are represented as desires. However, before a desire can be fulfilled, a number of criteria must be satisfied. Such criteria, referred to as beliefs, are continuously formulated and updated as the agent monitors its environment. In general, it is unlikely that an agent will be able to fulfill all its desires at any given time. However, it is realistic to assume that it could fulfill some of them and such desires are formulated as intentions which the agent then proceeds to undertake. BDI agents are a mature realisation of the strong intentional agent stance [4] and offer an

intuitive and computationally tractable model of intelligent agents. Naturally, this makes using the paradigm attractive to those designing and developing agent-based solutions.

Our framework used to component of supporting object group management for domain grouping in distributed object group framework, we consider the interaction of mobile devices and sensors. Also, for information collection and sharing in this environment, we adopted the TMO scheme and TMOSM [12] into the development environment of the healthcare application.

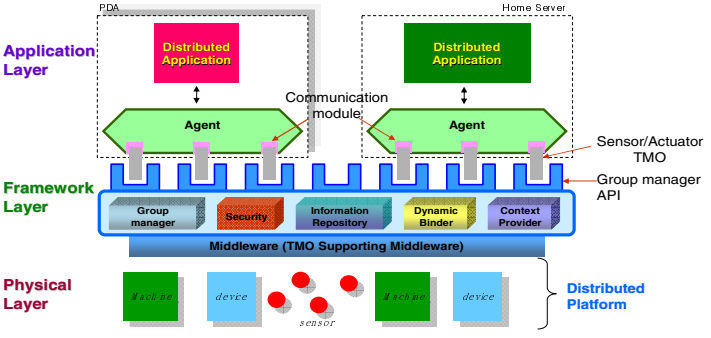


Fig. 1. Architecture of MC framework

Figure 1 shows its architecture. It supports a logical single view system environment by grouping them. The group manager API supports the execution of application of appropriate mobile collaboration service on upper layer by using the input information obtained from individual or grouped physical devices on the lower layer as distributed platform. That is, according to the mobile collaboration service or status of the logical domain for collaboration, our framework could configure new groups dynamically by integrating physical devices/sensors or machines on the distributed platform and distributed application and agent on the upper layer. For the middleware of interaction between distributed applications adapted the TMOSMS.

3.2 The Interactions of Component

In the whole mobile collaboration framework, we defined the interaction of components which interacts with the distributed application include the agent, sensors and components of framework. The group manager object provides the interaction of agent in distributed application on mobile devices or Server and sensor by APIs and service object reference which support collecting real time information from the sensor nodes. Also it supports the security service which is a security object that checks access right for client. When service object is replicated, dynamic binder object provides the reference of service object by binding algorithm. The stationary-typed agent on mobile devices obtains real time information of sensor node through service object reference. And, the interaction of objects in distributed application returned the result of service by framework components. The suggested framework

located at home server supports the security service and manages the devices and related information. Also, home server supports management of collecting information by stationary-typed agent and control of appliance by the context provider object.

3.3 The Agent Type of Mobile Collaboration

This paper suggested mobile collaboration framework, interacts with devices and sensors according to the type of devices. According to the function, we divide it into two classes; stationary and moving-typed agent.

The functions of a stationary-typed agent are collecting related patient health data and environment information about the ward from the healthcare sensors or devices. The moving-typed agent can support the nurse business. Also, its function is similar to the stationary-typed agent. The difference between stationary-typed agent and moving type agent is that the latter only collects the data whereas the moving type agent usually provides the interaction with other agents and home server.

We defined the mobile collaboration type about the interaction of devices. The information collected by sensors can be shared and exchanged by agents or home server in accordance to Push and Pull methods. The interaction between sensors and stationary-typed agent is using push methods that is collecting information using sensor send to moving-typed agent. And, the interaction between stationary-typed agent and moving-typed agent is using push/pull methods. In this case, the push methods are available to reconfiguration for sensor network and collecting the data from the sensor nodes. Then the pull methods are available to the exchange information between agents. Also, the mobile device and home server used push/pull methods of agents and communication ways which can use Bluetooth or wireless LAN. Firstly, Bluetooth way is to exchange data with the home server which is searched Bluetooth dongle. When mobile device has no Bluetooth dongle for the communication way, then mobile device can search the access points by IEEE 802.11b, g. After searching for several devices by AP, it is communicating with searched devices. The above interaction way is according to the system environment. The home server provides information management and service based on context recognition. Figure 2 shows the architectures for the agent type id mobile collaboration in this paper.

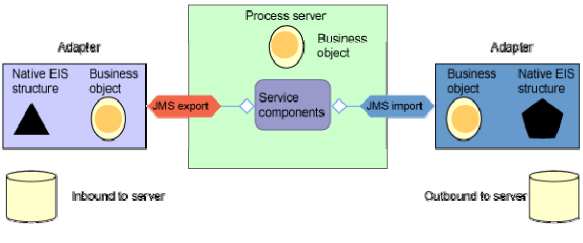


Fig. 2. Architectures for the agent type of mobile collaboration

4 Implementation of the Healthcare Application

In this section we describe a healthcare application of computer-supported mobile devices cooperation. Also, we implement the healthcare application based on TMO scheme and then show the monitoring of GUI results in collaboration environment.

4.1 Healthcare Application

The healthcare application focuses on the collaboration environment which includes the interaction of management server, patient and nurse.

The hospital has a three wards and each ward located stationary-typed agent on fixed hosts. When nurse's MOBILE DEVICES move about the ward, moving-typed agent adds network components and requests the environment information, patient health information to stationary-typed agent in the ward. And then, it provides information which is collected data from sensors, in which case the security information cannot access any information. The mobile devices of nurse's displays information that is collected data from the stationary type agent. With mobile collaboration application, we used the environment information from sensors such as temperature, humidity and illuminometer. Also, we used the location sensor for location tracking and healthcare sensor such as blood pressure, blood sugar, heart rate and body heat. They can be divided into two types such as security and public. The one is patient health information and the other is environment information in the wards. The moving-typed agent can be obtained public information in the wards, and provide the control of appliances.

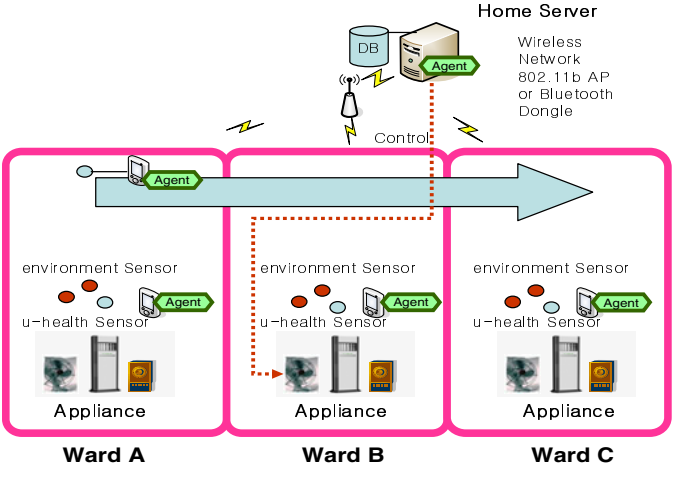


Fig. 3. The system environments of healthcare application for mobile collaboration

For example, if using the moving-typed agent, we want to control fan to regulate the temperature, it requires the certification for the control of appliance to home server. Also stationary-typed agent on the ward obtains the public information and

security information. And then, it is sent to home server. When information are transmitted a collecting data by agent on mobile devices, the home server manages the way data are stored to database. Figure 3 shows the physical environment for mobile collaboration.

4.2 Definition of Healthcare Application Component

The components of the healthcare application based on mobile collaboration framework are defined by the TMO scheme. And we used the TMOSM for interactions between distributed components. Figure 4 describes the interaction of the components for healthcare application.

For the healthcare application, distributed application components of the system(Client, Server) includes the distributed object implemented by TMO scheme and agents that autonomously act according to the perceived context information. In order to facilitate the implementation of autonomous agent for this healthcare Service, we used one more TMOs. Sensor TMO collects environment information, patient health information at the wards and also obtains the location information by agent. Monitoring GUI display the information about the collected data from Sensor TMO. Profile TMO manages user profile, user authority information and Control TMO have responsibility for appliance control. The Context TMO in home server, component for the appliance control service, monitors the action of all information appliances by receiving the information from corresponding appliances. The stationary-typed agents are located on MOBILE DEVICES s at wards and home server which are fixed host. The moving-typed agent is located on mobile devices which is mediate component that interact with mobile devices and PCs.

First, for the moving-typed agent on mobile devices, the Sensor TMO mapping to moving object, called a nurse, in hospital, and sensed by physical sensor (Cricket). It also senses the moving nurse by the periodic time description, stores the location information of home server into information repository. When detecting the moving object, Sensor TMO transfers the location information. And it obtains environment information and patient's health information from sensors or stationary-typed agent at ward. Control TMO controls peripheral appliances by sending those commands in their own proprietary language. In this case, the security object in DOGF needs to grant the access control right this privilege. All of the access control rights specified for the security object need this access privilege.

For stationary-typed agents on mobile devices at wards, they also collect data about the environment information and patient's health information from the environment sensors and health sensors. And then, the information transfers to the Sensor TMO in moving-typed agent.

4.3 Executing Results of Healthcare Applicaion

We have developed a set of initial healthcare agent services on the mobile collaboration framework to exercise a number of deployment scenarios and collaborative models. Figure 4 presents the healthcare application based on mobile collaboration framework as mentioned above.

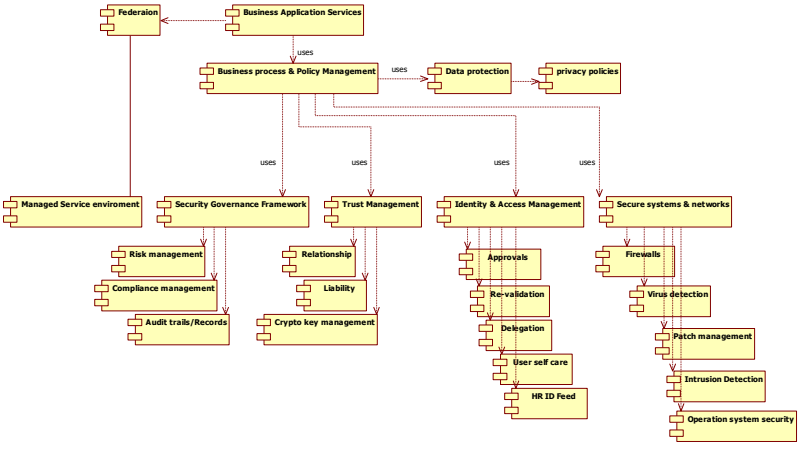


Fig. 4. Business Application Service Class diagram in our works

First, the GUI of mobile devices at ward displays collecting data from sensors by stationary type agent and setting environment which add members for collecting data from sensors. The mobile devices GUI for a nurse shows the information collected data by moving-typed agent for each ward according to location of mobile devices of nurse.

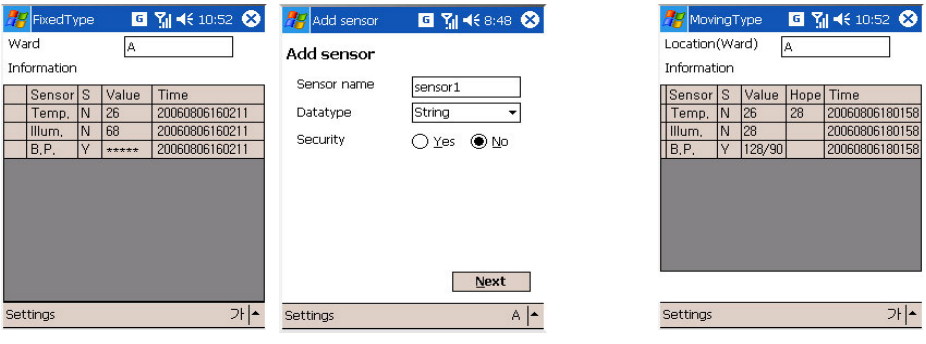


Fig. 5. GUI for mobile devices

For this, the nurse has to register the collected information to the home server. After then Figure 5 presents the home server GUI displays the collected information for each wards.

In details, patient health information and the environment information are managed in the home server. It also displays the location for mobile devices and the status information that is a live home appliance through the hope value on GUI.

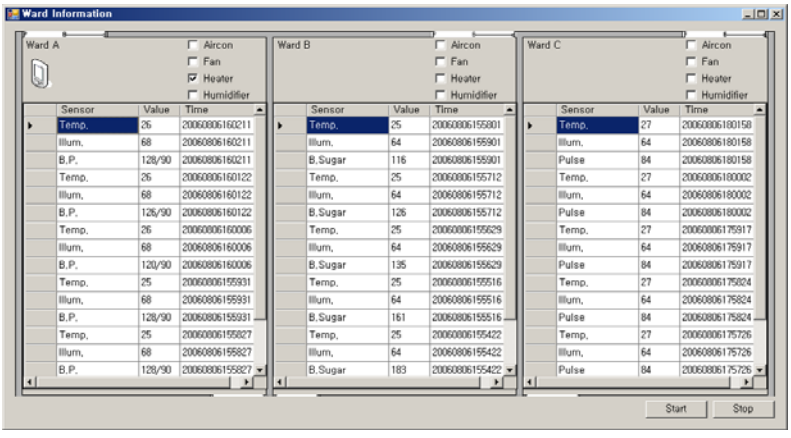


Fig. 6. GUI for integrated monitoring of Home Server

5 Conclusion and Future Works

Ubiquitous computing environment includes the various sensors, mobile devices and communication infrastructure. In this environment, research is towards the mobile collaboration that able to deal with the mobility and interactions of both users and devices. Hence, because of limitations on resource and platform, we suggest mobile collaboration framework. It is based on the distributed object group framework that supports the group service and real time service. The mobile collaboration environment includes the sensors, mobile devices and home server. We defined the interaction agent type that interacts with each other. And we applied healthcare service for mobile collaboration such as a hospital ward environment. Each component for executing functions of agent and an integrated monitoring is implemented by TMO scheme. we showed via GUI the executablility of healthcare application on our mobile collaboration framework. Our future work will apply different environments for ubiquitous healthcare service and improving the performance of the framework. We will include studying the mobile agent technologies and then, we will apply to the moving-typed agent.

Acknowledgement. This work was supported by the Korea National Research Foundation (NRF) granted funded by the Korea Government (Scientist of Regional University No. 2011-0013259).

References

1. Shin, C.-S., Jeong, C.-W., Joo, S.-C.: Construction of Distributed Object Group Framework and Its Execution Analysis Using Distributed Application Simulation. In: Yang, L.T., Guo, M., Gao, G.R., Jha, N.K. (eds.) EUC 2004. LNCS, vol. 3207, pp. 724–733. Springer, Heidelberg (2004)

2. Kim, D.-S., Jeong, C.-W., Joo, S.-C.: Implementation of Healthcare Application Service in Mobile Collaboration Environment. In: Proceedings of Korea Computer Congress 2006, vol. 33(1(D)), pp. 88–90 (2006)
3. Agarwal, D., McParland, C., Perry, M.: Supporting Collaborative Computing and Interaction. In: Proceedings of the Grace Hopper Celebration of Women in Computing 2002 Conference, Vancouver, Canada, October 9-12 (2002)
4. Heinemann, A., Kangasharju, J., Lyardet, F., Mühlhäuser, M.: iClouds – Peer-to-Peer Information Sharing in Mobile Environments. In: Kosch, H., Böszörményi, L., Hellwagner, H. (eds.) Euro-Par 2003. LNCS, vol. 2790, pp. 1038–1045. Springer, Heidelberg (2003)
5. Sun Microsystems. jxme: JXTA Java Micro Edition Project, http://jxme.jxta.org/
6. Collier, R., O'Hare, G.M.P., Lowen, T., Rooney, C.: Beyond Prototyping in the Factory of Agents. In: Mařík, V., Müller, J.P., Pěchouček, M. (eds.) CEEMAS 2003. LNCS (LNAI), vol. 2691, pp. 383–393. Springer, Heidelberg (2003)
7. Agent UML, http://www.auml.org
8. Kinny, D., Georgeff, M., Rao, A.: A methodology and modelling technique for systems of BDI agents. In: van Hoe, R. (ed.) 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (1996)
9. Wooldridge, M., Jennings, N.R.: Intelligent agents: theory and practice. The Knowledge Engineering Review 10(2), 115–152 (1995)
10. Bergenti, F., Poggi, A.: LEAP: A FIPA Platform for Handheld and Mobile Devices. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, pp. 436–446. Springer, Heidelberg (2002)
11. Bellifemine, F., Rimassa, G., Poggi, A.: JADE—a FIPA compliant agent framework. In: Proceedings of the 4th International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-Agents, London (1999)
12. Foundation for Intelligent Physical Agents (FIPA), http://www.fipa.org
13. Albuquerque, R.L., Hübner, J.F., de Paula, G.E., Sichman, J.S., Ramalho, G.L.: KSACI: A Handheld Device Infrastructure for Agents Communication. In: Meyer, J.-J.C., Tambe, M. (eds.) ATAL 2001. LNCS (LNAI), vol. 2333, pp. 423–435. Springer, Heidelberg (2002)
14. Johansen, D., et al.: A tacoma retrospective. Software Practice and Experience 32(6), 605–619 (2002)
15. Baümer, C., et al.: Grasshopper—a universal agent platform based on OMG MASIF and FIPA standards. In: Proceedings of the First International Workshop on Mobile Agents for Telecommunication Applications (MATA 1999), Ottawa, Canada, pp. 1–18 (1999)
16. The Agent Oriented Software Group, http://www.agent-software.com
17. Pister, K.S.J., Kahn, J.M., Boser, B.E.: Smart dust: wireless networks of millimeter-scale sensor nodes. Highlight Article in Electronics Research Laboratory Research Summary (1999)
18. O'Hare, G.M.P., O'Grady, M.J.: Gulliver's Genie: a multi-agent system for ubiquitous and intelligent content delivery. Computer Communications 26(11), 1177–1187 (2003)
19. Cheverst, K., Mitchell, K., Davies, N.: The role of adaptive hypermedia in a context-aware tourist guide. Communications of the ACM 45(5), 47–51 (2002)
20. Bauer, B.: UML Class Diagrams Revisited in the Context of Agent-Based Systems. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.) AOSE 2001. LNCS, vol. 2222, p. 101. Springer, Heidelberg (2002)
21. Foundation for Intelligent Physical Agents (FIPA), The FIPA 2000 Specifications (2000), http://www.fipa.org

22. Shen, S., O'Hare, G.M.P., Collier, R.: Decision-making of BDI agents: a fuzzy approach. In: Proceedings of the 4th International Conference on Computer and Information Technology (CIT), Wuhan, China (September 2004)
23. O'Grady, M.J., O'Hare, G.M.P., Sas, C.: Mobile tourists, mobile agents: a user evaluation, Gulliver's Genie, Interacting with Computers (submitted for publication)
24. O'Grady, M.J., O'Hare, G.M.P.: Just-in-time multimedia distribution in a mobile computing environment. IEEE Multimedia 11(4), 62–74 (2004)
25. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. IEEE Computer 36(1), 41–50 (2003)

Design and Implementation of Smart Meter Concentration Protocol for AMI

Byung-Seok Park[1], Cheoul-Shin Kang[2,*], and Young-Hun Lee[3]

[1] Green Growth Lab. KEPCO Research Institute, Munji-dong, Yusung-gu, Daejon 306-791, Korea
blueon@kepco.co.kr
[2] Dept. of Electronic Eng., Hannam University, Ojeong-dong, Daedeok-gu, Daejon 306-791, Korea
ckang@hnu.kr
[3] Dept. of Electronic Eng., Hannam University, Ojeong-dong, Daedeok-gu, Daejon 306-791, Korea
yhlee@hnu.kr

Abstract. The information of existing smart meter is transmitted by the fixed concentration and scheduling protocol on the request of the server. The connection configuration and the data transfer process are performed by DLMS/COSEM standard protocol. This protocol is applied to only point to point connections between the meter and the metering server. Therefore some constraints of transmission efficiency occur in large-scale networks. In this paper, we suggests an open protocol - Smart Meter Concentration Protocol(SMCP) which provide efficient transmission between concentrator and metering server. The AMI system was implemented with the proposed SMCP and has field tested in Jeju Island.

Keywords: AMI, SMCP, metering protocol, task script, job scheduling.

1 Introduction

Recently, as the concern over the greenhouse effect increases, smart grid, a type of electrical grid combined with IT technologies has been considered as a promising solution for the energy saving and the rational energy consumption. This has resulted in the development of many “smart” technologies. Among these, the AMI(Advanced Metering Infrastructure) system is the most focused technology.

The AMI system differs from traditional AMR system in that two-way communications are enabled with the meter in real-time in order to manage the energy demand with time-of-use pricing.[1]

In Korea, the traditional AMR system is coming into use and the AMR system polls the meter data. This polling method lacks in flexibility and thus makes upgrade of the AMI system hard.

* Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 179–187, 2011.

In this paper, we propose a new protocol appropriate for AMI system to support all the functions of the smart meter and to meet increasing requires of smart grid and introduce a field test to examine the protocol.

2 Introduction of AMI System

2.1 Network Architecture of AMI

Figure 1 shows a configuration of the AMR system using PLC and B-CDMA jointly. The DCU(Data Concentration Unit) installed on a pole periodically concentrates meter data from meters with PLC or B-CDMA modems. And the aggregated data is transmitted to the AMR server through WAN(wide area network) such as HFC and D-TRS.

The PLC modem and the B-CDMA modem achieve different performance depending on surroundings and power system configurations. The PLC modem is commonly appropriate for areas where power lines are overhead, while the B-CDMA modem is appropriate for rural areas or areas where power lines are underground.

One of the basic objectives of the smart grid network is bi-directional communication through the power transmission and distribution system. Recently, the metering data is concentrated at every 15 minutes in quasi-realtime in order to manage the energy demand with time-of-use pricing. Furthermore, AMI(Advanced Metering Infrastructure) system includes multiple devices such as IHD(In-home Display) and EMS(Energy Management System) in order to provide the real-time pricing and DR(Demand Response). Therefore higher speed communication is necessary, and consequently BPL(Broadband over Power Lines) system and high speed B-CDMA technology operating at 2.4GHz have been used in the industrial field.

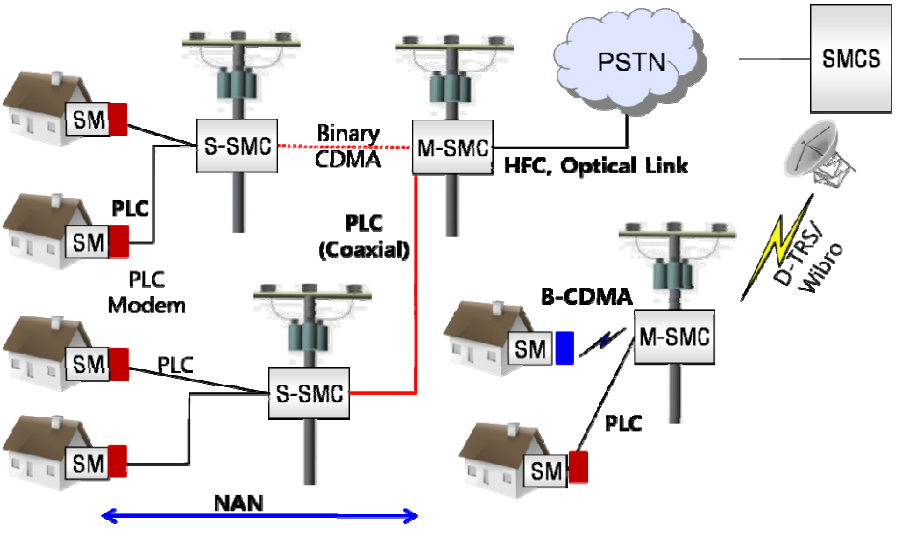


Fig. 1. Network structure of KEPCO AMI system

2.2 Function and Protocol of AMI Component

2.2.1 Function of Smart Meter(SM)

SM records consumption of electric energy and transmits that information to SMC(Smart Meter Concentrator). HDLC protocol referred in COSEM/DLMS published by IEC[2,3] is used for two-way communication between SM and SMC. For differential pricing depending on the weather and the time, SM loads TOU(Time-of-Use) table from SMC and reads usage information at times with a preset period.

2.2.2 Function of Smart Meter Concentrator (SMC)

SMC is an active device to collect meter data. Figure 2 shows the protocol stack between SM and SMC. PLC or B-CDMA modem only delivers meter data and never process the DLMS packet as a member of the NAN(Neighborhood Area Network). The modems are, therefore, a type of bridge and operate only as data link protocol and application protocol to encapsulate the DLMS packets. The functions of SMC include collecting meter data, monitoring abnormal consumption, reporting equipment failure, reporting interruption of power supply and reporting the result of SMC self-test.

Typical SMC are installed at field. Therefore, it is impossible to modify concentration strategies such as the kind of information to be collected, collecting period and retry method. The collected data are transmitted to SMCS in a lump.

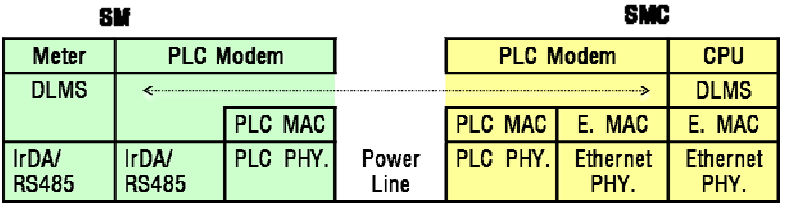


Fig. 2. Protocol stack of SM-SMC communication

2.2.3 Function SMCS

SMCS is connected with SMCs through long-distance wireless network and communicates using TCP/IP technology. The meter data from a number of SMCs are periodically polled by SMCS in consecutive order.

Trap packet to report unusual condition is the only case that SMC transmits spontaneously. Figure 3 shows the communication flow between SMCS and SMC. Polling requires scheduling of a number of meters and SMCs in the AMI system and thus imposes large burden on the system. And complex electric load dispatch is also needed.

Generally, the communication protocol between SMCS-SMC is privately defined and operated. KEPCO also defined and has operated SMCP(SMC Protocol).

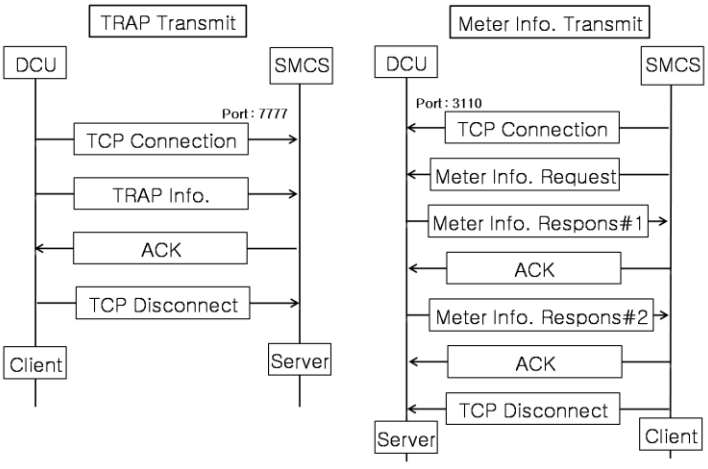


Fig. 3. Protocol flow of between SMC and SMCS

3 Design of SMCP

3.1 Concept of SMCP

Communication using SMCP, that is an object-oriented protocol, starts with the transmission of the task-script. SMCP has good parameter scalability and compatibility with other types of systems in smart grid due to exchangeability of object-oriented data such as CIM(Common Interface Model) and DLMS data[2-4].

Figure 4 shows the basic communication concept of the SMCP. SMC interprets the task-script from SMCS and operates SMCP according to the interpretation. The task-script includes the kind of information to be collected, collecting method, and list of SMs to communicate. SMC holds multiple task-script according to communication types. Meter data from SMs are collected and transmitted with the period and method defined in the task-script.[5]

3.2 Introduction of SMCP

SMCP is the text-based protocol using the ISO 10646 ASCII code in the UTF-8 encoding rules[6]. The text-base protocol facilitates editing and adding parameters and is used to HTTP, RTSP, SMTP, SIP and etc,. SMCP is composed of request messages and response messages. Request messages and response messages are linked with the same XID value marked on the head line. Generally, the character separator is used to recognize the end of message [6].

A task_script is a request message consisting of several detailed jobs. It specifies schedule and operation condition of each job. A job is a unit task to be performed by

SMC. It consists of get/set of SM and SMC, and report and notification to SMCS. Each job has a detailed attribute of {operation period, priority, execution time, report cycle} for efficient operation.

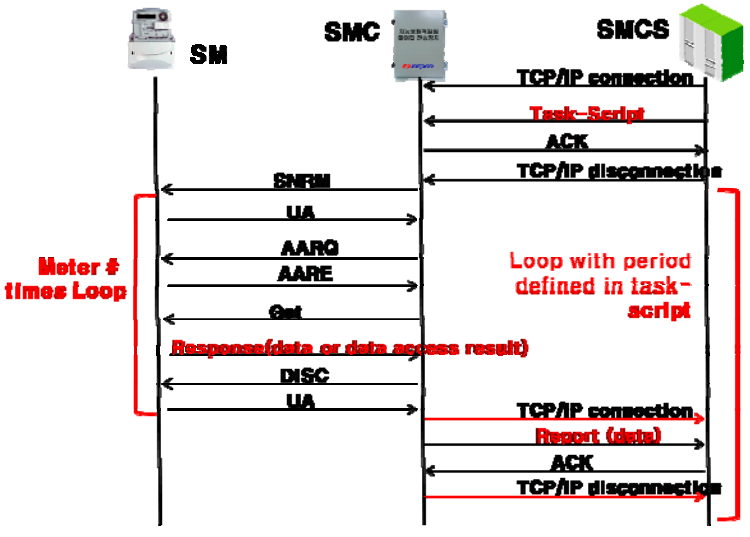


Fig. 4. Concept of SMCP

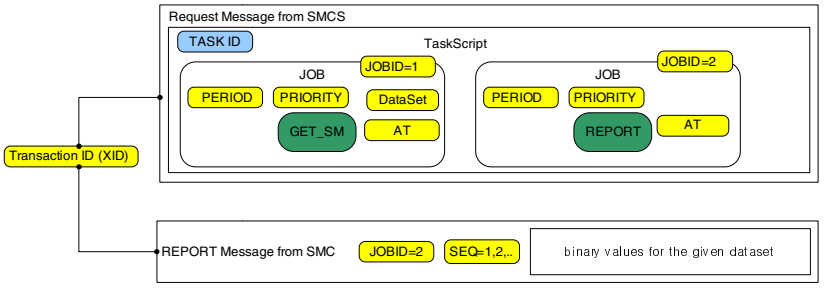


Fig. 5. Job configuration concept in SMCP

A task is formed with individual jobs. SMCP is a protocol which specifies a series of processes in which SMC performs the job based on the task_script received from the server.

Basically, the server sends the task_script on the request message to SMC. SMC interprets the task_script and collects SM data, and sends the result on the response message to SMCS. The following table lists the example of tasks. Extensionally, SMCP can deliver concentrator itself related data such as transformer monitoring data and SMC event and environment data as well as metering data from SMs.

Table 1. Type of JOB and TASK

Type Job	Description	Task Type (taskID)	Description
SM_GET	Extracts meter data periodically/non-periodically.	Periodic meter reading (0x1xxx)	Periodic data collection by SM (SM_GET) Periodic report to SMCS (REPORT)
SM_SET	Sets meter parameters.	On-demand meter reading (0x2xxx)	Selective data collection by SM (SM_GET) On-demand report to SMCS (NOTIFY)
SMC_GET TR_GET	Collects status of concentrator. Extracts transformer monitoring data.	SM setting (0x3xxx)	Notifies the result of setting of TOU, time, parameter (SM_SET) to SM (NOTIFY)
SMC_SET TR_SET	Sets concentrator action environment. Sets transformer monitoring action environment.	SMC related tr. monitoring (0x0xxx)	Notifies the result of SMC and TR status write and setting (SMC_GET/SET, TR_GET/SET) (NOTIFY)
REPORT	Reports the collected data to SMCS periodically.	Event notify /on demand reading (0x9xxx)	Notifies various events of SM and SMC (NOTIFY) On demand reading (RELAY)
NOTIFY	Notifies on-demand data under certain conditions.	-	-

The task is performed in the following procedures of scenario:

1. Upon receiving request for data from MDMS, SMCS makes view table from CIM to DLMS data.
2. SMCS creates the task_script in which SMs and metering data are configured.
3. SMCS requests to SMC to connect on TCP.
4. SMCS sends SMC task_script.
5. SMC parses the task_script.
6. SMC collects data from SM specified in the task_script using the DLMS communication protocol.
7. SMC makes the table with the collected data by SM.
8. Then, SMC sends the REPORT message to the server by task ID and transaction ID (XID) of requested task_script.

The example of request message of task_script and the response message of report are followed in Figure 6 and Figure 7, respectively.

```
SMCP/1.0 SMCID=0785A31101 XID=0x0003 TASKID=0x1200 MSGLEN=1234<CRLF>
SMCID=0785A31101 SMCP/1.0 XID=0x0003 TASKID=0x1200 MSGLEN=350<CRLF>
<CRLF>
DO=SM_GET JOBID=1 PRIO=0 EVERY=00-00-00-25-15 AFTER=30<CRLF><CRLF>
COMMENT= conduct JOB1 with 15 minute interval after 30 minute this script is
arrived at SMC <CRLF>
DATATYPE=smDataset/text DATALEN=200<CRLF>
Smlist=[08140009291,…… 08140009289,08140009287]
Objlist=[{1;1.0.0.0.0.255;2},….{1;0.127.94.82.0.8;2}]<CRLF>
DATATYPE=selectiveDataset/text DATALEN=50<CRLF> PGobj=[{7;1.0.99.1.0.255;2}]
Rel_entryrange=60 Col_range=[1,2,5,6] <CRLF>
COMMENT=collect 1,2,5,6 columns of LP data during 60 minute<CRLF>
DO=REPORT JOBID=2 AT=10-07-20-12-10 PERIOD=15 <CRLF>
<CRLF>
.<CRLF>
```

Fig. 6. Example of request packet of task_script

```
SMCP/1.0 SMCID=0785A31101 XID=0x0003 TASKID=0x1200 MSGLEN=1234<CRLF>
<CRLF>
WILL=REPORT JOBID=2 SEQ=1 STATUS=200<CRLF>
DATAVALUE=value/binary DATALEN=1100<CRLF>
--- transmit Data ----- <CRLF>
COMMENT=SMC time :10-7-20-12-10 <CRLF>
<CRLF>
.<CRLF>
```

Fig. 7. Example of response packet of report

4 Implementation and Field Test

Implementation of AMI system based on SMCP consists of newly developed SM and SMC. Microprocessor - UPLC 25 chip is used for the main component of SMC. UPLC 25 chip supports PLC communication and control application process such as AMI collecting engine. UPLC25 developed by KECO is composed of 32bit RISC processor, 24Mbps PLC module, analog to digital convertor for colleting power signal, various embedded peripheral interfaces such as I2C, SPI, Ethernet, and UART[7].

SMC is developed using C programming language based on Linux 2.6 operating system and supports PLC and B-CDMA NAN communication. SMC also supports D-TRS based on PPP interface and Ethernet interface for Wibro communication and HFC network connection for communicating with SMCS.

SMCS is developed using JAVA language. SMCS is composed of three major components. The first component is FEP(Front end Processor) for collecting and parsing SMCP packets transmitted from SMC. The second component is DB handling engine, which processes storing, searching and sorting of large bundle of SM information. The last component is Web engine for displaying data and operating system. FEP supports parallel processing of SMCP packet transmitted by a large number of SMC simultaneously, and DB engine supports data validation function due to network failure or device failure.

To examine the AMI system using SMCP, we configured the field test system shown in Figure 8 at Gujwa, Jeju Island. SMs of 10 houses are connected by PLC

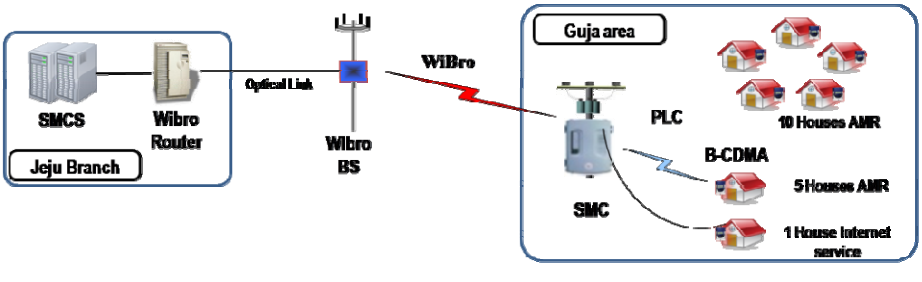


Fig. 8. Configuration of field test system

network and SMs of 5 houses are connected by B-CDMA network. SMC and SMCS are connected by Wibro network. SMC collects self reading information, cumulated active power, load profile and demand peak power with 15 minutes of period and transmits the collected data to SMCS with the same period.

Figure 9 shows the report result from SMC. The result shows the header of the response message and received meter data requested by the task-script. Finally, we can see that the meter data is successfully transmitted with the status=200.

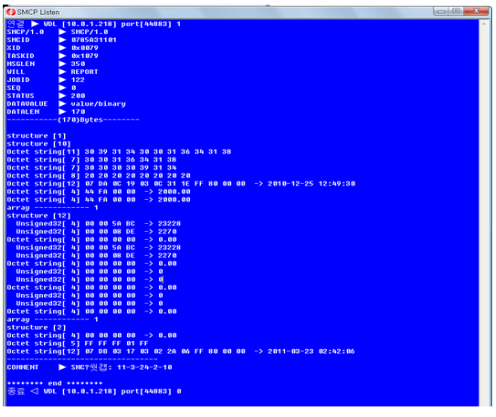


Fig. 9. Screen and Picture of field test system

5 Conclusion

In this paper, we proposed a new protocol for AMI system to overcome the limit of the traditional protocols. The proposed protocol is designed to support all the functions of the smart meter and to meet increasing requests of smart grid based task-script. We examined the AMI system using the proposed protocol in Jeju Island and the result demonstrates that the system works well without any problem.

Acknowledgments. This paper has been supported by 2011 Hannam University Research Fund.

References

[1] Advanced Metering Infrastructure, `http://en.wikipedia.org`
[2] IEC-62357, Power system control and associated communications Reference architecture for object models, services and protocols (2007)
[3] IEC-62056-53, Electricity metering - Data exchange for meter reading, tariff and load control. COSEM Application Layer (2007)

[4] IEC-62056-47, Electricity metering - Data exchange for meter reading, tariff and load control. COSEM Transport Layers for IP Networks (2007)
[5] Kim, M.-S., Ko, Y., Kim, Y., Park, B.: A Simulation Study of the PLC-MAC Performance using Network Simulator-2. In: ISPLC (2008)
[6] Choi, I., Park, B., Lee, S., Yoon, J.: Design of Smart Metering Control Protocol based on task_script for Advanced Metering Infrastructure. In: CEWIT (2010)
[7] Park, B., Kim, Y., Choi, I., Myung, N.: Development of AMI Data concentractor based on Ubiquitous PLC SoC. In: CICS, KIEE (2010)

Biologically-Inspired Optimal Video Streaming over Wireless LAN

Yakubu S. Baguda, Norsheila Fisal, Rozeha A. Rashid, Sharifah K. Yusof, Sharifah H. Syed, and Dahiru S. Shuaibu

UTM-MIMOS Centre of Excellence,
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia
{baguda_pg,sheila,rozeha,kamila,hafiza}@fke.utm.my

Abstract. There is dramatic need to achieve optimal performance in wireless multimedia network due to its heterogeneous nature, media content and different quality of service (QoS) requirements from different applications. It is very obvious that supporting multimedia applications and services over wireless is very challenging task, and it requires low complexity and highly efficient scheme to cope with the unpredictable channel condition. In this paper, we develop a biologically-inspired strategy for optimal video streaming application. The optimal parameters configuration selected provide the best QoS settings to enhance the video streaming quality over wireless LAN. The scenario has been simulated in NS-2 environment, it clearly shows that the video quality has been improve by selecting minimum configuration to ultimately support video application. The PSO-based approach outperforms other techniques used to compare the performance of the develop scheme in terms of perceived video quality by more than 0.5dB. The experimental simulation has been used to verify the efficiency and potential application of the PSO in wireless multimedia communication.

Keywords: Particle swarm optimization (PSO), Quality of service (QoS), optimal configuration, biologically-inspired, channel condition.

1 Introduction

Optimization is becoming an increasingly important tool in order to achieve optimal solution and effective resource allocation especially in time-varying transmission environment. This is primarily due to dramatic need for efficient and economical design solution, and allocating limited resources [1]. With rapid growth of optimization techniques in recent years, more complicated communication and networking problems can be solved with relative ease and sophistication. More importantly, advancement in computing has contributed tremendously toward achieving more remarkable results with high precision and accuracy. Also, the geometric increase in size and complexity of problems as a result of technological advancement has necessitated the need for

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 188–197, 2011.

more systematic and efficient approach. In a nut shell, optimization is extremely important tool used to determine the best solution to a given problem within a set of constraints.

Wireless multimedia communication has witnessed an exponential growth due to high demand; and hence it requires efficient strategies to ultimately support multimedia applications over erroneous channel. The alarming increase in utilization and number of users with different QoS requirements increase the computational complexity and time. For delay sensitive application, it is very challenging to achieve the necessary QoS required to ultimately support wireless multimedia communication [17,18,19,20]. Bio-inspired algorithms have developed mainly based on the successful evolutionary behavior of natural system - which naturally mimic nature and adapts dynamically with the environment. These algorithms can be used in solving complex problems with high precision and sophistication. In order to perform more complex computations within shortest possible time, there is need for highly efficient and sophisticated optimization technique which converges very fast and simple to be implemented. Particle swarm optimization (PSO) [2][3][21][22] has been extremely important tool which dramatically reduce the processing time and complexity as well. This is primarily due to its convergence and searching capability. The aforementioned characteristics of PSO when fully exploited can play more significant role in minimizing the time required to make critical decision. PSO is a well-known optimization technique which is based on social and cognitive behavior.

The unpredictable nature of wireless environment makes it difficult to manage the resources available within the network more effectively. Hence, there tremendous need for optimization to effectively manage the network resources properly due increasing number of wireless station which have detrimental impact on the network performance [4]. It is a known fact that multimedia applications are very much sensitive to delay and hence high priority is needed to be given to such application in order to be transported over unpredicted channel condition.

In order to support wireless multimedia communication, more than one QoS parameters should be considered to enhance the performance of the network. Hence, multi-criteria decision approach is needed and PSO is used to properly manage the conflicting multi-objective functions. Real-time optimization problems are multi-objective in nature and therefore multi-objective PSO can simultaneously and efficiently determines the best optimal solution within the packet deadline [13][14][15]. The ability to meet up with packet deadline is very important in delay sensitive application where any delay can seriously affect the multimedia content.

The remainder of this paper is organized as follows. Section 2 mainly focuses on the general overview of PSO. Section 3 introduces our proposed strategy using particle swarm optimization. Simulations results are presented in Section 4. Finally, conclusions are enumerated in Section 5.

2 Overview of Particle Swarm Optimization

Swarm intelligence has been used extensively in solving optimization problems and it primarily uses biologically inspired approach to select best optimal solution. Its

principle is purely based on simulating the movement of flock of birds or school of fish. It has been extremely important tool and can potentially tackle complex optimization problems. Many different swarm optimization techniques exist today and this primarily includes Ant, Bee, Glow warm etc. In fact, PSO has been very promising due to its fast convergence and simplicity. This can be applicable in supporting delay sensitive applications which require highly efficient and low computation complexity. It is very obvious that in order to enhance the performance of wireless multimedia applications, many parameters and factors should be considered. Hence, multi-objective optimization PSO can eventually solve variety of multi-objective problems to achieve optimal performance. More importantly, multi-objective PSO can effectively search and determine the set of optimal solutions simultaneously. The wireless channel characteristic can be represented by highly non-linear objective and constraint in order to genuinely consider its impact on the transmitted data or information.

Particle swarm optimization mainly used stochastic and population based optimization approach to mimic the natural phenomenon. It primarily produces a set of solution and further modified as the process continues. This has been very efficient and powerful tool for searching the optimal solution to a given particular problem. In general, PSO normally used to have one objective function but due to confliction between the functions require an efficient strategy to determine the best possible solution. It has been a major challenge because by increasing one of the functions can have detrimental impact on the other function. Therefore, this subsequently leads to development of multi-objective optimization to properly handle the situation. Problems dealing with multiple objective function can be solve using multi-objective PSO with relative ease and high searching capability for optimum solution [5][6]. Due to aforementioned reasons, multi-objective PSO can be used in wireless network to ultimately support multimedia applications and services.

It is important to note that PSO is based on velocity and position equations. The velocity and position of a particle at time t can be represented by v(t) and x(t). It is used in selecting the best position and fitness. The velocity and position of the particle at time t+1 can be determined using equation (1) and (2)

$$v(t+1) = \omega v(t) + c_1\big(P_l + x(t)\big) + c_2\left(P_g + x(t)\right) \tag{1}$$

$$x(t+1) = x(t) + v(t+1) \tag{2}$$

The velocity equation basically describes the velocity of the particles at time t+1. v(t) keeps track of the particle flight direction and it prevents the particle from sudden change in direction. $c_1\big(P_l + x(t)\big)$ normally measures the performance of the particle relative to the past performance. In a nutshell, it draws the particles to their best known position. $c_2\left(P_g + x(t)\right)$ measures the performance of particle relative to the neighbours. Generally, it serves as standard in which individuals want to reach. Both the cognitive and social components depend greatly on c_1 and c_2 respectively.

The global best (P_g) determines the best possible solution for the entire neighborhood for each particle in the entire swarm. It is just a typical star topology in which the social component of the particle velocity update depends on the information obtained from all the particles in the swarm [12]. In star structure, the particles are interconnected and hence they can communicate with one another more efficiently and effectively. All the particles tend towards the best solution found in the swarm. The global best used star structure which converges faster, but can be trapped in local minima. The best possible position is determined by P_g. After the velocity is determined, the position is then computed using the x(t+1). It is very important to note that this will determine best possible position and velocity discovered by any particle at time (t+1).

However, it is very important to note that meeting up with the delay constraint in wireless multimedia network is a major issue of concern. Multi-objective cross layer optimizer using PSO can potentially select the optimal configuration within the time deadline. In the next section, the multi-objective decision criterion will be elaborated.

3 Concept of Optimal Video Streaming Using PSO-Based Approach

Generally, optimization problems which have more than one objective function termed as multi-objective optimization. It is absolutely possible that the objective functions conflict with one another while optimizing such a problem [7][8]. Multi-objective optimization has different possible solution to particular problem. In this particular case, the multi-objective optimization primarily consists of objective function, constraint and optimal solution. The main optimization task is achieved through multi-objective optimization which is primarily based on the concept of particle swarm optimization. The equation (1) and (2) describes the velocity and position of the particle at any particular time. More importantly, the PSO capability depends greatly on the aforementioned functions. At any particular time, the values for these functions are computed and best possible value is selected. Also, it is very much necessary to include the constraint in order to set boundary for the search space.

In fact, wireless environment require multi-objective optimization due to multiple objective functions describing various parameters of the network. The minimization problem can be represented mathematically as

$$\text{Minimize} \quad f(x) = (f(x_1), f(x_2), \ldots., f(x_n)) \quad \text{Subject to} \quad x \in X \tag{3}$$

Let us assumed that $f: R^n \rightarrow R$ and hence $x \in R^n$ can be determine if the condition $f(x) \leq f(\boldsymbol{x}), \forall x \epsilon R^n$ has been satisfied.

3.1 Problem Formulation

The objective functions are selected to provide the necessary QoS support for multimedia application over wireless network. The optimal video quality problem primarily determine the best parameter settings for minimum queue size, and delay at particular and subsequently assign how the policy to use in transmitting the packet. It is aimed at setting the configurations which yields optimum performance at relative low queue size, and delay.

The joint sets of optimal solutions of the optimization problem can be represented mathematically in matrix form by

$$\begin{bmatrix} \varphi_1 \\ . \\ \varphi_n \end{bmatrix} = \begin{bmatrix} Q_1 & D_1 \\ . & . \\ Q_n & D_n \end{bmatrix} \tag{4}$$

φ represents the optimal solution and n is the total number of solutions obtained. At each particular time, sets of optimal solutions are obtained and any of the solution can be potentially yields optimal in video quality.

Equation (5), and (6) describes the queue size, and delay.

$$\boldsymbol{Q_i} = \frac{1}{N}\left(\sum_{i=1}^{n} q_i\right) \tag{5}$$

$$\boldsymbol{D_i} = \frac{1}{N}\left(\sum_{i=1}^{n} d_i\right) \tag{6}$$

The queue size and delay objective function are minimize subject to the delay constraint as in equation (8). The joint optimal video streaming quality ϑ_{opt} can be achieve by finding the minimum values for the queue, and delay at a particular time. The objective function is given as $f(q,d) = 0.5 * f(q) + 0.5 * f(d)$.

$$\text{Minimize} \quad \varphi_{opt} = argmin\, f(q,d) \quad \text{Subject to:} \tag{7}$$

$$0 \leq d \leq d_{total} \tag{8}$$

The above parameters are computed within the time delay limit d which should be less than the overall packet deadline d_{total} for the video application. It is very clearly that the algorithm shows more remarkable results by converging below the packet deadlines which indicated that the algorithm can used for delay sensitive applications.

The biologically inspired algorithm for optimal video streaming application can be summarized as follows

Algorithm: PSO-Based algorithm for optimal video streaming quality

Objective function: ***min*** *f(q,d)*
Initialize swarm with random positions and velocities of n particles

for loop over all n particles

Compute the values of $\boldsymbol{Q_i}$, *and* $\boldsymbol{D_i}$ *at time t*
Generate new velocity
Calculate new position based on the new velocity

// Evaluate objective functions f(q,d)
Check delay constraints if less than packet deadline;
Compute the fitness of the particle based on f(q, d)
Select local best;
Select global best;

if current fitness of particle is less than previous and global, replace the previous
Update particles velocities and positions
end
Increase the loop counter, $\boldsymbol{t+1}$
end // once stopping condition has been reached;
Output $\boldsymbol{\varphi_{opt}} = argmin\, f(q,d)$

4 Simulation Result and Discussion

In this section, the simulation results for PSO-based optimal streaming scheme is presented. The optimizer parameters have been set based on the particle swarm to suit the QoS requirement for video application. The maximum number of iterations and inertia weight were set to 30 and 0.5 respectively within which the optimal solution should determine or terminated. The cognitive c_1 and social c_2 constants have been set to 1. The delay constraint is checked before evaluating the fitness function for each objective function. Both local and global best are determined and subsequently the position and velocity of the particles are updated. If the condition for termination has been reached, the current best optimal parameter configurations are selected. When the termination condition has not been reached, the loop counter is compared with the number of particles. The process is repeated until the termination condition is reached.

The simulation has been conducted using NS-2 based on the settings in table 1 and 2. The code for the PSO-based algorithm is integrated into NS-2 environment in order to select the best solution to optimization problem. We simulated the scenario using three QCIF (176X144) video samples which are encoded at 30fps. Hall and News video samples were used in the simulation to effectively analyze the performance of the scheme. The proposed scheme is compared with static and rate distortion optimization schemes to verify the performance of the developed approach.

Table 1. PSO parameter settings

Parameters	Value
Number of particles	30
Number of iteration	30
Learning factors C1 & C2	1
Inertia weight	0.5

Table 2. Network settings

Parameters	Value
Propagation model	Shadowing
PhyType	Phy/WirelessPhy/802_11
MacType	Mac/802_11
freq_	2.4 GHz
Data rate	1 – 11 Mbps
CSThresh_	1-13
RXThresh_	1.10765e-11
Video Samples (QCIF)	Hall, News
Number of frames	300

Based on the experimental results obtained, it is very obvious that the scheme can be applicable for wireless video streaming application which eventually requires reliable and robust mechanism to meet up with the delay constraint. Simulations show that the proposed technique increases the media quality. The scheme has improved video streaming quality caused mainly due to time varying channel condition, heterogeneity and congestion. The proposed scheme uses queue and delay statistic in real-time and subsequently adjusts the transmission rate in order to meet up with the constraint of the video content. When compared to RDO, the PSO-based has improved the video quality of hall and news video samples by 0.99 and 0.5 dB respectively.

Table 3. Video quality comparison

Different Samples Video Quality (dB)		
Technique	**Hall**	**News**
Static	31.29	31.87
RDO	31.88	32.79
PSO-Based	33.87	33.29

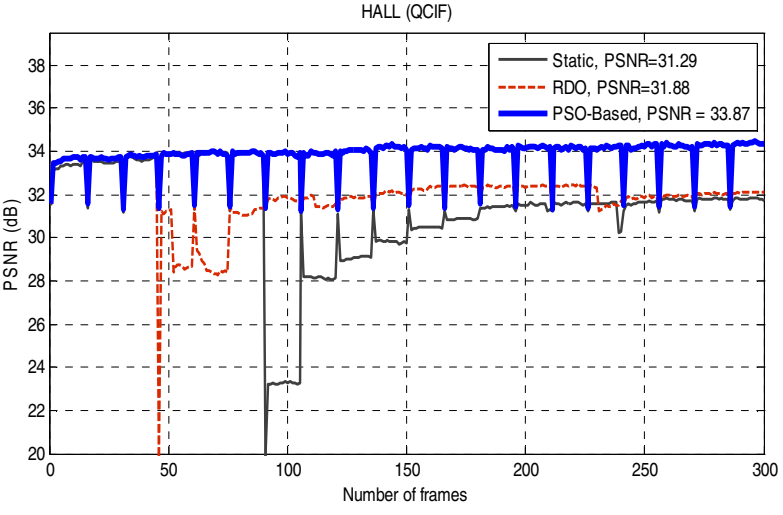


Fig. 1. Comparison of PSO-Based optimal video streaming scheme with other strategies HALL (176X144 @ 30fps)

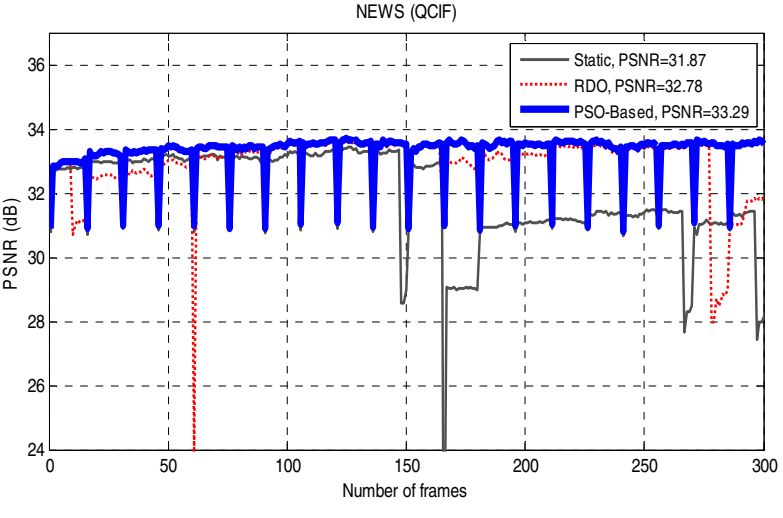


Fig. 2. Comparison of PSO-Based optimal video streaming scheme with other strategies NEWS (176X144 @ 30fps)

From fig. 1 and 2, it can be seen that the video quality has been reasonably maintained in both hall and news video samples. This is primarily due to adaptive capability of the scheme to time varying environment. Also, it can be notice that as the complexity of the video samples increases, the perceived video quality decreases. The PSO-Based scheme features such as flexibility, adaptability, modularity and scalability has tremendously assisted in adapting with wireless channel condition.

5 Conclusion

In this paper, we proposed an efficient strategy to select best possible optimal solution in multi-criteria decision problem in order to ultimately support delay sensitive application. It fundamentally presents a new approach to reduce the complexity of cross layer and at the same time achieved better video quality. More importantly, the PSO-Based approach has shown more promising result for achieving high convergence which meets up with the time constraint for video application. The biologically inspired approach using PSO is very effective, simple, flexible and high searching capability which can potentially be used in making critical decision in error prone transmission environment. Our future work will adapts this technique in developing multi-objective cross layer optimization for wireless video streaming application and services.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Non-linear programming: Theory and Algorithms. Wiley-Interscience Publishing (2006)
2. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proc. IEEE International Conference on Neural Networks, Australia (1995)
3. Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kauffman Publishers, California (2001)
4. Van der Schaar, M., Chou, P.A.: Multimedia over IP and Wireless network: Compression, networking and systems. Academic Press (2007)
5. Coello, C.A., Pulido, G.T., Lechuga, M.S.: Handling Multiple Objectives with Particle Swarm Optimization. IEEE, Evolutionary Computing (2004)
6. Huang, V.L., Suganthan, P.N., Liang, J.J.: Comprehensive Learning Particle Swarm Optimizer for Solving Multi-objective Optimization Problems. International Journal of Intelligent System (2006)
7. Zitzler, E., Laumanns, M., Bleuler, S.: A tutorial on evolutionary multi-objective optimization (2004)
8. Coello, C.A.C., Lamont, G.B.: Application of Multi-objective Evolutionary Algorithms. World Scientific Publishing (2004)
9. Van der Schaar, M., Tekalp, M.: Integrated multi-objective cross-layer optimization for wireless multimedia transmission. IEEE (2005)
10. Deb, K.: Multi-objective Optimization Using Evolutionary Algorithm. Wiley (2008)
11. Baguda, Y.S., Fisal, N., Shuaibu, D.S.: Multi-objective Particle Swarm Optimization for Wireless video Support. International Journal of Recent Trends in Engineering (2009)
12. Løvberg, M., Krink, T.: Extending Particle Swarm Optimisers with Self-Organized Criticality. In: Proceedings of the IEEE Congress on Evolutionary Computation (2002)
13. Fieldsend, J., Everson, R., Singh, S.: Using unconstrained elite archives for multiobjective optimization. IEEE Transactions on Evolutionary Computation (2003)
14. Hu, X., Eberhart, R.C.: Multiobjective optimization using dynamic neighborhood particle swarm optimization. In: Proceedings of the Evolutionary Computation (2002)
15. Parsopoulos, K., Vrahatis, M.: Particle swarm optimization method in multiobjective problems. ACM (2000)
16. NS-2 Network Simulator, `http://www.isi.edu/snamlns/`

17. Licandro, F., Schembra, G.: Wireless Mesh Networks to Support Video Surveillance: Architecture, Protocol, and Implementation Issues. Eurasip Wireless Communication & Networking (2007)
18. Bourouha, M., Ci, S., Brahim, G.B., Guizani, M.: A Cross Layer Design for QoS Support in the 3GPP2 Wireless Systems. In: IEEE Communication, Globecom Workshop (2004)
19. Kofler, I., Timmerer, C., Hellwagner, H., Ahmed, T.: Toward a MPEG-21 Based Cross layer Multimedia Content Adaptation. IEEE Computer Society (2007)
20. Nafaa, A.: Provisioning of multimedia services in 801.11 – Based Networks: Facts & Challenges. IEEE Wireless Communication (2007)
21. Baguda, Y.S., Fisal, N., Syed, S.H., Latiff, L.A., Yusof, S.K., Rashid, R.A., Shuaibu, D.S.: Low Complexity PSO-Based Multi-Objective Algorithm for Delay-Constraint Application. In: ICIEIS 2011, Kuala Lumpur, Malaysia (2011)
22. Rashid, R.A., Baguda, Y.S., Fisal, N., Sarijari, M.A., Yusof, S.K.S., Ariffin, S.H.S.: Optimizing Achievable Throughput for Cognitive Radio Network using Swarm Intelligence. In: IEEE Proc. APCC 2011, Kota Kinabalu, Sabah, Malaysia (2011)

Emerging of Mobile Ad-Hoc Networks and New Generation Technology for Best QOS and 5G Technology

Jahangir Khan[1], Zoran S. Bojkovic[2], and Muhammad Imran Khan Marwat[3]

[1] Department of Computer Science & IT
Sarhad university of science and Information Technology peshawar 25000, Pakistan
[2] Faculty of Transport and Traffic Engineering
University of Belgrade, Serbia
[3] Research & Development
Higher Education Commission
Sector H-9, Islamabad 44000 Pakistan
jahangir.csit@suit.ed.pk, z.bojkovic@yahoo.com,
mimarwat@hec.gov.pk

Abstract. Enhancing in performance, strong bandwidth utilization and merging of MANET (3G) and 4G make a plate- form to get best Quality of services. Purpose of this research is to understand the functioning of ad hoc networks and implement proposed routing protocols for mobile ad hoc networks. After analyzing the existing QOS framework with respect to the dynamic and rapidly changing behavior of ad hoc networks. This research attempts to present some of the fundamental routing issues such as decentralization, bandwidth, MANETS delay in DSR and AODV, data delivery ratio in intermediate nodes in static and dynamic position and resource constrained in ad hoc network. Routing of data packets between source to destination is difficult task, So to overcome on these problems we have to extend and evaluate proposed routing protocols (AODV and DSR) suitable for mobile ad hoc networks for best QoS, In order to prove its correctness and efficiency evaluation of the proposed protocols should be done theoretically and implemented through simulation using OPNET simulator. The research will compares the statistics of ad hoc routing protocols by different researches with different simulating tools. 5G can be achieved by making a revolutionary attempt on the basis of this research ideology (merging of ad hoc and cellular technology under 4G umbrella with user centric concept wirelessly with multi terminals at a time), which will assure optimization in the next generation technology.

Keywords: MANETS, DSR, AODV, 4G, QOS, Performance evolution, comparisons.

1 Introduction

Mobile Ad hoc Networks are the organized and enhanced form of the original Mobile Packets Radio Network (PRNET), where packet switching was used as the main source of mobile network. The history of mobile ad hoc network can be classified

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 198–208, 2011.

into three categories; i.e. first generation, second generation and third generation of Ad hoc network. Nowadays mobile ad hoc network comes in third generation.

The first generation of mobile ad hoc network starts from 1971 with the Packets Radio Network (PRNET). The second generation of Ad hoc networks was advanced and enhanced implementation of Survivable Adaptive Radio Networks (SURAN) in 1980s. This enhanced Ad hoc network communication provides environment for mobile nodes in battle field without any fixed infrastructure or central location. The second generation of Ad hoc brings a lot of changes in the improvement of radio performance by making them smaller, cheaper, and resilient to other electronic equipments. The development in computer equipments and the commercial arrival of laptop computers and other enhancements in mobile technology during 1990s gave birth to third generation of Ad hoc networks. The researchers proposed the idea of a collection of mobile nodes in ad hoc network. Mobile Ad hoc network working group was formed by Internet Engineering Task Force (IETF) in mid 1990s to standardized routing protocols for the newly adopted technology. Mobile Ad hoc working group proposed reactive and proactive routing protocols for MANETS [1].

Now turning to 4G networks is a combination of various wireless access networks. To combine different wireless access networks for 4G is to combine different techniques that is including Mobile IP and cellular IP integration, fast handoff , mobility management for all-IP networks, end-to-end Multi-path and routing optimization[2] .

The simulations have to show the sudden change in delay, traffic, and link breakage occurs due to nodes mobility. In my six nodes scenario, we have only two mobile nodes which effect greatly in routing. The size of the network, traffic load, and delay effects both AODV and DSR routing protocols. DSR routing protocol is not efficient for large networks with many mobile nodes and high load in terms of traffic and delay which will increase overhead. In such situation AODV routing protocol will be ideal because of its hop-by-hop routing. Simulation of TORA is in progress as well as combined simulation of these routing protocols for best performance and data throughput.

The paper is organized in such a fashion, section 2 briefs for new generation networks and QOS issues, section 3 shows performance based routing protocols , section 4 discuss simulation analysis, second last section briefs about 4G Technology and DSR/AODV Routing Protocols and last one concludes the paper.

2 New Generation Networks and QOS Issues

Like 4G, Mobile Ad-hoc Networks (MANETS) is also in a developing stage. While defining the MANET standard, the Internet Engineering Task Force(IETF) is working on routing techniques, defined in allowing self configuring network of mobile nodes with routing capabilities. MANET standardizes the static and mobile techniques of creating mesh networks using available wireless technology. Currently, 802.11a/b/g/n

wireless networks defined by the IEEE standards are being used in applications such at homes, offices, as disaster recovery, conferences, lectures, emergency situation in hospitals, meetings, crowd control, and battle fields. and also could be found in the initial MANET infrastructure. Thus, what limits cellular networks and WLANs will limit MANET as well. 4G is the emerging technology of future wireless networks. For the past years, many researchers and scientists from all over the world have been working on projects funded by governments and business institutions whose goals are efficient wireless networks by merging all current technologies and adapting new solutions for the enhanced telecommunication which provides superior quality, efficiency, and opportunities where wireless communications were not feasible. Some researchers define 4G as a significant improvement of 3G where current cellular networks' issues will be solved and data transfer will play more significant role. For others, 4G unifies cellular and wireless local area networks and introduces new routing techniques, efficient solutions for sharing dedicated frequency band, and increases mobility and bandwidth capacity.

The fourth generation of cellular communication systems, generally known as 4G, is the emerging technology of future wireless networks. For the past years, many researchers and scientists from all over the world have been working on projects funded by governments and business institutions whose goals are efficient wireless networks by merging all current technologies and adapting new solutions for the enhanced telecommunication which provides superior quality, efficiency, and opportunities where wireless communications were not feasible. Some researchers define 4G as a significant improvement of 3G where current cellular networks' issues will be solved and data transfer will play more significant role. For others, 4G unifies cellular and wireless local area networks and introduces new routing techniques, efficient solutions for sharing dedicated frequency band, and increases mobility and bandwidth capacity [5] as shown in fig 1.

Fig. 1. Multiple faces of 4G

There are two paths leading toward each other whose goal is 4G. One path defines the evolution of 3G cellular systems into more advanced 4G technology that will recognize and be supported by Wi-Fi standards and upcoming wireless networks technologies. The other path successfully deploys high bandwidth and introduces high speed mobility emerging from currently popular Wi-Fi technology and upcoming standards such as WiMax 802.16 supported by mobility amendment 802.16e and additional projects like 802.20 considering mobile broadband wireless networks. For the second group of people leading toward 4G, cellular networks are additional supporting component offering complete integrity with all available wireless connections [5] as given in fig 2.

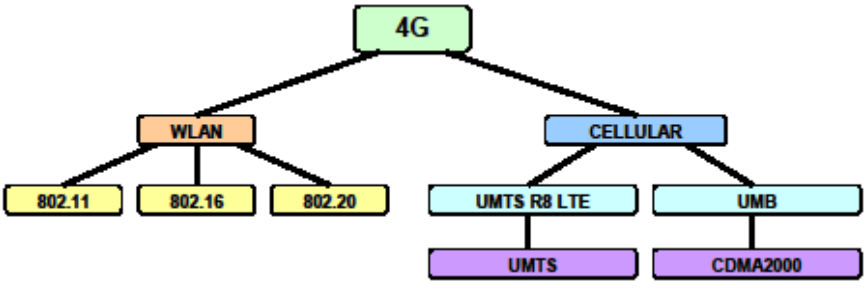


Fig. 2. Evolution of 4G

Like 4G, Mobile Ad-hoc Network is also in a developing stage. While defining the MANET standard, the Internet Engineering Task Force(IETF) is working on routing techniques, like Ad-hoc On-Demand Distance Vector (AODV) defined in and Dynamic source Routing Protocol (DSR) defined in allowing self configuring network of mobile nodes with routing capabilities. MANET standardizes the static and mobile techniques of creating mesh networks using available wireless technology. Currently, 802.11a/b/g/n wireless networks defined by the IEEE standards are being used in applications such at homes, offices, as disaster recovery, conferences, lectures, emergency situation in hospitals, meetings, crowd control, and battle fields. and also could be found in the initial MANET infrastructure. Thus, what limits cellular networks and WLANs will limit MANET as well.

The widely accepted definition of QoS is defined by the consultative committee for international telephony and telegraph (CCIT) recommendation E.800 as “the collective effect of service performance which determines satisfaction degree of a service user”. The maturity of wireless mobile technologies and the evolution of different applications provide a reason for the introduction of QoS in wireless ad hoc networks. The goal of QoS routing in MANET is to select routes with sufficient resources for data packets with QoS requirements to increase possibility that network will be capable of supporting and maintaining them. The following fig 3. shows the QoS components.

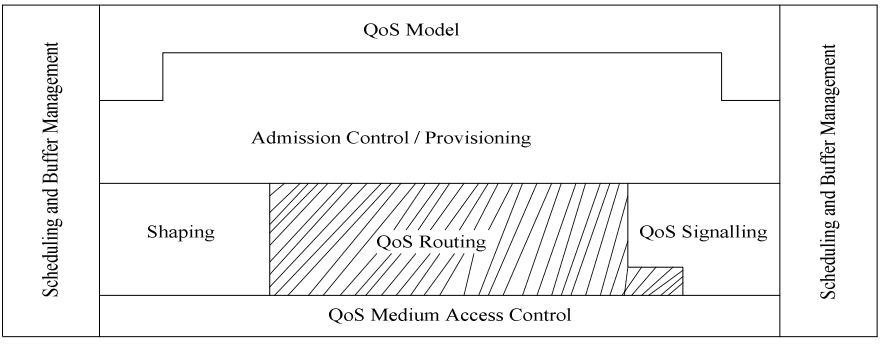


Fig. 3. QoS Components in Routing [6]

The QoS routing in mobile ad hoc network is measured with respect to the following attribute.

- Available bandwidth
- Control message overhead
- Data throughput
- Probability of packet loss
- Table storage overhead
- Delay (setup connection)
- Delay (end-to-end)
- Service coverage area
- Low power operation

Thus from above discussion it is clear that achieving QoS in mobile ad hoc network corresponds to a real need and is difficult as compared with traditional wired networks. QoS is essential element in routing which informs source node about successful availability of destination node. QoS guarantees in mobile ad hoc networks are difficult because of the following elements;

- Dynamic network topology
- Imprecise state information
- Lack of central administration
- No pre-existing base station / infrastructure less network
- Hidden terminal problem
- Limited resources availability
- Energy constrained operation
- Multicasting
- Variable capacity link
- Weak physical security

Understanding the significance of unifying Wi-Fi, WiMax and Cellular networks into one product and to propose that the most important factor of 4G will be “seamless integration of wireless networks” based on flexibility of the software radio

technology, with improved bandwidth capacity, and improved routing techniques allowing multi-hop peer-to-peer networks. So the proposed handover management system make a very strong contribution in these technology, to switch from one cell to another without dropping call, keep on the QOS that may compromise with 4G as well as Long Term Evolution-Advanced up to 100Mbps bandwidth.

3 Performance Based Routing Protocol Model

For best routing between wireless nodes of ad hoc network a number of qualitative properties is desirable. For best and effective routing loop freedom is very necessary for data packets to avoid collision and waste of time. MANET uses time to live (TTL) to avoid such loops, but more structured approach is required to get best results. The MANET delay effects on routing very badly, as the mobile nodes are battery powered and scarce memory. The routing protocol needs to be routed intelligently and utilize network bandwidth and energy resources in a better way in case of network route delay.

One of the important characteristics of MANET is dynamic topology based on node mobility. Mobile nodes in ad hoc network are free to change position frequently, therefore routing protocol needs to quickly adopt topology changes. Due to dynamic mobility and limited energy resources a unique wireless node in MANET cannot be trusted for auto configuration in case of break down. Therefore support of distributed operation is required by routing protocols to solve such type of problems. Due to lack of physical security, MANET protocols are highly exposed to different types of attacks. In MANET it is very easy for attacker to disturb network traffic, corrupt packet header, change addresses of routing messages, and increase traffic to waste bandwidth. To avoid all these threats a sufficient security protection is highly desirable in all MANET routing protocols.

Limited energy resources become scarce when the protocol does not support sleep period. As MANET nodes are battery powered and the energy used in both sending and receiving of packets. The routing protocol should be smart enough to adjust sleep period by itself without user command. As discussed, that nodes of MANET are generally limited in resources such as bandwidth, energy, storage capacity, security, and CPU capacity. Therefore the routing protocol should design to eliminate or minimize control traffic and other periodic messages. Source routing protocol DSR is quite good in route establishment and route maintenance phases but not as good in routing data packets. Where hop-by-hop routing protocol is good in transmission of data packets and also to design and implements a routing protocol that is a combination of source routing and distance vector hop-by-hop routing.

4 Simulation Analysis

4.1 Global Statistical Analysis of DSR and AODV Routing Protocols

Researcher's simulation results show that DSR performs well for medium mobility of nodes while TORA is suitable for very dense networks having high mobility ratio due

to provisions for multipaths and multicast routing. AODV is the winner being suitable for networks of different density and mobility [6] use a simulated network of GloMoSim simulator (a simulator by UCLA) with 50 nodes covering an area of 1000x1000 m2 for CBR traffic. They compare the performance of AODV and DSR for average packet delivery rate under the situation of multiple source -- common destination communication sessions. The performance of both protocols degrades i.e. performance of DSR drops by 10% and that of AODV by 30% under such constrained situation. This performance degradation is due to the fact that both protocols drop packets when they are unable to transmit data on active routes. In case of AODV, the congestion on routes to the same destination result in packet drops which is perceived as link breakage by the protocol. RERR message is sent to the source. Since stale routes are removed from routing tables and only the updated routes are maintained, source node initiates the rout discovery procedure. This degrades the packet delivery ratio. In DSR protocol however, alternate routes in maintained in route caches. The congestion problem is therefore solved locally by referring to alternate routes from route caches. Therefore the performance degradation of DSR is not as severe as that of AODV and hence DSR outperforms AODV in constrained situations [7-13] compare the performance of DSR and AODV routing protocols. They perform a number of experiments while changing the network size, load, and mobility. The experiments are simulated in ns2 simulator with two configurations i) a network size of 50 nodes in an area of 1500 x 300 m2 and ii) a network size of 100 nodes in an area of 2200 x 600 m2. The two protocols are evaluated against the metrics of average end-to-end delay, packet delivery ratio, and normalized routing load. DSR shows a better performance in terms of delay and throughput for small networks with low mobility and lesser load.

AODV on the other hand, is suitable for larger network or networks with high load or mobility. The reason of performance degradation of DSR for larger networks is due to the fact that DSR maintains information about multiple routes to a destination and there is no mechanism to remove information about stale routes. We use OPNET simulation model for comparison of DSR and AODV routing protocols using the metric of packet delivery ratio. A network of 6 nodes is simulated to cover an area of 4000x3000 m2. We conclude that size of network, mobility of node, and traffic load affect the performance of both protocols. Simulation results show that performance of DSR degrades with an increase in the size of network as well as traffic load. In such situations, AODV shows better performance due to hop-by-hop routing.

The high delay AODV occurs during first attempt for route discovery is very less, which is 0.10 per second. To compare this delay with the routing data discussed above, it is looking very less. The graph also shows that delay occur during second and third time route discovery occurs in ad hoc network. The simulation results shows that during communication there are some times where the delay is zero, such as once after the first route establishment and second in between 230 and 240 seconds. The MANET delay of AODV routing protocol is shown in following fig 4.

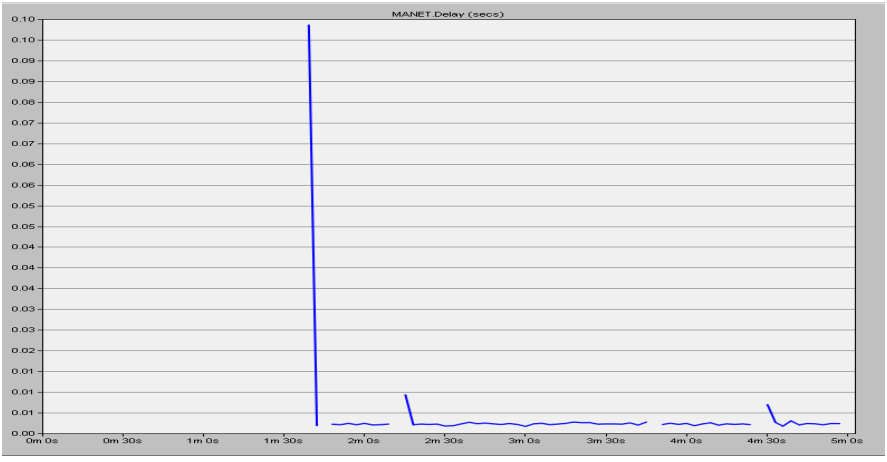


Fig. 4. MANET Delay AODV

The total delay occurs in DSR protocol data transmission during simulation period is shown in the following fig 5. The delay is a bit high in start of the route establishment process, which is 0.0065 per second. The graph shows that delay becomes suddenly high and low during whole simulation period. The DSR delay is low because there is no Hello messages to send by nodes and routing traffic wait for that. MANET delay of DSR routing protocol is very low as compared to AODV routing protocol as shown in the following fig 5.

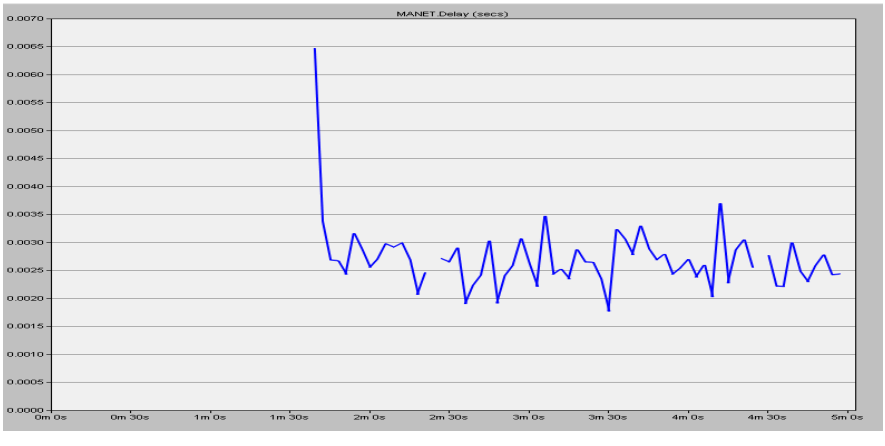


Fig. 5. MANET Delay in DSR

4.2 Research Routing Protocols Simulation Analysis

Here we propose the simulated analysis for AODV and DSR using OPNET v.12 simulator in order to make a contribution for best and efficient results in the field of

MANETS for best performance, data throughput, data delivery ratio from source to destination. Here the statistics on which we are going to improve the performance of each protocols, having Movement space 4000m x 3000m, Maximum speed is 2, 10, and 20 m/s each, Maximum pause time are 0 and 200s and Transmission rates are 2, 5, and 10 packets/s with packet size is 512 bytes for 300s evaluates the performance of important TCP parameters for AODV based 30 nodes in a small campus network.

The lifetime of a routing table entry is initialized to active route timeout and is defined as 5 seconds for this scenario. If a route is not used and refreshed within defined period of time, AODV marks route as invalid and removes it from IP common table. To establish a route for the first time, originating node uses a TTL start value which is 1 for the first RREQ packet. The individual statistics of source and destination nodes in simulated scenario. The number of hops covered by source node is 2 where destination node is 1. The AODV packets are queued during route establishment, when route is established then packets are transmitted without any waiting in queue. Both nodes routing table contains same number of nodes information. The routing traffic sent and received by both source and destination nodes are also shown in the figures. The routing traffic received by all wireless nodes of ad hoc network during simulation time. The traffic received individually by destination node is less than source node. The high traffic received by nodes occur in selected route are mobile node 2 and intermediate node 2. As discussed above that intermediate nodes have high routing traffic because of Hello messages. During simulation time these nodes receive more than 600 bits/sec traffic.

During transmission of data in simulated ad hoc network routing traffic sent by all wireless nodes. The route selected by DSR protocol in simulated scenario for transmission of data between source and destination nodes consists of four nodes, i.e. source, mobile node 2, intermediate node 2, and destination. Source node sent traffic only once during start of transmission. During reply by destination node a lot of traffic is added to the data. The intermediate node 2 and mobile node 2 adds nearly same data traffic during simulation time. The source node sent less than 100 bits/sec traffic where a mobile and intermediate node 2 sent more than 500 bits/sec traffic.

5 4G Technology and DSR/AODV Routing Protocols

4G networks are the fourth generation cellular communication systems and are still in the development stage. To provide efficient, fast, and secure communication over the cellular networks, 4G networks promise meeting these demands by the use of Internet Protocol (IP). The International Telecommunication Union (ITU) has set the speed standard for 4G networks as 100Mbps for high mobility and 1Gbps for low mobility situations.

Till this time, cellular networks have seen the transition from first generation to the fourth with advancement in the transmission technology. 1st generation (1G) voice-only networks, being completely analog, expired in the wake of digital era. Replacing the 1G networks, 2G networks are based on GSM technology and encode the radio

signals digitally, enabling the transmission of voice as well as data. With 2G networks, short messaging-service (SMS), conference calling, and web browsing facilities are introduced to the users of 2G networks. These networks enable communication to take place at the speeds of 10-20 Kbps. 2.5G and 2.75G networks, which are GPRS and EDGE based systems respectively, provide enhanced speed for faster communications i.e. around 384 and 473 Kbps, respectively. However, these extensions of 2G networks still cannot improve the speed to allow for applications such as video conferencing, voice over IP, live streaming of video, and the alike. The "need for speed" thirst for these multimedia applications is satisfied by 3G networks which provide speeds of DSL standard i.e. beyond 2 Mbps. 3.5G networks further improve the speed many folds (around 14.4 Mbps) using the HSDPA (High Speed Downlink Packet Access) technology. However, this quest for speed is never ending. While 3G networks have not yet become pervasive because the infrastructure is still on the built, 4G networks have evolved in the mean while, which are supposed to provide a bandwidth of 100Mbps – 1Gbps to allow for high quality streaming of live videos, 3D gaming, and faster web surfing. Here we discuss 4G networks and the implication of AODV and DSR protocols for 4G networks which may contribute for future generation technology.

6 Conclusion

Still the fast growing technology needs attention in many areas such as routing, bandwidth, security, power consumption, collisions, simulations, and topology control due to moving nodes. In mobile ad hoc network we cannot use the existing techniques for better quality as we are using in fixed wireless network for better efficiency. The limited transmission range of mobile wireless network nodes relies on multiple hops to exchange information between the different nodes in network.

In this master research we conclude that DSR and AODV can improve the correctness, delay, performance and efficiency by using OPNET simulator and also combined performance of AODV and DSR routing protocol could be best solution for routing in MANET instead of separate performance to get QOS.

Based on IEEE 802.xx wireless mobile networks, the main contribution of this paper is definition of 5G mobile network concept, which is user-centric concept instead of operator-centric as in 3G and service-centric for 4G. The 5G terminals have software defined radios and modulation scheme along with new error-control schemes in moving position to introduce the wireless world wide web(www). The development is seen towards the user terminals as a focus of the 5G mobile networks. The terminals will have access to different wireless technologies at the same time and the terminal should be able to combine different flows from different technologies. Each network will be responsible for handling user mobility, hands-off while the terminal will make the final choice among different wireless/mobile access network providers for a specific service.

References

[1] Bakht, H.: Wireless Infrastructure: Critical ah-hoc networking features (2005), http://www.zatz.com/authors/authorpages/humayunbakht.htm

[2] Salleh, R., Li, X.: Handoff Technique for 4G Mobile Wireless Internet. In: SETIT 2005 3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, Tunisia, March 27-31 (2005)

[3] Ram, S.C., Manoj, S.B.: Ad-Hoc wireless Networks, Architecture and Protocols, 1st edn. Prentice Hall, NJ (2004)

[4] Performance Evaluation: Running DSR and TORA Routing Protocol Concurrently. This work was supported by the National Science Foundation-Division of Undergraduate Education under grant # 0516432 from IEEE digital library

[5] Szczodrak, M., Kim, J., Baek, Y.: 4GM@4GW: Implementing 4G in the Military Mobile Ad-Hoc Network Environment. IJCSNS International Journal of Computer Science and Network Security 7(4) (April 2007)

[6] Feeney, M.L.: Introduction to MANET Routing.Swedish Institute of Computer Science (2005), http://www.nada.kth.se/kurser/kth/2D1490/05/lectures/feeney_mobile_adhoc_routing.pdf

[7] Lin, C.: (N/A) AODV Routing Implementation for Scalable Wireless Ad Hoc Network Simulation (SWANS), http://jist.ece.cornell.edu/docs/040421-swans-aodv.pdf

[8] OPNET, Modeler Wireless Suite for Defence (2007), http://www.opnet.com/products/modeler/home.html

[9] Koziniec, T.: OPNET Modeler My research and experience (2007), http://www.koziniec.com/research/presentations/opnet.PDF

[10] Frodigh, M., Johansson, P., Larsson, P.: Wireless Ad Hoc Networking –The Art of Networking without a Network. Ericsson Review 77(4), 248–263 (2000)

[11] Misra, R., Manda, C.R.: Performance Comparison of AODV/DSR On-demand RoutingProtocols for Ad Hoc Networks in Constrained Situation. Indian Institute of Technology, Kharagpur (India)

[12] Das, S.R., Perkins, C.E., Royer, E.M.: Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks. In: Proceedings of the IEEE Conference on Computer Communications (INFOCOM), March 2000, pp. 3–12 (2000)

[13] Lee, S.: Routing and Multicasting Strategies in Wireless Mobile Ad Hoc Networks, University of California Los Angeles (2000), http://www.sigmobile.org/phd/2000/theses/sjlee.pdf

Intelligent Hybrid Anomaly Network Intrusion Detection System

Heba F. Eid[1], Ashraf Darwish[2], Aboul Ella Hassanien[3], and Tai-hoon Kim[4]

[1] Al-Azhar University, Faculty of Science, Cairo, Egypt
heba.fathy@yahoo.com
[2] Helwan University, Faculty of Science, Cairo, Egypt
amodarwish@yahoo.com
[3] Cairo University, Faculty of Computers and Information, Cairo, Egypt
aboitegypt@gmail.com
[4] Hannam University, Korea
taihoonn@hannam.ac.kr

Abstract. Intrusion detection systems (IDSs) is an essential key for network defense. The hybrid intrusion detection system combines the individual base classifiers and feature selection algorithm to maximize detection accuracy and minimize computational complexity. We investigated the performance of Genetic algorithm-based feature selection system to reduce the data features space and then the hidden naïve bays (HNB) system were adapted to classify the network intrusion into five outcomes: normal, and four anomaly types including denial of service, user-to-root, remote-to-local, and probing. In order to evaluate the performance of introduced hybrid intrusion system, several groups of experiments are conducted and demonstrated on NSL-KDD dataset. Moreover, the performances of intelligent hybrid intrusion system have been compared with the results of well-known feature selection algorithms. It is found that, hybrid intrusion system produces consistently better performances on selecting the subsets of features which resulting better classification accuracies (98.63%).

Keywords: Network security, Intrusion detection system, Feature selection, Hidden naïve bays, Genetic algorithm.

1 Introduction

Network security needs to be carefully concerned to provide secure information channels. Intrusion detection system (IDS) is a major research problem in network security. The concept of IDS was proposed by Anderson in 1980 [1]. The IDS goal is to dynamically identify unusual access or attacks to secure the networks [2,3]. IDS systems divided into two techniques: misuse detection and anomaly detection [4,5]. Misuse detection (signature detection) uses well-defined patterns of the attacks to identify intrusions [6,7]. Anomaly detection attempts use normal behavior profile. It identifies attacks based on the deviations from

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 209–218, 2011.

that normal profile [8]. The anomaly detection techniques have the advantage over the misuse detection technique by detecting unknown attacks [9].

Several machine-learning techniques including fuzzy logic [10], neural networks [11], support vector machines (SVM) [8,10] and Bayesian networks [12] have been proposed for the design of IDS. In particular, these techniques are developed to classify whether the incoming network trances are normal or intruder.

One of the most important problems for ID is dealing with data containing large number of features. The classification accuracy is effected by the features subset selected for training dataset. Therefore, feature selection is required to deal with a large feature set. Different feature selection methods are proposed to enhance the performance of IDS [13]. Genetic algorithm (GA)[14,15,16] is one of successfully global search algorithm used to solve the feature selection tasks.

In this paper, we propose an anomaly hybrid network intrusion detection system using two machine learning techniques. Genetic algorithm is used first as feature selection followed by Hidden naïve bays (HNB) classifier. The effectiveness of the proposed hybrid anamoly network intrusion detection system is evaluated by conducting several experiments on NSL-KDD network intrusion dataset. Also, GA as a feature selection method is compared with other well known feature selection techniques. The rest of this paper is organized as follows: Section 2 gives an overview of Genetic algorithm and HNB architecture. Section 3 describes the proposed hybrid anolamly intrusion detection system. The experimental results and conclusions are presented in Section 4 and 5 respectively.

2 Genetic Algorithm and Hidden Naïve Bays: An Overview

This section give an overview of genetic algorithm and HNB algorithm.

2.1 Genetic Algorithm

Genetic algorithm (GA) is an adaptive search technique initially introduced by Holland [18]. It is computational model designed to simulate the evolutionary processes in the nature [19] GA includes three fundamental operators: selection, crossover and mutation within chromosomes.

1. **Selection:** A population is created with a group of randomly individuals. The individuals in the population are then evaluated by fitness function. Two individuals (offspring) are selected for the next generation based on their fitness.
2. **Crossover:** crossover randomly chooses a point in the two selected parents and exchanging the remaining segments of them to create the new individuals.

3. **Mutation:** mutation randomly changes one or more components of a selected individual. This process continues until a suitable solution has been found or a certain number of generations have passed [20].

Given a well bounded problem GAs can find a global optimum which makes them well suited to feature selection problems.

2.2 Hidden Naïve Bays

Jiang et al [17] proposed hidden nave Bayes (HNB). HNB inherits its structural from naive Bayes. It creates a hidden parent for each attribute to combine the influences from all other attributes. Hidden parents are defined by the average of weighted one-dependence estimators.

Assume that $A_1, A_2, ..., A_n$ are n attributes. An instance E is represented by a vector $< a_1, a_2, ..., a_n >$, where a_i is the value of A_i. C is the class variable and c(E)represent the class of E.

In HNB each attribute A_i has a hidden parent A_{hpi} , $i = 1, 2, ..., n$, The joint distribution represented by an HNB is given by:

$$P(A_1, ..., A_n, C) = P(C) \prod_{i=1} P(A_i \mid A_{hpi}, C) \tag{1}$$

where

$$P(A_i \mid A_{hpi}, C) = \sum_{j=1, j \neq i}^{n} W_{ij} * P(A_i \mid A_j, C) \tag{2}$$

and

$$\sum_{j=1, j \neq i}^{n} W_{ij} = 1 \tag{3}$$

the weight W_{ij} is compute directly from the conditional mutual information between two attributes A_i and A_j

$$W_{ij} = \frac{I_p(A_i; A_j \mid C)}{\sum_{j=1, j \neq i}^{n} I_p(A_i; A_j \mid C)} \tag{4}$$

where

$$I_p(A_i; A_j \mid C) = \sum_{a_i, a_j, c} P(a_i, a_j, c) log \frac{P(a_i, a_j \mid c)}{P(a_i \mid c) P(a_j \mid c)} \tag{5}$$

The HNB classifier on $E = (a_1, ..., a_n)$ is define as follows

$$c(E) = arg \max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i \mid a_{hpi}, c) \tag{6}$$

3 An Intelligent Hybrid Network Intrusion Detction System (GA-HNB System)

The proposed hybrid anomaly intrusion detection system is using the advantages of hidden naïve bayes in conjunction with genetic algorithm-based feature extraction to detect and classify the network intrusions into five outcomes: normal and four anomaly intrusion types. It is comprised of the following three fundamental building phases: (1) Preprocessing In the first phase of the investigation, a preprocessing algorithms were used to map symbolic features to numeric value and attack names were mapped to one of the five classes. It is adopted and used to improve the quality of the data and to make the feature extraction phase more reliable. (2) Feature extraction-based genetic algorithm phase: In the second phase, genetic algorithm has been used as feature selection to reduce the dimensionality of the dataset. (3) detection and classification using hidden naïve bays phase: The last phase is the intrusion detection and classification of a new intrusion into five outcome. These three phases are described in detail in the following section along with the steps involved and the characteristics feature for each phase. Figure (1) Depicts the overall architecture of the introduced system.

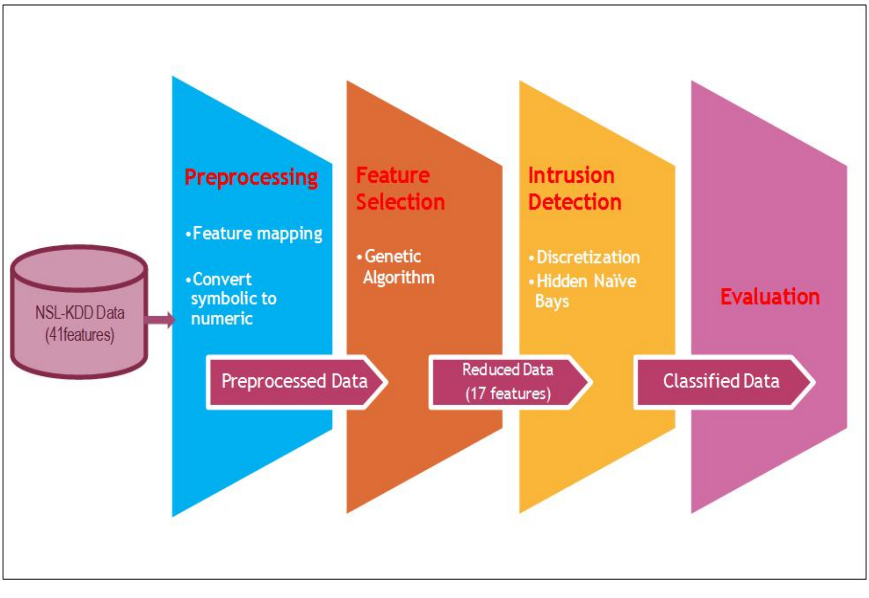


Fig. 1. The overall architecture of the anomaly hybrid network intrusion detection system

3.1 Preprocessing Phase

The following two pre-processing stages has been done on NSL-KDD dataset:

1. Mapping symbolic features to numeric value.
2. Attack names were mapped to one of the five classes, 0 for $Normal$, 1 for DoS (Denial of Service), 2 for $U2R$ (user-to-root: unauthorized access to root privileges), 3 for $R2L$ (remote-to-local: unauthorized access to local from a remote machine), and 4 for $Probe$ (probing:information gathering attacks

3.2 GA Feature Selection Phase

Feature selection is a process that reduces the feature space by eliminating unnecessary features to classification. There are two main feature selection approaches, filter methods and Wrapper methods [13]. Filter method does not rely on any particular learning algorithm. It assesses each feature independently, and then a subset of features from the feature space is selected according to the ranking between the features. Wrapper method relies on some learning algorithm to estimate the features subset. Using cross-validation it calculates the accuracy of the learning algorithm for each feature from the feature space to predict the benefits of adding or removing this feature from the feature subset.In this paper, genetic algorithm has been used as feature selection method to reduce the dimensionality of the dataset. GA efficiently reduces the NSL-KDD dataset from 41 features to 17 features, which reduces 58.5% of the feature space. Algorithm (1) shows the main steps of the genetic algorithm-based feature selection.

Algorithm 1. Genetic algorithm-based feature selection

1: Randomly Generate an initial population $M(0)$ of 41 NSL-KDD features.
2: **for** maximum number of generation or fitness threshold **do**
3: Evaluate the fitness $f(m)$ of each individual m in the current population $M(t)$
4: select the best-fit individuals for reproduction, using selection probabilities $P(m)$ for each individual m in $M(t)$
5: Generate new individuals $M(t+1)$ through crossover and mutation operations to produce offspring.
6: Replace least-fit individual with new ones.
7: **end for**
8: Return the best n features of NSL-KDD dataset.

3.3 HNB Intrusion Detection Phase

The 17 features output from the GA where discritized by the Entropy Minimization Discretization method [21]. Then, the reduced discritized dataset is passed to the HNB classifier to be classified. The parameters of the HNB classifier are estimated from the training data. The GA-HNB model is described in algorithm 2.

Algorithm 2. hybrid GA-HNB intrusion detection approach

1: Use GA to reduce the training set.
2: Discritize the reduced training set by Entropy Minimization Discretization method.
3: Use the output of the GA layer to build HNB classifier.
4: **for** each $c \in C$ **do**
5: compute P(c) from training set.
6: **end for**
7: **for** each pair of attributes A_i and A_j **do**
8: **for** each assignment a_i , a_j and c to A_i , A_j and C **do**
9: compute $P(a_i \mid a_j$, c) from training set
10: **end for**
11: **end for**
12: **for** each pair of attributes A_i and A_j **do**
13: compute $I_p(A_i; A_j \mid C)$ and W_{ij} from training set
14: **end for**
15: Run the testing dataset through the network (GA-Discritized-trained HNB) to assign a class label for each instance
16: **if** Assigned class label is equal to actual class label **then**
17: object is classifier correctly
18: **end if**
19: Calculate the classification accuracy.

4 Experimental Results and Analysis

4.1 Dataset Characteristics

NSL-KDD [22]is a dataset used for the evaluation of researches in computer network intrusion detection systems. NSL-KDD consists of selected records of the complete KDD'99 dataset [23]. Each NSL-KDD connection record contains 41 features (e.g., protocol type, service, and flag) and is labeled as either normal or an attack. The training set contains a total of 22 training attack types, with additional to 17 types of attacks in the testing set. The attacks fall into four categories:DoS e.g Neptune, Smurf, Pod and Teardrop, R2L e.g Guess-password, Ftp-write, Imap and Phf, U2R e.g Buffer-overflow, Load-module, Perl and Spy, and Probing eg. Port-sweep, IP-sweep, Nmap and Satan.

4.2 Experiments and Analysis

The classification performance of the hybrid anomaly network intrusion detection system is measured by the *precision*, *recall* and $F - measure$; which are calculated based on the confusion matrix given in Table 1.

True negatives (TN) as well as True positives (TP) correspond to a correct prediction of the IDS. TN and TP indicates that normal and attacks events are successfully labeled as normal and attacks, respectively. False positives (FP)

Table 1. Confusion Matrix

		Predicted Class	
		Normal	Attake
Actual Class	Normal	True positives (TP)	False negatives (FN)
	Attake	False positives (FP)	True negatives (TN)

refer to normal events being predicted as attacks; False negatives (FN) are attack events incorrectly predicted as normal [18].

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{9}$$

An IDS should achieve a high recall without loss of precision, where F-measure is a weighted mean that assesses the trade-off between them.

The proposed hybrid anomaly network intrusion detection system is evaluated using the NSL- KDD dataset, where 59586 records are randomly taken. All experiments have been performed using Intel Core 2 Duo 2.26 GHz processor with 2 GB of RAM.

The classification performance measurements are shown in Table 2 and 3. Table 2 shows the accuracy measurements achieved for HNB using full dimension data (41 features). Table 3 gives the accuracy measurements for the proposed hybrid anomaly network intrusion detection system with 17 dimension feature.

Table 2. HNB accuracy measurements (41-dimension feature)

Class name	TP Rate	FP Rate	Precision	Recall	F-Measure
Normal	0.881	0.008	0.962	0.881	0.92
DoS	0.994	0.003	0.995	0.994	0.994
U2R	0.992	0.002	0.984	0.992	0.988
R2L	0.979	0.01	0.934	0.979	0.956
Probe	0.993	0.013	0.953	0.993	0.973

Table 3. GHNB accuracy measurements (17-dimension feature)

Class name	TP Rate	FP Rate	Precision	Recall	F-Measure
Normal	0.98	0.008	0.989	0.98	0.984
DoS	0.995	0.003	0.995	0.995	0.995
U2R	0.981	0.001	0.982	0.981	0.981
R2L	0.97	0.005	0.913	0.97	0.941
Probe	0.991	0.002	0.985	0.991	0.988

From table 4, the testing speed of hybrid anomaly network intrusion detection system is improved which is important for real time network applications. Also, the classification accuracy achieved using hybrid anomaly network intrusion detection system is improved than using HNB as standalone classier.

Table 4. Timing and Testing accuracy comparison

	Time to build model (sec)	Test accuracy
HNB model with 41-dimension feature	1.12	97.10 %
GA-HNB model with 17-dimension feature	0.26	98.63%

Various well known filter and wrapper feature selection methods are applied to NSL-KDD dataset. The effectiveness of these methods is evaluated by HNB classifier. Table 5 gives the comparison results based on 10 fold cross-validation.

Figures 2 shows the *Precision*, *Recall*, $F-Measure$ and *overallaccuracy* of the hybrid anomaly neywork intrusion detection system versus different feature selection methods.

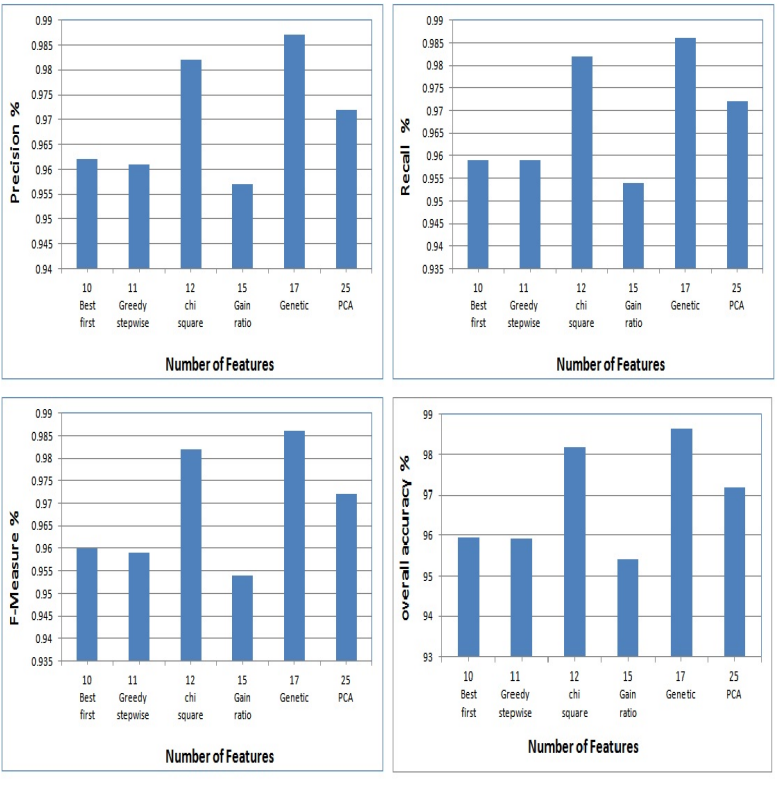


Fig. 2. An overall accuracy of the hybrid network intrusion detection system

Table 5. Different feature selection methods Performance accuracy with HNB classifier

feature selection method	Precision	Recall	F-Measure	overall accuracy %
Best first (10 features)	0.962	0.959	0.96	95.93
Greedy stepwise (11 features)	0.961	0.959	0.959	95.92
Gain ratio (15 features)	0.957	0.954	0.954	95.4019
chi square (12 features)	0.982	0.982	0.982	98.19
Genetic (17 features)	0.987	0.986	0.986	98.63

5 Conclusions and Future Works

In this paper we proposed a hybrid Genetic algorithm and HNB anomaly network intrusion detection system, where genetic algorithm is used as a feature selection and then classify the reduced data by HNB classifier. The proposed intelligent hybrid anomaly network intrusion detection system reduced the 41-dimensional of NSL-KDD dataset by 58.5% of its original size. The proposed hybrid anomaly network intrusion detection system shows high percentage of classification and enhances the testing speed due to data dimension reduction. Also, we compare the performance of the GA as a feature selection method with different filter and wrapper feature selection method as Best first, Greedy stepwise, Gain Ratio and Chi-Square.

References

1. Anderson, J.P.: Computer security threat monitoring and surveillance. Technical Report, James P. Anderson Co., Fort Washington, PA (April 1980)
2. Tsai, C., Hsu, Y., Lin, C., Lin, W.: Intrusion detection by machine learning: A review. Expert Systems with Applications 36, 11994–12000 (2009)
3. Debar, H., Dacier, M., Wespi, A.: Towards a taxonomy of intrusion-detection systems. Computer Networks 31, 805–822 (1999)
4. Biermann, E., Cloete, E., Venter, L.M.: A comparison of intrusion detection Systems. Computer and Security 20, 676–683 (2001)
5. Verwoerd, T., Hunt, R.: Intrusion detection techniques and approaches. Computer Communications 25, 1356–1365 (2002)
6. Ilgun, K., Kemmerer, R.A., Porras, P.A.: State transition analysis: A rule-based intrusion detection approach. IEEE Trans. Software Eng. 21, 181–199 (1995)
7. Marchette, D.: A statistical method for profiling network traffic. In: Proceedings of the First USENIX Workshop on Intrusion Detection and Network Monitoring (Santa Clara), CA, pp. 119–128 (1999)
8. Mukkamala, S., Janoski, G., Sung, A.: Intrusion detection: support vector machines and neural networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), St. Louis, MO, pp. 1702–1707 (2002)
9. Lundin, E., Jonsson, E.: Anomaly-based intrusion detection: privacy concerns and other problems. Computer Networks 34, 623–640 (2002)
10. Wu, S.X., Banzhaf, W.: The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing 10, 1–35 (2010)
11. Wang, G., Hao, J., Ma, J., Huang, L.: A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. Expert Systems with Applications 37, 6225–6232 (2010)

12. Jemili, F., Zaghdoud, M., Ahmed, M.: Intrusion detection based on Hybrid propagation in Bayesian Networks. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics, pp. 137–142 (2009)
13. Tsang, C., Kwong, S., Wang, H.: Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognition 40, 2373–2391 (2007)
14. Chan, K.Y., Kwong, C.K., Tsim, Y.C., Aydin, M.E., Fogarty, T.C.: A new orthogonal array based crossover with analysis of gene interactions for evolutionary algorithms and its application to car door design. Expert Systems with Applications 37, 3853–3862 (2010)
15. Zhu, Z., Ong, Y.S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition 49, 3236–3248 (2007)
16. Li, Y., Zhang, S., Zeng, X.: Research of multi-population agent genetic algorithm for feature selection. Expert Systems with Applications 36, 11570–11581 (2009)
17. Jiang, L., Zhang, H., Cai, Z.: A Novel Bayes Model: Hidden Naive Bayes. IEEE Tran. on Knowledge and Data Engineering 21, 1361–1371 (2009)
18. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. JohnWiley & Sons, USA (2001)
19. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
20. Jiang, B., Ding, X., Ma, L., He, Y., Wang, T., Xie, W.: A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms. In: The Second International Symposium on Optimization and Systems Biology (OSB 2008), China, pp. 152–157 (2008)
21. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuousvalued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, pp. 1022–1027 (1993)
22. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A.: A Detailed Analysis of the KDD CUP 1999 Data Set. In: Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Application, CISDA (2009)
23. KDD 1999 dataset Irvine, CA, USA (July 2010), http://kdd.ics.uci.edu/databases

Remote Data Acquisition and Touch-Based Control of a Mobile Robot Using a Smart Phone

Yong-Ho Seo, Hyo-Young Jung, Chung-Sub Lee, and Tae-Kyu Yang

Department of Intelligent Robot Engineering, Mokwon University,
Mokwon Gil 21, Seo-gu, Daejeon, Republic of Korea
yhseo@mokwon.ac.kr

Abstract. This paper proposes a methodology for remote data acquisition of sensor information and touch-based control of a mobile robot using a smart phone. The acquired sensor information is remotely processed by a smart phone and employed to conduct autonomous navigation. By touching the screen of the smart phone, a series of points obtained from designated curve traces are analyzed and provide control of a robot. This study develops a mobile application that acquires and handles data from a mobile robot and sends appropriate action commands through remote control using Bluetooth communication with a smart phone. The utility and performance of the proposed control scheme have been successfully verified through experimental tasks using an actual smart phone and a mobile robot.

Keywords: Remote Data Acquisition, Touch-based Control, Mobile Robot.

1 Introduction

A mobile robot is typically equipped with various sensors to recognize the surrounding environment, and its main task is generally to travel to a given destination while avoiding obstacles. In mobile robots, a PC-level computer is usually equipped for autonomous navigation. However, with the rapid increase of smart phone users, the calculation performance by smart phones has also been remarkably enhanced, and the development and application of robots using smart phones instead of PCs has become a focal point of many recent studies. In addition, current smart phones have various sensors, and research on applying this feature in robot development is being conducted. Furthermore, many robot S/W developers are also interested in application development for smart phones [1].

The present study focuses on data acquisition and navigation control of a mobile robot using a smart phone application that can be used anytime anywhere, unlike the conventional approach of employing additional high-performance computing equipment to control a mobile robot [2][3]. This study realized a remote application by using a smart phone running on Android, which is the most widely used among current smart phone OS platforms. CRX10 of CNRobot Inc. was used for a mobile robot [4]. Bluetooth communication was employed as the remote data transmission method between the mobile robot and the smart phone.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 219–226, 2011.

For the smart phone application, touch-based GUI was designed for visualization and remote control of acquired sensor information. Specifically, this study suggests a method of drawing a curved track, akin to drawing a picture by one's finger's, on the touch screen of a smart phone and applying this input to robot navigation. The suggested approach was carried out using a real smart phone for verification of the developed algorithm's performance. The smart phone robot application was developed through the use of JAVA for an Android Dalvik virtual machine in an Eclipse environment [5].

2 Mobile Robot Platform

CRX10 of CNRobot, the mobile robot platform used in this study, has numerous sensor inputs and expression functions and is appropriate for development of high-level robot applications. Accordingly, CRX10 can be controlled by an Android-based smart phone through Bluetooth. The appearance of the mobile robot platform is shown in Fig. 1 and its key specifications are listed in Table 1.

CRX10 uses Bluetooth communication as a wireless interface with an Android-based smart phone or smart pad, and sends sensor information on a real-time basis. The communication speed between the mobile robot and the smart phone is 119600bps and the information of the mobile robot is transmitted at a speed of 20ms per packet. In addition, two-way communication is possible from a minimum of 20m up to a maximum of 100m.

Fig. 1. Appearance of the mobile robot platform, CRX10

Table 1. Hardware specifications of the mobile robot platform, CRX10

H/W Item	Specification
Microcontroller	ATmega128 (main), ATmega88 (sensor)
Wheel	Four wheels driven by two DC motors
Display	8x8 LED dot-matrix
Sensor	PSD sensor - 3EA
	Flower IR sensor - 7EA
	Magnetic encoder - 2EA
	Bumper Sensor - 3EA
Communication	Bluetooth 2.4Gh / Serial RS232
Transport speed	MAX. 1m/s
Battery/Run time	Li-ion cell 11.1V 2600mA / Up to 4 hours
Charging time	2.5Hours

3 User Interface Design for the Smart Phone Application

3.1 Main Menu of the Mobile Robot Application

Fig. 2 shows the screen of the main menu of the application to choose the output functions and control types of the mobile robot through real-time Bluetooth data communication after data acquisition. The control types are composed of buttons for joy stick mode, LED mode, control mode using the smart phone's acceleration sensor, and the mobile robot's autonomous navigation mode [6]. When a button of any mode is touched, the composed layout screen appears, and action can be performed via connection through Bluetooth to the mobile robot.

Fig. 2. Main menu of mobile robot application for information output and control

3.2 Sensor Data Visualization and LED Dot-Matrix Interfaces

PSD sensors are placed on the left and right sides and at the center, on the front side of the mobile robot. The distance values between the mobile robot and an obstacle are acquired from the mobile robot and are shown on the screen of the smart phone. In addition, the information on surrounding obstacles is acquired and an avoidance algorithm is applied to realize autonomous navigation application through motor control. The basic layout of the user interface is shown in Fig. 3. The arrow in the figure indicates the proceeding direction of the mobile robot, which depends on the location of the motor. Odometric information is obtained by acquiring encoder signals when the mobile robot moves.

Among the features of the mobile robot, there is an 8x8 LED dot-matrix display. For use of this display, a touch event is received from the screen of the smart phone. Coordinate information corresponding to the touched LED is turned on both in the smart phone screen and the real dot-matrix on the mobile robot.

In the 64 LEDs comprising the dot-matrix, an 8 bit control address is assigned and images are arranged in accordance with the LED address on the screen of the smart phone. Fig. 3 shows the LED control layout for the 8x8 dot-matrix location output on the screen of the smart phone and the bit address according to the LED coordinates sent to the mobile robot through Bluetooth communication.

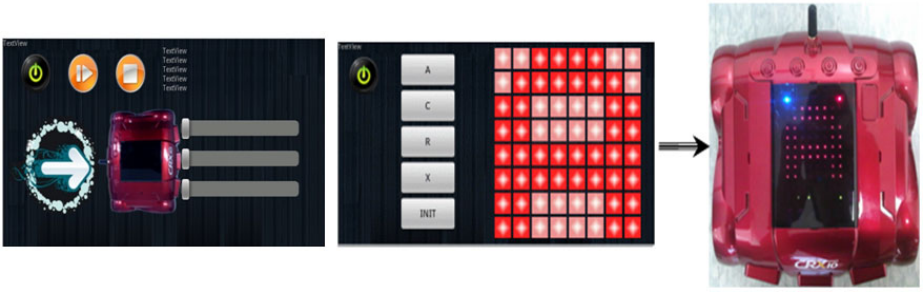


Fig. 3. PSD, Encoder sensor data output (left) UI, and 8x8 dot-matrix location output UI and indication (right)

3.3 Mobile Robot Control Interface

For controlling the mobile robot, real-time data transmission was used to send motion commands to the robot with a simple avoidance algorithm depending on the surrounding environment. Fig. 4 shows the user interface, which is comprised of a track ball and direction buttons for remote mobile robot control. In addition, buttons are arranged in line with each direction and the speed value for the motor is sent to the mobile robot only when the relevant button is pushed.

Fig. 4. Track ball and direction buttons for mobile robot remote control

3.4 Touch-Based Odometic Information Generation for Mobile Robot Navigation

The smart phone touch events that we defined are shown in Table 2, and the input order of a curve inputted by the user's touch is shown in Fig. 5.

Table 2. Types of Touch Events

Action Item	Detailed Explanation
Action_DOWN	Push the screen by a finger
Action_UP	Move a finger away from the screen
Action_MOVE	Move the finger on the screen while the finger is contacting the screen

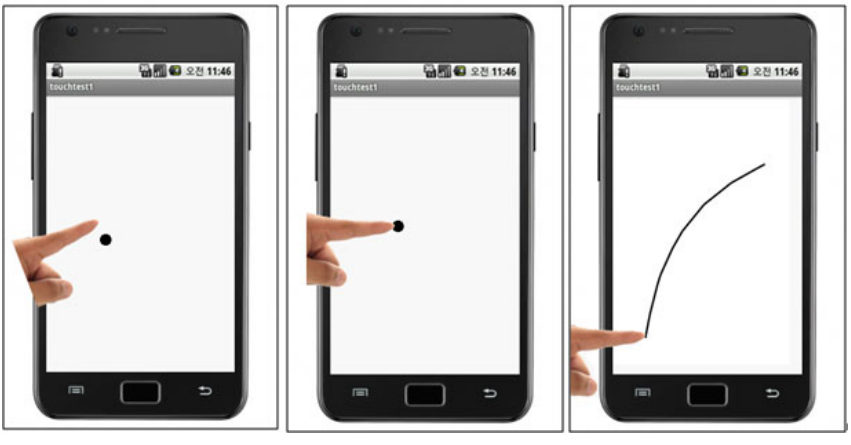


Fig. 5. Input order of a curve from a user's touch event

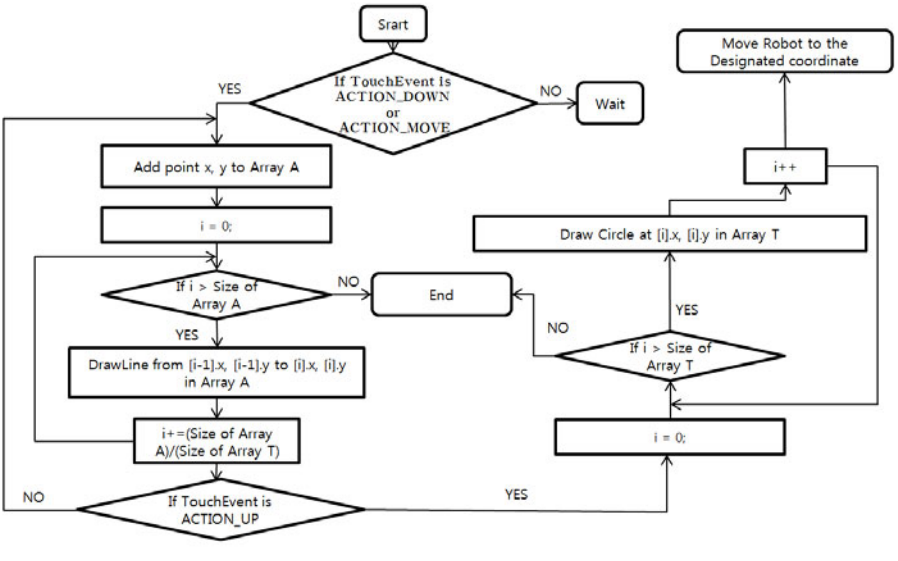


Fig. 6. Flowchart for generation of odometric information

From the point when the user begins to touch the screen with a finger, to the point when the finger is removed from the screen, the points are saved in an array. Then, from the touch curve drawn by the user, five major points are chosen and marked [7]. The straight line segments are then extracted from the saved points. Equation (1) is used to obtain a angle in consideration of n-th straight line located on cross coordinates from the values of marked points. In addition, equation (2) is used to acquire the length of each line segment. Flowchart to generate odometric information for navigation of the mobile robot is shown in Fig. 6.

$$\theta_n = \arctan(\frac{\Delta y}{\Delta x}) \tag{1}$$

$$l_n = \sqrt{\Delta x^2 + \Delta y^2} \tag{2}$$

4 Experimental Results

PSD sensors equipped in the mobile robot are used to acquire the distance value from obstacles in front of the robot, and these values are outputted on the screen of the smart phone. An obstacle avoidance experiment was conducted, as shown in Fig. 7.

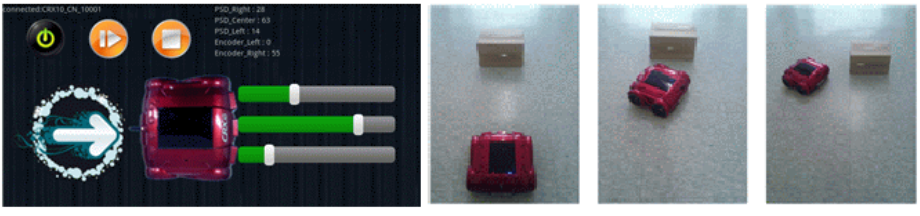


Fig. 7. Remote sensor data acquisition and obstacle avoidance experiment

When the application was tested, a touch event on the smart phone screen was registered as Action_UP, described in Table 2, and the five major points were displayed as red dots, as shown in Figure 8. At the same time, as shown in Fig. 8, after the mobile robot moved to the location designated by the coordinates and a message indicating that the mobile robot arrived at the point was sent from the smart phone. The smart phone, which received this message, sends 32 bit data by using the value of the coordinates and the angle of the next point through via Bluetooth communication. Fig. 9 shows an experiment where an actual smart phone is used to enter the robot track by touch, and followed the given odometric information of the mobile robot.

Fig. 8. Extraction of major points from the drawn curve experiment

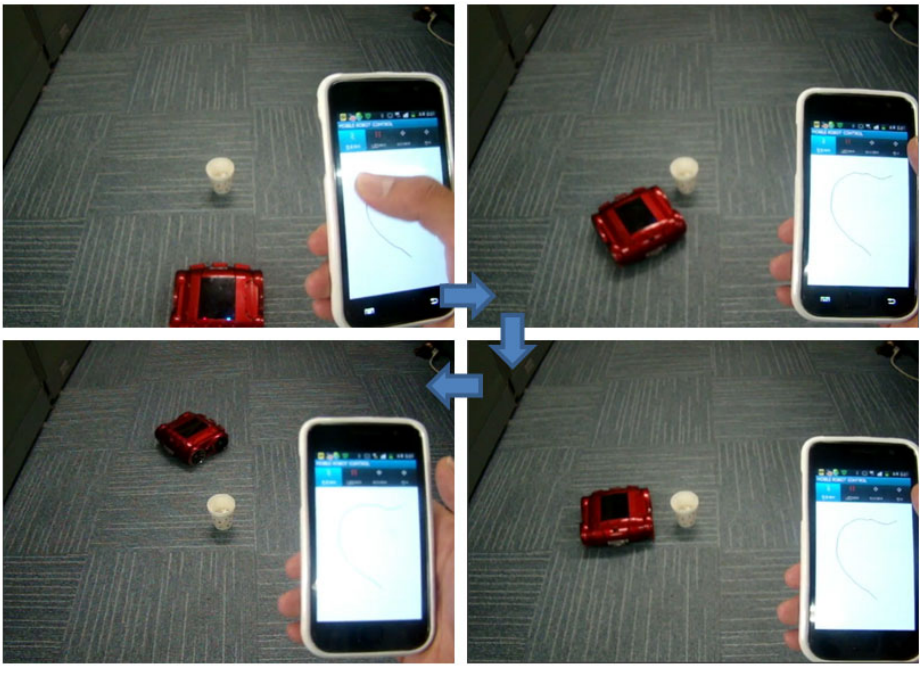


Fig. 9. Touch-based odometric information generation and mobile robot navigation experiment

5 Conclusion

This paper proposed a method to remotely acquire information on a remote robot and remotely control the robot. In addition, this study suggested technology for generation of odometric information of a mobile robot and following technology based on the use of a smart phone touch event. Simply drawing a curve with a finger provides easier way to control a robot than clicking a button or using additional sensors. In addition, as the robot moves according to marked points, the path of motion could be

accurately recognized. The robot could be controlled accurately by designating the path for the robot to follow in advance.

Lastly, from the experiments, we confirmed that it is possible to control a mobile robot through the use of remote data acquisition and the acquired the path of a mobile robot depending on the intention of users with the Android OS-based smart phone environment. Study on localization of a robot using a smart phone's GPS and acceleration sensor information will be the focus of future research.

Acknowledgments. This work was supported by the research funding program (Matching Fund) funded by the Korea Sanhak Foundation (2011).

References

1. Santos, A.C., Tarrataca, L.C., João, M.P.: The Feasibility of Navigation Algorithms on Smartphones using J2ME. Mobile Networks and Applications (MONET) 15(6), 819–830 (2010)
2. Nasereddin, H.H.O., Abdelkarim, A.A.: Smartphone Control Robots Through Bluetooth. International Journal of Research and Reviews in Applied Sciences 4(4) (2010)
3. Sparks, D.: The Smartphone Controlled Robotic Ball by Orbotix, http://walyou.com/smartphone-controlled-robot
4. CNRobot Inc., http://www.cnrobot.co.kr
5. Android developers, http://developer.android.com/guide/basics/what-is-android.html
6. de Souza, M., Carvalho, D.D.B., Barth, P., Ramos, J.V., Comunello, E., von Wangenheim, A.: Using Acceleration Data from Smartphones to Interact with 3D Medical Data. In: Proc. of the SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI 2010), Gramado, pp. 339–345 (2010)
7. Singh, S., Shin, D.H.: Position Based Path Tracking for Wheeled Mobile Robots. In: Proc. of IEEE IROS 1989, pp. 386–391 (1989)

The Performance Analysis of LT Codes*

Ling Yang, ShiLi Song, Wei Wei Su, Yi Fan Wang, and Hong Wen

National Key Lab of Communication of UESTC, Chengdu 611731, China
yang_ling106@126.com, sunlike@uestc.edu.cn

Abstract. LT codes are designed for the purpose of scalable and fault-tolerant distribution of data over wireless networks. This paper introduces the four design methods of LT codes degree distribution. Analyzing and comparison results of the short and long of LT code performance under the different channel probability are given, which show that in the case of large delete the probability, the original data packets can be recovered with high probability as long as the decoder receives a sufficient number of packets. The results of this paper will reference for system choose a different LT code parameters.

Keywords: LT Codes, Fountain Codes, Degree Distribution.

1 Introduction

Fountain codes are a class of graph-based linear erasure codes. In the broadcast communication system, the transmitter is just like a fountain, which encodes the original information to get potentially infinite encoded symbols and sends them out. The receivers can reconstruct the original information only if they have received enough encoded symbols. Fountain code is a good candidate to replace the retransmission error control method, which is especially useful for deep space communication environment.

In 1998, the concept of fountain codes was first proposed by Michael Luby[1]. Then he introduced a sparse random linear fountain codes in 2002[2]. LT (Luby Transform Codes) codes are the first realization of fountain codes on Binary Erasure Channel (BEC). In 2006, on the basis of the LT codes, Shokrollahi propose better performance of Raptor codes.

2 LT Codes Degree Distribution

Given k input symbols, fountain codes generate a potentially limitless stream of encoding symbols. Each encoding symbol is generated independently by sampling form a degree distribution and randomly exclusive-or the input symbols corresponding to sampled degree. It is highly possible to recover all the input symbols if received encoding packets are slightly more than input symbols.

* This work is supported by the NSFC (Project No. 61032003 and 61071100) and Sichuan Sci. & Tech. Support Plan (Project No. 2011GZ0183).

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 227–235, 2011.

The behavior of LT code is completely determined by the degree distribution, $\rho(\bullet)$, and the received number of encoding symbols, N, by receiver. The overhead $\varepsilon = N/K$ denotes the performance of LT code, and ε depends on a given degree distribution. First of all, we introduce the concept of degree and degree distribution. The degree of the LT codes is the number of original symbols associated with encoding symbol. $\rho(d)$ denotes the degree distribution which is the probability that an encoding symbol has degree d.

Currently, there are four kind of degree distribution design methods, which is uniform distribution, all-at-once distribution, ideal Soliton distribution, robust Soliton distribution.

2.1 Uniform Distribution

The degree distribution is a uniform probability distribution.

$$\rho(i) = \frac{1}{n}, i = 1, 2 \cdots, n \tag{1}$$

The feature of this kind of the degree distribution is that no important information packet is used larger the degree distribution probability than other packets. Therefore, the successfully decoding probability of to those extents is affected and this degree distribution method is not commonly used in practical coding.

2.2 All-at-Once Distribution

Equation (2) shows that degree of all code words is one, which is also not commonly used in practical coding. This distribution tactics can not produce good result [1, 9].

$$\rho(d) = \begin{cases} 1, & d = 1 \\ 0, & others \end{cases} \tag{2}$$

2.3 Ideal Soliton Distribution

Based on his theoretical analysis, Luby proposed the ideal soliton distribution of which the overhead is 1, the best performance, in the ideal case. The Ideal Soliton distribution is $\rho(1) \cdots \rho(k)$, where

$$\rho(1) = \frac{1}{k} \tag{3}$$

$$\rho(d) = \frac{1}{d(d-1)}, d = 2 \cdots, k \tag{4}$$

The ideal soliton distribution guarantees that all the release probabilities are identical to $1/k$ at each subsequent step. Hence, there is one expected ripple generated at each processing step when the encoding symbol size is k. After k processing step, the source data can be ideally recovered. Figure 1, shows an example of the ideal Soliton distribution, in which blue curve denotes the degree distribution probability of the ideal Soliton distribution.

However, ideal Soliton distribution works poorly in practice. Belief propagation may be suspended by a small variance of the stochastic encoding and decoding situation in which no ripple exists, because the expected ripple size is only one at any moment. According to the theory of random walk, the probability with which a random walk of length k deviates form its mean by more than $\ln(k/\sigma)\sqrt{k}$ is at most σ .It is a baseline of the ripple queue size which must be maintained to complete a decoding process. Hence, in the same paper by Luby, a modified version called robust soliton distribution.

2.4 Robust Soliton Distribution

The robust Soliton distribution is $\mu(.)$ defined as follows. Let

$$S = c \ln\left(\frac{k}{\sigma}\right)\sqrt{k} \tag{5}$$

Define

$$\tau(d) = \begin{cases} \frac{s}{k}\frac{1}{d} & d = 1, 2 \cdots (k/s - 1) \\ \frac{s}{k}\ln\left(\frac{s}{\sigma}\right) & d = k/s \\ 0 & d > k/s \end{cases} \tag{6}$$

Add the ideal Soliton distribution $\rho(.)$ to $\tau(.)$ and normalize to obtain $\mu(.)$:

$$\beta = \sum_{d=1}^{k} \rho(d) + \tau(d) \tag{7}$$

$$\mu(d) = \frac{(\rho(d) + \tau(d))}{\beta}, \text{d=1...k} \tag{8}$$

where c and σ are two parameters for tuning robust soliton distribution. c control the mean of the degree distribution. Smaller values of c increase the probability of low degrees, and larger ones decrease it. σ estimates that there are $\ln(k/\sigma)\sqrt{k}$

expected ripples as described. Figure 1 green curve is an example of robust soliton distribution with c=0.2, σ =0.05.Robust soliton distribution can ensure that only $N = k + O(\ln^2(k/\sigma)\sqrt{k})$ encoding symbols are required to recover the source data with a successful probability at least $1-\sigma$.

Robust soliton distribution is not only viable but also practical. The analysis of robust soliton distribution based on probability and statistics is sound if k is infinite. However, in practice, source data cannot be divided into infinite pieces, and as a consequence, the behavior of LT code will not exactly match the mathematical analysis, especially when k is small. Furthermore, robust soliton distribution is a general purpose design. It provides a convenient way to construct a distribution works well but not optimally.

In Figure 1, we set k=255,c=0.2,δ=0.05, S=27, k/S=9. Red curve denotes Uniform Distribution, Purple curve denotes All-At-Once Distribution, Blue curve denotes ideal Soliton distribution, Green curve denotes robust Soliton distribution. As we can see, robust Soliton distribution has a maximum value in d=2 and d=9. This shows that there is lots of 2 and 9 degree, which is beneficial to successful decoding. Ideal Soliton distribution has only maximum value in d=2, which is not beneficial to successful decoding. Consequently, we choose robust Soliton distribution.

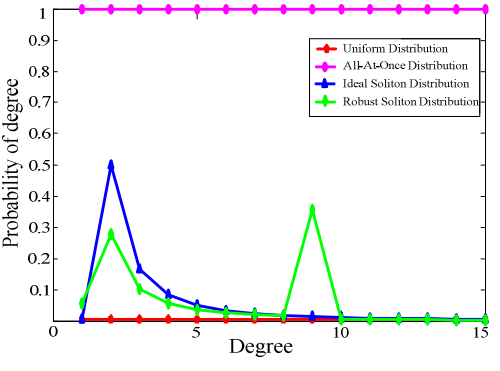


Fig. 1. Four of the probability distribution of degree distribution

3 Encoding and Decoding Procedures

In this section, utilizing the distribution defined in (8), we present the encoding and decoding procedures.

3.1 Encoding Procedure

With k input symbols $s_1, s_2, s_3, \cdots, s_k$, LT codes will produce a unlimited number of output symbols $t_1, t_2, t_3, \cdots, t_k, \cdots$.

The encoding procedure can be summarized as follows:

Step 1: Randomly choose the degree d($1 \le d \le k$) of the encoding symbol from The Robust Soliton Distribution.

Step 2: Choose uniformly at random d distinct input symbols

Step 3: The value of the encoding symbol is the exclusive-or of the d symbols selected in step 2.

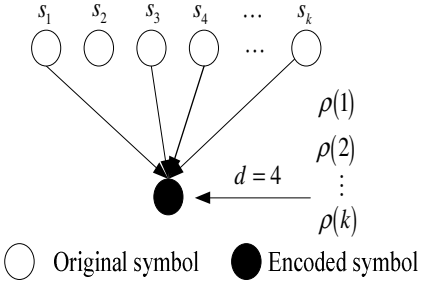


Fig. 2. The Generation process of encoding symbols

Take figure 2 as example, First of all, randomly choose the degree 4 according to function of the robust soliton distribution. Then choose uniformly at random 4 distinct input symbol form k original symbol, in this figure, we choose s_1、s_3、s_4 and s_k. Finally, the value of the encoding symbol is exclusive-or of the 4 symbols selected.

3.2 Decoding Procedure

At present, there are two algorithms used for the decoding of LT codes. The first method is Message Passing. the second method is Gaussian Elimination Decoding.

3.2.1 Message Passing Algorithm

In Binary Erasure Channel(BEC), decoding method of LT code utilizes the MP algorithm, it makes well use of the encoding nature of LT codes to decode the input symbols. In the following part, we shall explain the decoding process of LT codes.

Step 1: Find a encoded node t_n that is connected to only one source packet s_k (if there is no such encoded node, this decoding algorithm halts at this point, and fails to recover all the source packets).

a) Set $s_k = t_n$;

b) Add s_k to all encoded node $t_{n'}$ that are connected to s_k : $t_{n'} = t_{n'} + s_k$ for all n' such that $G_{n'k} = 1$;

c) Remove all the edges connected to the source packet s_k ;

Step 2: Repeat (1) until all s_k are determined.

The complexity of both encoding and decoding of LT codes is only $O(k\ln(k/\sigma))$ times of computation, which is much lower than their ascender.

3.2.2 Gaussian Elimination Algorithm

The Gaussian Elimination decoding Algorithm process as follows:

Step 1: Columns transform to generator matrix G, carry on the same transformation to the received packets of the corresponding columns, and transform G as $[E_{k*k} \mid P]$ form and corresponding $t_1, t_2, \cdots, t_n$ as $t_1^{'}, t_2^{'}, \cdots, t_k^{'}$.

Step 2: $s_1, s_2 \cdots s_k$ is equal to $t_1^{'}, t_2^{'}, \cdots, t_k^{'}$.

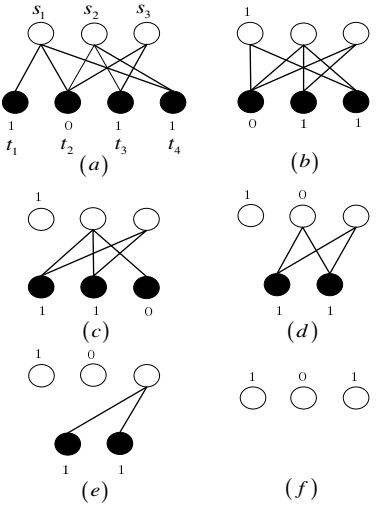


Fig. 3. The Decoding Procedure Of LT Codes

In Figure 3(a), we can see that the degree of t_1 is 1, associated with t_1 is s_1.Therefore, set $t_1 = s_1$.In Figure 3(b), we can see that after set $t_1 = s_1$, s_1 is recovered, release t_1. In Figure 3(c), remove all the edges connected to s_1 and let the degree of t_4 be 1. Set $t_4 = s_2$, then s_2 is recovered and t_4 is deleted. In Figure 3(d), add s_2 to t_2 and t_3. Remove all the edges connected to s_2. In Figure 3(e), the degree of t_2 and t_3 are 1, therefore, s_3 is recovered. In Fig 3(f) show that recover all the original packets.

4 Simulation Model and Result Analysis

4.1 Simulation Model

The simulation model of performance of LT codes is shown in Figure 4. The source produces the original information, which is encoded by using LT codes in the transmitter. The encoded data are received by the receiver after going through binary erasure channel. Decoder began to decode when it has received enough encoded data. Finally, by comparing the original data with decoding data, we get the analysis results of the performance of LT codes.

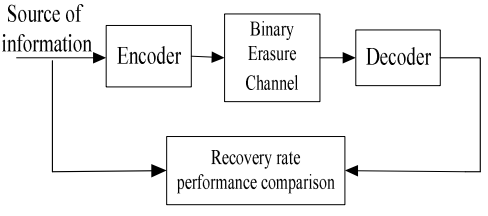


Fig. 4. LT codes system simulation model

4.2 Performance Analysis

We choose VC 6.0 and Matlab as the simulation Software and degree distribution as robust Soliton distribution. The length of the original symbol are K=255, 2000, 4000, 6000, 8000, respectively. The delete the probability of the channel is from 0.05 to 0.5 and the step length is 0.05. Probability of decoding failure δ is 0.5. The number of iterations is 200.

The main performance metric that we investigate is the successful decoding times versus redundancy. We consider a decoding operation is failed if any packet is deleted after the end of decoding operation.

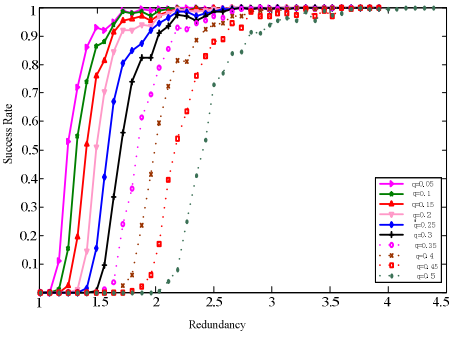


Fig. 5. Success Rate versus Redundancy k=255

The curves in Figure 5 displays the redundancy and success rate of LT codes. The first curve shows the successful recovering probability under 0.05 delete probability. From left to right the curves denote the performance of decoder when the probability in turn increases 0.05. These curves indicate that only when the steady increase redundancy the entire original packet can be recovered.

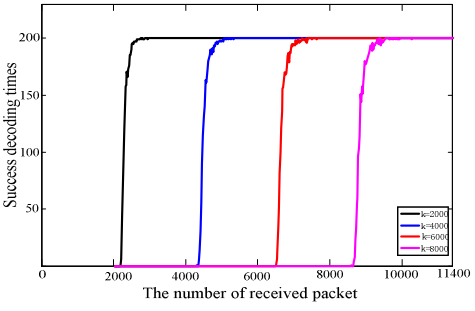


Fig. 6. The number of received packet versus success decoding times

From figure 6 we can conclude that when the original packets number is 2000, the original packets can be successfully decoded as long as the receiver receives about 2800 encoded packets. At the same time, with the increase in the number of received packets, the receiver achieves complete decoding with probability of 100%. In a word, the more the number of data packets is sent, the better the performance can be, but the delay and decoding complexity increased.

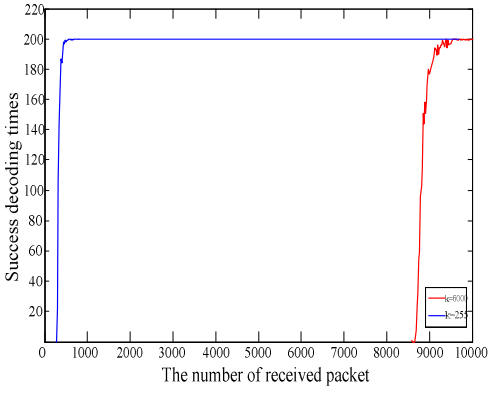


Fig. 7. Successfully decoding times versus the number of received packet under different original information packets of LT codes

In Figure 7, blue and red curve represent the length of code 255 and 6000, respectively. We can see that blue curve quickly successful decoding. Yet red curve need more decoding time, At the same time, it also increase decoding complexity.

5 Conclusion

This paper introduces the four design methods of LT codes degree distribution and simulation with MATLAB. According to the performance results, we can conclude that the robust Soliton distribution is the best method. The performance of fountain code greatly is influenced by the code length. Therefore, it usually need long code to achieve good performance. However, the long code length will increase the decoding complexity and redundancy is also greatly increased. Although the LT code performance is good, but the encoding and decoding complexity is nonlinear. In the decoding process, when the number of encoded symbol is close to the number of original symbol, LT codes can not decode based on the cost of time and space consumption. To solve these problems, Raptor code is proposed by Shokrollahi. It effectively solves contradiction between the encoding and decoding complexity of LT codes and the transmission efficiency. In the future, we will further study the performance of Raptor codes.

References

1. Mackay, D.J.C.: Fountain codes. In: IEEE Workshop on Discrete Event Systems, pp. 26–28 (1998)
2. Luby, M.: LT Codes. In: Proceedings of the ACM Symposium on Foundations of Computer Science (FOCS), vol. 1, pp. 6–7 (2002)
3. Shokrollahi, A.: Raptor Codes. IEEE Transactions on Information Theory 52(6), 2551–2555 (2006)
4. Finamore, W.A., Ramos, M.C.: Improving the Performance of LT Codes. In: 7th International Symposium on Wireless Communication Systems (ISWCS), pp. 566–570 (2010)
5. Zhu, H., Zhang, C., Lu, J.: Designing of Fountain Codes with Short Code-Length. In: Proceedings of 3rd International Workshop on Signal Design and Its Applications in Communications (IWSDA), pp. 65–68 (2007)
6. Zhang, F., Xu, L., Pan, X.: Comparison of BP and Gauss code base on Fountain Code Measuring. In: Third International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), vol. 1, pp. 737–740 (2011)
7. Chen, C.-M., Shen, T.-C., Zao, J.K.: Optimizing Degree Distributions in LT Codes by Using The Multiobjective Evolutionary Algorithm Based on Decomposition. In: IEEE Congresson on Evolutionary Computation (CEC), pp. 1–8 (2010)
8. Yao, W., Chen, L., Li, H., Xu, H.: Research on Fountain Codes in Deep Space Communication. Congress on Image and Signal Processing 2, 219–224 (2008)
9. Li, L., Zhao, J.: LT Codes with a New Degree Distribution. In: International Conference on Multimedia Information Networking and Security (MINES), pp. 531–535 (2010)

Unseen Visible Watermarking for Gray Level Images Based on Gamma Correction

Chu-Hsing Lin[1], Chen-Yu Lee[2], Shu-Yuan Lu[1], and Shih-Pei Chien[1]

[1] Department of Computer Science, Tunghai University,
Taichung 407, Taiwan
chlin@go.thu.edu.tw, g99350011@thu.edu.tw
[2] Department of Computer Science, National Chiao-Tung University
1001 Ta-Hsueh Road, HsinChu, 30050, Taiwan
chenyu@cs.nctu.edu.tw

Abstract. In this paper, we proposed a novel watermarking technology with the properties that from visible watermarking and invisible watermarking. In the common situation our approach can be used to transmit the multimedia files, according to the gamma correction, the watermark can be revealed. Our proposed method has a unique feature that we can obtain a watermarked-image similar to the original image after embedding watermarks. The naked eye hardly realized the watermarks, but after adjusting the viewing angle, it then demonstrated the embedded watermarks. We embed the unseen visible watermarks by adjusting degree of saturation and value of HSV model in the color space. The experiment result shows that the PSNR values of the watermarked images are more than 45 dB, sometimes even as high as 55 dB.

Keywords: Unseen visible watermark, Gamma Correction.

1 Introduction

In the digital era, any digital data is possible to be easily copied or modified, so the digital data protection is an extremely important area of research in the information security. Digital watermarking [7] is used to verify an image is a falsification of an effective way. The most of current commercial signs or notations are color images, of which the complexity is more than the binary image used as watermark traditionally. If a colorful signs or notations are transformed to a binary image, it would be hard to recognize by human eyes. Therefore, a gray image converted from a colorful sign or a notation could display its outline of original image by the dark degree. As a result, the gray-scale watermark has become the main watermark technology recently.

Watermark technology is divided into visible and invisible watermark [2]. Visible watermark is visible by naked eye easily, and its main purpose is to declare the copyright or ownership. The disadvantage is easy to destroy the beauty of original image and is easy removed additionally. On the other hand, the embedded invisible watermark is not easily identified by naked eyes, but the watermark needs special methods to be retrieved [3] [6]. Although this method increases the security of the watermark [4], but the identification is not as convenient as the visible watermark. In

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 236–243, 2011.

2007, Huang et al proposed an unseen visible watermarking, namely UVW [1][5], which takes the both advantages of visible- and invisible watermark. However, UVW only accept binary images as watermark to embed into cover images, it would cause the limitation of the practical applications. This paper proposed an unseen visible gray watermark, namely UVGW, improved from UVW to take gray images as watermark without loss of any characteristics of UVW. UVGW is more suitable for modern applications.

The rest of this article is organized as follows: Section 2 introduces UVW method simply and section 3 proposes our unseen visible gray watermark scheme. The results are presented in Section 4 and we have a conclusion in Section 5.

2 Introduction of UVW

Huang et al proposed an unseen visible watermark technology in 2007 [1] [5]. The hidden watermark can only be seen by naked eyes at a specific angle without any image processing, but it remains visible in other angles. This method did not affect the visual quality of embedded watermark image at normal viewing angle. Table1 lists the comparison between UVW and the invisible watermark.

UVW adjusts the intensity of original image and takes the bright difference of dark squared to achieve the effect of unseen embedded watermark. However, the method only applies to the dark part of images, and embedded blocks must be in the same hue. Otherwise UVW will not embed correctly. In addition, UVW only accept binary image as watermark to be embedded.

Table 1. Functional and analysis for UVW

Feature	1. New watermark embedding technology. 2. Using the human visual system Characteristic. 3. High practical value.
Advantage	1. Not reduce the visual quality. 2. The watermark extraction does not require additional process. 3. Both including invisible and visible watermark advantages.
Shortcoming	Less able to resist attacks.

3 Comparison of Binary and Gray Scale Images

Before introducing our scheme, the section simply introduces the differences between grayscale and binary images to explain the reasons of proposing the UVGW. Binary image takes 0 and 1 bit to present the white and black pixel and it highlights the contrast and costs less storage space. Grayscale image takes 0 to 255 to present the levels form black to white. It is easy to display the brightness of color.

3.1 Grayscale and Binary Watermark Advantage and Shortcoming

We found a problem that some binary and grayscale images remain distinguishable converted form the same color image, but some were not. Figure 1 showed that the recognition effect was almost the same, but the identification result was significantly different in Figure 2. As the result, the grayscale image has a better characteristic of identification than binary image and is more suitable to be embedded as watermark.

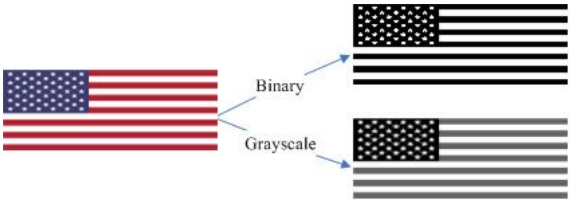


Fig. 1. Binary image with grayscale image recognition minor differences

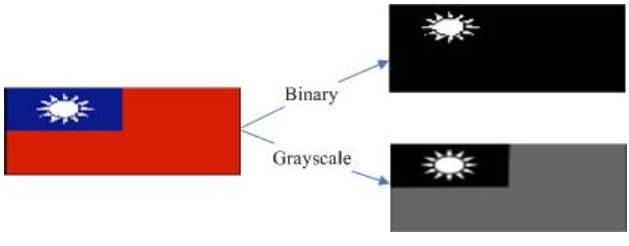


Fig. 2. Binary images with grayscale image recognition large differences

4 Unseen Visible Grayscale Watermark

Unseen visible gray watermarking, namely UVGW, improved from UVW retained all the characteristics of UVW and was more easy to recognize the embedded watermark. Tables 2 and 3 defined the symbols used in the scheme.

Table 2. Description the original image

Original image	I	Original image range	$0 \le I(x,y) \le 255$
Length	I_l	Length range	$1 \le x \le I_l$
Width	I_w	Width range	$1 \le y \le I_w$
Watermark	W	Watermark range	$0 \le W(x,y) \le 255$
Length	W_l	Length range	$1 \le x \le W_l$
Width	W_w	Width range	$1 \le y \le W_w$

4.1 Grayscale Watermark Embedding Method

First, we performed gamma correction on the original image I to obtain $G(i)$ and find the largest gradient value i^* of $G(i)$.

$$G(i), 0 \le i \le 255, 0 \le G(i) \le 255 \tag{1}$$

$$i^* = \arg\max_i \nabla G(i) \tag{2}$$

The noise of the original image I would be the amplified after gamma correction. To ensure not to reduce image quality seriously, the de-noise operation was performed to get I'_j in Eq. 3 where δ_j value was the parameter of de-noising under the maximum value of δ_{max}.

$$I'_j = D(I_G, \delta_j), \delta_j \le \delta_{\max} \setminus \tag{3}$$

We chose a block starting at the position P with the same size of grayscale watermark and the maximum difference between largest gradient value i^*. The best embedding region R starting at position P was defined in Eq. (5). The de-noising was performed again until the Eq. (6) satisfied of which T was a predefined constant value.

$$P = \arg\max \sum_{x=x_0}^{x_0+W_l-1} \sum_{y=y_0}^{y_0+W_w-1} \left| I'_j(x, y) - i^* \right| \tag{4}$$

$$\begin{aligned} R = \{ I'_j(x, y) \mid & x_0 \le x \le x_0 + W_l - 1, \\ & y_0 \le y \le y_0 + W_w - 1 \} \end{aligned} \tag{5}$$

$$S = \sum_{x=x_0}^{x_0+W_l-1} \sum_{y=y_0}^{y_0+W_w-1} D\left(I'_j(x, y) - i^* \right) \le W_l \times W_w \times T \tag{6}$$

$$D(k) = \begin{cases} 1, if\ k = 0 \\ 0, otherwise \end{cases}$$

In RGB color mode, saturation is the proportion of R, G, B values. One of them was chosen as the color index, which had the maximum value among them in the embedded region. If a pixel had the same R, G, B value, the index was determined as the color appeared most around the pixel.

$$\begin{aligned} MAX &= \max(R, G, B) \\ MIN &= \min(R, G, B) \end{aligned} \tag{7}$$

Perform the analysis on grayscale watermark, which is used to identify the distribution of gray watermark as Eq.(8).

$$L=\left(avg\sum_{x=x_0}^{W_l-1}\sum_{y=y_0}^{W_w-1} W(x,y) \right)\times\sum_{x=x_0}^{W_l-1}\sum_{y=y_0}^{W_w-1} W(x,y) \tag{8}$$

Intensity adjustment of each pixel is according to its index in the watermark embedded in the region. The dynamic adjustment prevent the adjusted color from being too great to over than 255. In the grayscale watermark, the pixel intensity value must be adjusted in accordance with gray level, so the adjustment valuable ω was determined by *L*. While the above completed, the grayscale watermark was embedded in the original image successfully.

$$\tilde{I}(x,y)=\begin{cases} I_j'(x,y)-\omega, & if\ (I_j'(x,y)+\omega)>255 \\ I_j'(x,y)+\omega, & otherwise \end{cases} \tag{9}$$

$$\omega=W(x-x_0,\ y-y_0)\pm L$$

4.2 Watermark Extraction

There are three ways to extract or display the watermark from the embedded image: gamma correction, image processing tools, and human vision.

The first way was Gamma Correction processing which could increase the intensity of embedding the image and made the watermark appear obviously. The second way was to use some image processing softwares, such as Photoshop, to adjust the saturation to display the embedded watermark. The last way, also is the easiest way, is to see the embedded image at a specific angle by naked eyes.

4.3 Experimental Results

The section explains our experimental results. Figure 3(a) was a 400 × 400 pixels original image and Figure 3(b) was a 160 × 60 pixels grayscale image as the watermark in our experiments. The embedded image was shown in Figure 4(a) in which the watermark is embedded at the region shown in Figure 4(b). Figure 4(c) and 4(d) were the embedded image after gamma correct and proceed by image processing software respectively. Figure 4(e) and Figure 4(f) were the results viewing computer screen from different angles. The embedded watermark would be seen only at specific angle as in Figure 4(f). The experiment result showed that the PSNR values of the watermarked images were more than 45 dB, sometimes even as high as 55 dB.

The original watermark is a grayscale image, but the embedded watermark image became a color image shown in Figure 5. The pixels of embedded watermark were changed in accordance with the background color to increase the effect of hiding.

(a)

(b)

Fig. 3. (a) Original image (400 × 400) pixels; (b) Grayscale watermark (160 × 60) pixels. From: www.taipei-101.com.tw/index_en.htm.

Fig. 4. Experimental results

Fig. 5. Grayscale watermark embedding Figure

4.4 Comparison of UVGW and UVW

Our experiments performed six different attacks on UVGW and UVW, including rotation attack, print-scan attack, contrast attacks, sharpening attacks, emboss attacks and spherical attacks. The test results were summarized in Table 3:

Table 3. Following attack against the stability of the comparison table UVGW and UVW

Attacks	UVGW	UVW
Rotation	●	●
Print -Scan	●	●
Contrast	●	○
Sharpen	●	○
Emboss	●	○
Spherical	●	●

※ ● that is able to withstand attack, ○ that is able to not withstand attack

5 Conclusion

In this paper, we proposed a unseen visible grayscale watermarking for grayscale watermark to hide a gray watermark and it can be seen by human eye easily. The results showed that the embedded grayscale watermark in the normal viewing angle is not easy to be detected. Therefore, the embedded watermark could be easily displayed by adjusting the angle of the screen. Our UVGW resisted more attacks then the original UVW and it is more practical in many applications.

Acknowledgments. This work was supported in part by the National Science Council under the grant NSC99-2221-E029-034-MY3.

References

1. Huang, C.H., Chuang, S.C., Huang, Y.L., Wu, J.L.: Unseen Visible Watermarking: A Novel Methodology for Auxiliary Information Delivery via Visual Contents. IEEE Trans on Information Forensics and Security 4(2), 193–206 (2009)
2. Das, S., Bandyopadhyay, P., Paul, S., Ray, A.S., Banerjee, M.: A New Introduction Towards Invisible Image Watermarking on Color Image. In: IEEE International Advance Computing Conference (IACC 2009), pp. 1224–1229 (2009)
3. Tsai, H.M., Chang, L.W.: A High Secure Reversible Visible Watermarking Scheme. In: IEEE International Conference on Multimedia and Expo, pp. 2106–2109 (2007)

4. Khattak, F.U., Bais, A., Khawaja, Y.M.: An Invisible Dual Watermarking Scheme for Authentication and Copyrights Protection. In: International Conference on Emerging Technologies (ICET 2009), pp. 247–251 (2009)
5. Chuang, S.C., Huang, C.H., Wu, J.L.: Unseen Visible Watermarking. In: IEEE International Conference on Image Processing (ICIP 2007), vol. 3, pp. III-261–III-264 (2007)
6. Mohanty, S.P., Guturu, P., Kougianos, E., Pati, N.: A Novel Invisible Color Image Watermarking Scheme Using Image Adaptive Watermark Creation and Robust Insertion-Extraction. In: Eighth IEEE International Symposium on Multimedia (ISM 2006), pp. 153–160 (2006)
7. Zhang, Y.: Digital Watermarking Technology: A Review. In:International Conference on Future Computer and Communicatio, FCC 2009, pp. 250–252 (2009)

People Counting Using Object Detection and Grid Size Estimation*

Oliver C. Agustin and Byung-Joo Oh**

Department of Electronics Engineering, Hannam University
133 Ojeong-dong, Daedeok-Gu, Daejeon, Korea
oliver.agustin@gmail.com, bjoh@hnu.kr

Abstract. The proposed system estimates the number of people in indoor or outdoor location. A background model of the scene is constructed to detect moving objects (people) in a video stream. Detected foreground objects are preprocessed to eliminate pixel noise and small artifacts by performing opening morphology operation in the foreground image. Counting persons in the occluded foreground object is estimated by counting the human blobs by estimating the grid size. The accurately normalized and calibrated reference grid model is necessary for the grid estimation.

Keywords: People counting, foreground segmentation, grid size estimation.

1 Introduction

Counting people is an important task in automatic surveillance systems. It gives information as to how many people entered a room at specified span of time or determine the number of people in an area at any time. The problem in people counting is about estimation of the density of people in crowded environment such as train platform or busy corridors. To solve this problem, various types of sensors that are strategically positioned can be used. Some types of these sensors are electro-optical, thermic, and passive-optics directional sensors. Among these sensors, video cameras are getting popular and becoming an increasingly commercial off-the-shelf solution for surveillance purposes. Moreover, video cameras used for surveillance can be exploited to perform more important tasks like tracking and crowd estimations. This leads us to the problem of estimating people count in videos.

An obvious solution to the problem of estimating crowd size in a video frame would be to perform head count. While this would be a tiresome but feasible procedure for a human to perform manually, it would certainly be very difficult problem in vision-based systems to accomplish. In [1], head search and model matching technique is used to solve the people counting problem. However, the approach uses color digital camera positioned vertically above the ceiling viewing downwards to the heads similar to the camera configuration in [2]. In this way, occlusion is not much of a problem. Similar idea is exploited in [3] where face

* This work was supported by the 2011 Hannam University grant.
** Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 244–253, 2011.

detection scheme determine the person count. The downside of this approach is that people can only be counted in one direction where faces are exposed to the camera. RAM-based neural network is proposed in [4] to create a score image, people counting are then made possible by serial search for peaks in the bright contiguous regions in the image. In the approach suggested in [5], instead of determining the exact number of people, estimation of the crowd density is obtained using a single-layer neural network fed with features (edge count, densities of the background, and crowd objects). The crowd density estimation of the neural network is then refined by the hybrid global learning algorithm.

In this paper, we present a method for automatic people counting in conventional closed-circuit television (CCTV) using computer vision technique. In section 2, we review various background segmentation techniques. The proposed people counting system is presented in section 3. We also provided a grid size estimation technique. Section 4 presents the result and discussion about what we have done and our on-going research. In section 5 we give the conclusion and discuss some difficulties and constraints in this work.

2 Background and Foreground Modeling

The problems inherent to people counting system both needed to detect the presence of motions in order to activate the mechanism that tracks and processes received video signals. However, the capability to extract moving objects from video sequence is crucial and a fundamental problem in vision-based people counting and surveillance systems. Background subtraction is the most widely used method to segment moving regions in the image sequences. This is achieved by comparing new frames to a model of the scene background[6]. However, background modeling that employs statistical approach has been found to be more successful in diverse conditions.

Because background modeling is one of the critical components to accurately estimate the number of people present in the image, different methods were explored by researchers in the field. Some authors use a probability density function to characterize the background model by means of nonparametric kernel density estimation technique [6,7]. Optical flow as features in high dimensional space was exploited in [8] to model the background via data-dependent bandwidth in kernel density estimation. However, other techniques similar to [9] do not depend on background model to count people. Instead, feature set trajectories are tracked and preprocessed through spatio-temporal conditioning. The number of people is determined by clustering these trajectories.

In [10], a method has been proposed that employs the color co-occurrence to describe the moving background. Initial results showed that features are more effective to model the dynamic parts of the background than those employed by previous methods. However, this technique is still incapable of recovering from sudden changes in the video. In [11], foreground objects were extracted from complex video were developed using the Bayes decision framework. In this framework, Bayes decision rule are employed for classifying the general feature vector into foreground

and background. Statistics of different feature vectors is maintained and learned so it can be used for modeling the complex background which comprised of stationary and motion objects. Statistics of most significant colors described the stationary parts of the background while the most significant color co-occurrences are used to describe the motion objects of the background. Foreground objects are then extracted by fusing the detection results from both the stationary and motion pixels. The details of this section are given in [11].

2.1 Bayesian Approach for Modeling of Background and Foreground

In the framework in [11], Bayes decision theory are used to classify image pixels in each video frames to identify foreground and background objects even with the presence of non-stationary background objects and suddenly changing lighting condition. Their work is able to handle gradual and sudden changes in background. Stationary background objects are characterized by color features, while moving background object is represented by color co-occurrence features. The foreground objects can be extracted by combining the result of classifying the stationary and moving pixels. There are two features used to model the complex background:

1. statistics of most significant colors (stationary parts of the background)
2. most significant color co-occurrences (motion objects of the background)

2.1.1 Formulation of Classification Rule

Let v_t be a discrete value feature vector from the image sequence at pixel $s = (x, y)$ and instant of time *t*. By using Bayes rule, the a posterior probability of v_t from the background *b* or foreground f is [11]

$$P(C \mid v_t, s) = \frac{P(v_t \mid C, s)\, P(C \mid s)}{P(v_t \mid s)}, \quad C = b \text{ or } f. \tag{1}$$

By the Bayes decision rule, pixel is classified as background if the feature vector satisfies

$$P(b \mid v_t, s) > P(f \mid v_t, s) \cdot \tag{2}$$

Since feature vector associates the pixel *s* as pixels in foreground or background objects,

$$P(v_t \mid s) = P(v_t \mid b, s) \cdot P(b \mid s) + P(v_t \mid f, s) \cdot P(f \mid s) \cdot \tag{3}$$

Substitution of (1) and (3) to (2) leads to

$$2\, P(v_t \mid b, s) \cdot P(b \mid s) > P(v_t \mid s) \tag{4}$$

Equation (4) indicates that by learning the *a prior* probability $P(b \mid s)$, the probability $P(v_t \mid s)$ and the conditional probability $P(\mathbf{v}_t \mid b, s)$ in advance, we may classify $\mathbf{v}_t$ as either associated with foreground or background.

2.1.2 Representation of Feature Statistics

Since the mathematical form of $P(v_t \mid s)$ and $P(\mathbf{v}_t \mid b, s)$ in (4) are unknown in general cases. They could be represented by the histograms of feature vectors over the entire feature space. If the selected features are effective to represent background, at a pixel *s*, the feature vectors from the background would concentrate in a very small subspace of the feature histogram, while the feature vectors from foreground objects would distribute widely in the feature space.

For each type of feature vectors, statistics of $P(v_t^i|s)$, $P(v_t^i|b,s)$, and v_t^i for $i = 1, \ldots, N_2$ $(N_2 > N_1)$ is maintained at pixel *s* and time *t* to record the statistics for the N_2 most significant values for each type of feature vectors. To adapt to the variations of busyness in the scene over different time duration, we maintain $p_b^{s,t} = P(b \mid s)$ at each time *t*.

2.2 Foreground Object Detection

The procedure for foreground object detection from a real-time video containing complex background includes change detection, change classification, foreground object segmentation, and background maintenance, and described in the following subsections[11].

2.2.1 Change Detection and Classification

The background and temporal difference are performed as follows:

First, a simple picture differencing is performed for each color component using adaptive thresholding. The results from the three components are combined to generate the background difference and temporal difference respectively. The image differencing is used to remove the imaging noise. The remainder changes will be classified based on the background features.

The temporal differences classify the change pixels into two types. If temporal difference $F_{td}(s,t) = 1$ is detected, pixel is classified as motion pixel that belongs to a moving object. Otherwise, this pixel belongs to a stationary pixel which is associated with stationary object. Morphological operation *open* and *close* is performed to eliminate the scattered pixel errors and connect the foreground pixels. The segmented foreground objects form a binary output image.

2.2.2 Background Maintenance

In [11], the background maintenance includes two parts, updating the tables of feature statistics and a reference background image. Two tables of color and color co-occurrence statistics are maintained at each pixel. Two updating strategies are available to adapt them to both gradual and abrupt background changes.

Updating to gradual background changes. If the feature vector $\mathbf{v}_t$ is used to classify the pixel *s* as foreground or background at time *t*, the statistics of the color distribution and color co-occurrence is gradually updated recursively using the learning rate and Boolean values.

Updating to abrupt background changes. When a sudden background change has occurred, the features of the new background appearance become dominated immediately after the change. From (3) and the relations between the background and foreground, new background features at *s* is detected if

$$P(f \mid s)\sum_{i=1}^{N_1} P(\mathbf{v}_t^i \mid f, s) > T \tag{5}$$

where *T* is a percentage value which decides when the new features can be identified as new background appearance. The term $P(f \mid s)$ prevents updating from small number of features. With this operation during "abrupt" background changes, the observed domination features are converted as the learned background features.

Updating the reference background Image. The reference background image[11] is updated gradually for stationary objects by using Infinite Impulse Response (IIR) filter as:

$$B_c(s,t+1) = (1-\alpha)B_c(s,t) + \alpha I_c(s,t) \tag{6}$$

for *c=r, g, b* and α is the parameter of the IIR filter. A small positive number of α is chosen to smooth out any perturbation caused by image noise.

If a background change has been detected. The pixel is updated with the new background appearance

$$B_c(s,t+1) = I_c(s,t), \text{ for } c = r, g, b. \tag{7}$$

This operation ensures that the reference background image can track the background motion.

3 People Counting Algorithm

This section describes our proposed people counting system. First, we present the components that make up the people counting system and describe each component individually. We also present an idea for dealing with occlusion and improved estimation of crowd density.

3.1 Characterizing Foreground Objects

The goal of foreground classification is to distinguish objects of interest from the background scene and illumination artifacts such as shadows and highlights, and to exclude these background objects from further processing. This is usually accomplished by evaluating the color distribution within a foreground region as in [7]. A *shadow* region has similar hue and saturation to the background but lower intensity. *Highlight* region has similar hue and saturation to the background but a higher intensity. Consequently, an object region has a different hue and saturation than the background. In people counting system, we are particularly interested in identifying humans from the classified foreground regions.

Since, detected foreground regions may still include objects other than humans, e.g. trolleys, push carts, bag; we propose a human detector that takes care of determining the presence of persons in the video frame. The proposed system should process each found region separately and take care of occlusion problems when found in the image. This approach prevents unnecessary processing of detected non-occluding persons that could drastically reduce reliability of the system.

In Fig. 1, foreground regions are contiguous objects that were classified as foreground pixels described in section 2.1. Human detector is a classifier built for detecting the presence of persons in foreground regions. If a person cannot be detected in foreground region, it is discarded.

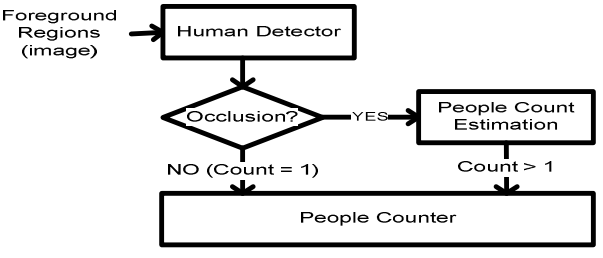


Fig. 1. Block diagram of people counting system

In each video frame, more than one region could have been present and the operation in Fig. 1 is repeated for each regions found in each frame. The presence of at least one person in foreground object under processing suggests human occlusion which activates the next stage for estimating the number of people in occluded foreground region. The people counter sums and stores all detected singular humans and people count estimation. The people count estimation stage is a regression model that correlates the number of occluded people in a foreground region based on the number of pixels and shape of the regions. This process is performed for each frame in a surveillance video. Additional shape and color features can be added to further increase the correct estimation ability of the model.

In the next section, we describe the strategy for people count estimation when human occlusion is detected in the image.

3.2 People Count Estimation

Our proposed scheme is similar to [12] but instead a support vector machine regression handles the estimation of the people count at any video frame. In our approach, we correlate the features of the foreground object to the number of people found in the occlusion. The block diagram of the algorithm is presented in Fig. 2.

When this operation was activated, a foreground region with human occlusion is processed as follows: Subdivide the image into smaller $n \times n$ grids, extract relevant features, finally, use these extracted features as input to grid regression to get the estimated number of people. Note that in this approach, processing is confined only

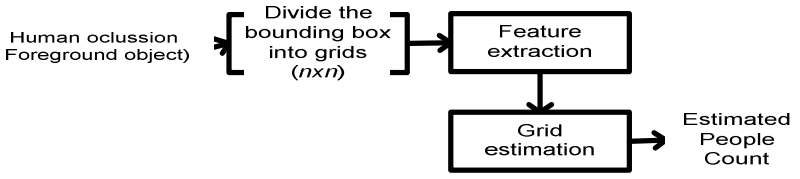


Fig. 2. People count estimation using grid estimation

to each foreground regions in which occlusion of people are present as opposed to the approach in [13].

Grid Size Selection. Grid size selection is critical to the accuracy of estimation result of the grid estimation in Figure 2. Depending on the position and angle of the camera grid size can be made constant so that size of each cell is the same for the whole image. Another approach is to make the cell size proportional to the horizontal distance of the foreground region from the camera, this way, large foreground objects will not appear larger than foreground objects far away from the camera. In addition, we can choose to normalize all foreground objects to the same size so that fixed cell can be used in straightforward manner.

Feature Extraction. The ratio(r*)* of foreground pixels in each cell of the grid is obtained by counting foreground pixels and then dividing it to the area of the cell. Regardless of how the image grid is selected (fixed or variable size), the number of cells will be fixed for the whole duration of the video streams. Thus, there will be a total of *N* features set whose values are real numbers in the range $\{0,\ldots,1\}$ which corresponds to the proportion of foreground pixels in each cell.

Grid Size Estimation. The number of the people is countered depending on the position and the size of the objects occupied. The size of the occupied object for the same persons will be varied depending on the distance from the camera position. If the persons are far away from the camera, the size is small, and if the distance is near the size will be large. Here we propose a reference model representing the relations between the number of people and the size of the object.

Assume *c* is the distance from the camera to the center of the block, *x* and *y* are the vertical and horizontal size of the block, respectively. For example, the height *z* and the width y of the of a person standing at distance *x* is

$$z = x\ tan\ \theta_1 \text{ and } y = x\ tan\ \theta_2 \tag{8}$$

where θ_1 is the vertical angle representing the ratio of distance *c* with height of the camera, and θ_2 is the horizontal angle representing the ratio of distance *c* with reference vertical size. The number of persons *n* in the occluded block is estimated *n = f(c, x, y).* The function *f* should be adjusted depending on the position of the camera.

Figure 3(left) shows a sample image for the original scene obtained from CCTV. After noise filtering, foreground segmentation, we obtain foreground feature blocks. Finally we obtain the feature blocks with reference grid estimation. The reference grid is given as in the following Fig. 3(right).

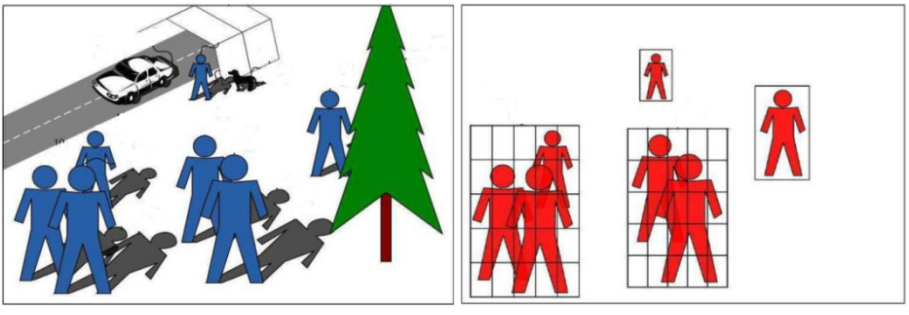


Fig. 3. (left) A sample image for the original scene obtained from video camera. (right) Reference grid model representing the people and block occupied.

4 Test Results

Experiments have been performed on video image of subway station. The test video include light illuminations, shadows of people on the ground surface, people carrying baggage, people walking together closely. Fig. 4 shows six persons are moving separately, and two persons are closely together. The foreground objects were detected and segmented and the results are given in Fig. 4(right) and in Fig. 5(right).

In Fig. 4(right), four persons are counted easily since the small blocks are separated with reasonable distance. But the block located in the center below occupied a large size of grid. So it can be counted as a person carrying a baggage or two persons walking closely. In our test this was counted as two persons.

The image in Fig. 5 includes four persons, The person standing away is not detected because he does not move around and the color is similar to the background image at the time the image is captured. The person carrying a baggage is counted as a person even the small object is detected since the position and the size of the object is too small to be counted as a person. The large block in the right below is counted as two persons.

Here the test have been done on the relatively simple images. We have faced many difficulties for the busy complex images including the occluded crowed people. Because of the diverse variety and complexity of the video image, it is not easy to say the performance of the proposed algorithm yet. To raise the accuracy of counting the occluded persons, reference grid model must be normalized and calibrated accurately depending on the position and angle of the CCTV.

Fig. 4. (left) Six persons are moving, and two persons are close together.
(right) Five blocks in addition to one large block.

Fig. 5. (left) Four persons, and one person has a drawing a baggage.
(right) One big block and small two blocks.

5 Conclusions

The presented system estimates the number of people in indoor location. The image obtained from CCTV is preprocessed for noise filtering, segmentation, and foreground feature blocks. The count of persons in the occluded image is estimated by counting the human blobs by grid size estimation. Here the test have been done on the relatively simple images. We have faced many difficulties for the busy complex images including the occluded crowed people. To raise the accuracy of counting the occluded persons, reference grid model must be normalized and calibrated accurately depending on the position and angle of the CCTV.

In future, there are lots of challenges to be addressed. One is the development of background model that would prove effective in people counting domain robust enough for use in any indoor situation. Second is the development of *support vector regression (SVR)* that will estimate the count of people when an occlusion is present.

References

1. Pang, K.-H., Ng, C.-K.: Automated People Counting Using Template Matching and Head Search. In: The 9th IEEE Conf. on Mechatronics and Machine Vision in Practice, Chiang Mai, Thailand, pp. 323–331 (2002)
2. Wang, J.M., Chen, S.W.: People Counting Based on Top-View Video Sequence. In: 18th IPPR Conf. on Computer Vision, Graphics and Image Processing (CVGIP 2005), Taipei, ROC (August 2005)
3. Sweeney, L., Gross, R.: Mining images in publicly-available cameras for homeland security. In: Spring Symposium on AI Technologies for Homeland Security, Palo Alto (2005)
4. Schofield, A.J., Mehta, P.A., Stonham, T.J.: A system for counting people in video images using neural networks to identify the background scene. Pattern Recognition 29, 1421–1428 (1996)
5. Cho, S.Y., Chow, T.W.S., Leung, C.T.: A neural-based crowd estimation by hybrid global learning algorithm. IEEE Trans. on Systems, Man and Cybernetics, Part B 29, 535–541 (1999)
6. Elgammal, A., Harwood, D., Davis, L.: Non-parametric Model for Background Subtraction. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 751–767. Springer, Heidelberg (2000)
7. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proc. of the IEEE 90, 1151–1163 (2002)
8. Mittal, A., Paragios, N.: Motion-Based Background Subtraction Using Adaptive Kernel Density Estimation. In: IEEE Conf. on CVPR, vol. 2, pp. 302–309 (2004)
9. Rabaud, V., Belongie, S.: Counting Crowded Moving Objects. In: IEEE Computer Society Conf. on CVPR, vol. 1, pp. 705–711 (2006)
10. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Foreground object detection in changing background based on color cooccurrence statistics. In: Proc. of the Sixth IEEE Workshop on Applications of Computer Vision, p. 269 (2002)
11. Li, L., Huang, W., Gu, I.Y.H., Tian, Q.: Foreground object detection from videos containing complex background. In: Proc. of the Eleventh ACM Intern. Conf. on Multimedia, pp. 2–10 (2003)
12. Kilambi, P., Ribnick, E., Joshi, A.J., Masoud, O., Papanikolopoulos, N.: Estimating pedestrian counts in groups. Computer Vision and Image Understanding 110, 43–59 (2008)
13. Roqueiro, D., Petrushin, V.A.: Counting people using video cameras. Inter. Journal of Parallel, Emergent and Distributed Systems 22, 193–209 (2007)

The Research of Serially Concatenated FQPSK Demodulation Based on LDPC Codes

Gao Yuan Zhang, Hong Wen, Liang Zhou, Ling Yang, and Yi Fan Wang

National Key Lab of Communication of UESTC, Chengdu 611731, China
zhanggaoyuan407@163.com, sunlike@uestc.edu.cn

Abstract. As a kind of modulation with high degree of both power and bandwidth efficiency, Filtered QPSK (FQPSK) has been widely applied in the satellite and deep space communication. This paper first investigates the Maximum-A-Posteriori-Probability (MAP) demodulation of FQPSK-B. Then a serially concatenated system of FQPSK and LDPC codes is built, which gives both the advantages of LDPC codes and FQPSK and can easily be considered to save the power and spectral utilization ratio of the system. We demonstrate that the new concatenated system can achieve high power and spectral efficiency through extensive analysis and simulation.

Keywords: Filtered QPSK, LDPC Codes, MAP, weighted process.

1 Introduction

To be a kind of modulation with high degree of both power and bandwidth efficiency, FQPSK modulation has been widely applied in the deep space communication [1-4], Filtered and hard-limited variants of FQPSK referred as FQPSK-B have been implemented [5-6] to achieve additional spectral efficiency and true constant envelope.

The optimum receiver for FQPSK is the Viterbi receiver which uses the 16-state trellis diagram. There are also some other suboptimum detection schemes [4] with lower complexity. The MAP algorithm [7], also based on the representation of FQPSK as a trellis-coded modulation, is not investigated previously.

With their good performance which is near to Shannon limit, LDPC codes brought out by Gallager [8], and rediscovered by Mackay [9], is widely used in the satellite communication and deep space communication. With the soft information abstracted by the MAP demodulation, the Sum-Product Algorithm (SPA) can be easily used.

In this paper, taken the view of both the advantages of LDPC code and FQPSK, we can easily consider the combination of them to save the power and spectral utilization ratio of the system and to enhance the system capacity. With the soft information which is prone to abstracted by the MAP demodulation of FQPSK, the Sum-Product Algorithm (SPA) can be easily used to decode the LDPC codes, and then the serially concatenated system of FQPSK and LDPC code is realized.

This paper is organized as follows. In section 2, the deduction of the MAP demodulation of FQPSK-B is presented. We explain the serially concatenated system

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 254–260, 2011.

used in section 3, and the weighted process is also detailedly described in this section. In section 4, the simulation results are given. Finally, the conclusion is given in section 5.

2 The Map Demodulation of FQPSK

2.1 The Simplified MAP Receiver Structure

The simplified MAP receiver is similar to the simplified Viterbi type receiver in [5]. The only difference is the detection algorithm as shown in Figure 1. So, the preliminary introduction is omitted, which can be found in [5].

2.2 The MAP Demodulation Process

The received signal is

$$r_i(t) = q_m(t) + n(t), \quad (m = 0,1,2,3) \tag{1}$$

where $n(t)$ is an AWGN noise with zero mean and $N_0/2$ W/Hz two-sided power spectral density. Let s_t represents the state of the trellis at time t, x_t is the transmitted bit at time t, the corresponding received vector is $\mathbf{r}_t = (r_{1t}, r_{2t}, r_{3t}, r_{4t})$, where $r_{it}\ (i = 1, \cdots 4)$ is the output of the four correlators, in the time interval $[-T_s/2, T_s/2]$. $\mathbf{r}_t^{t'}$ is the received sequence from time t to t', $\mathbf{r}_1^L$ is the whole receive sequence.

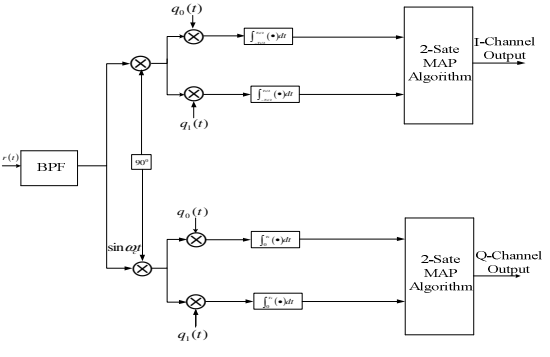


Fig. 1. Simplified MAP receiver

The conditional probability is

$$p(\mathbf{r}_t / q_m(t)) = \prod_{i=1}^{4} p(r_{it} / q_m(t)) \quad (m = 0,1,2,3). \tag{2}$$

The component of the received vector $\mathbf{r}_t$ satisfies

$$r_{it} = q_{mit} + N_{it} \left(i = 1, 2, ..., 4\right) \tag{3}$$

where q_{mit} is the projection of the transmitted signal $q_m(t)$ on $q_i\left(t\right)$, N_{it} is the noise at the output of the correlators. Now we define four constants:

$$E_0 = \int_{-\frac{T_S}{2}}^{\frac{T_s}{2}} q_0{}^2(t)d(t),\ E_1 = \int_{-\frac{T_S}{2}}^{\frac{T_s}{2}} q_1{}^2(t)d(t), \tag{4a}$$

$$E_2 = \int_{-\frac{T_S}{2}}^{\frac{T_s}{2}} q_2{}^2(t)d(t),\ E_3 = \int_{-\frac{T_S}{2}}^{\frac{T_s}{2}} q_3{}^2(t)d(t). \tag{4b}$$

So (2) is given by

$$p\left(\mathbf{r_t}/q_{\mathbf{m}}(t)\right) = \frac{1}{\left(2\pi\delta^2 E_i\right)^2} \exp\left[-\sum_{k=1}^{4} \frac{\left(r_{it} - q_{mit}\right)^2}{\delta^2 E_i}\right] \tag{5}$$

where δ^2 is the variance of the noise at the input of the correlators.

We suppose that the baseband signal $q_0(t)$ is sent. Now we define some functions as follow:

$$\begin{cases} \alpha_t(m) = pr\left\{s_t = m; \mathbf{r}_1^t\right\} \\ \beta_t(m) = pr\left\{\mathbf{r}_{t+1}^L \middle| s_t = m\right\} \\ \gamma_t(m', m) = pr\left\{s_t = m; \mathbf{r}_t \middle| s_{t-1} = m'\right\} \end{cases}, \tag{6}$$

$$\begin{cases} \lambda_t(0) = \sum_{(s',s)\in\sum_l^+} p(s_l = m', s_{l+1} = m, \mathbf{r_t}) \\ \lambda_t(1) = \sum_{(s',s)\in\sum_l^-} p(s_l = m', s_{l+1} = m, \mathbf{r_t}) \end{cases}, \tag{7}$$

$$p(s_l = \mathrm{m}', s_{l+1} = m, \mathbf{r_t}) = \alpha_t(m) \cdot \beta_t(m) \cdot \gamma_t(m', m). \tag{8}$$

The simplified MAP demodulation process consists of the following steps:

1) The initialization is

$$\alpha_0(0) = 1, \alpha_0(m) = 0 (m \neq 0), \beta_L(m) = \frac{1}{2}. \tag{9}$$

2) The computation of $\gamma_t(m',m)$:

$$\gamma_t(m',m)=p\{m/m'\}*p\{q_i(t)/m,m'\}*p\{\mathbf{r}_t/q_i(t)\}. \tag{10}$$

The bits sent at the source is usually of equiprobability, and omitting the portions which make no difference, (10) can be further simplified as

$$\begin{cases} \gamma_t'(0,0)=e^{\frac{2r_{1,t}-E_0}{\delta^2}} & \gamma_t'(1,0)=e^{-\frac{2r_{2,t}+E_1}{\delta^2}} \\ \gamma_t'(0,1)=e^{\frac{2r_{2,t}-E_1}{\delta^2}} & \gamma_t'(1,1)=e^{-\frac{2r_{1,t}+E_0}{\delta^2}} \end{cases}. \tag{11}$$

3) The computation of forward variable:

$$a_t(m)=\sum_{m'} a_{t-1}(m')\gamma_t'(m',m). \tag{12}$$

4) The computation of backward variable:

$$\beta_t(m)=\sum_{m'} \beta_{t+1}(m')\gamma_{t+1}'(m,m'). \tag{13}$$

The bit-level detection scheme at time t is

$$s_t=\begin{cases} 1, & \lambda_t(1)\geq\lambda_t(0); \\ 0, & \lambda_t(0)\geq\lambda_t(1). \end{cases} \tag{14}$$

3 The Serially Concatention

With the soft information abstracted by the MAP demodulation, the Sum-Product Algorithm (SPA) can be easily used, and then a serially concatenated system of FQPSK and LDPC code is realized. In addition, the phenomenon of positive feedback is existence in our serially concatenated system. One of the methods to abate the infection of the positive feedback is named the weighted process[10]. With the weighted process, the volatility of the intrinsic information is reduced and the speed of the system convergence is quickened. The block diagram of the modified system is shown in Figure 2.

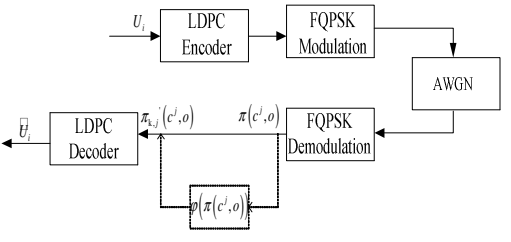


Fig. 2. The serially concatenated system of FQPSK and LDPC code with weighted process

4 Simulation Perforamance

The simulations described in this section were developed in Matlab using the bit error probabilities

4.1 The Simplified FQPSK-B MAP Receivers Performance

The length of the bit sequence in MAP receiver is set as 1000 in our simulation. Figure 3 illustrates the simulation result and compares it with some other suboptimum detection schemes.

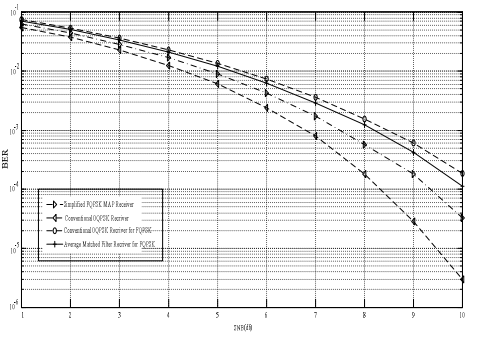


Fig. 3. Bit-error probability performance of FQPSK-B conventional OQPSK, average matched filter and MAP receiver

It is convinced that, when compared with the full MAP receiver, the simplified MAP receiver suffers a slight degradation but is still better than the conventional OQPSK receiver and the average matched filter receiver.

4.2 Detailed Introduction of the Weighed Process

The LDPC used in our simulation is (n,k)=(2000,1000). And the bitwise probability after the weighted process is given as

$$\pi_{\mathrm{k},j}{}'\left(c^j,o\right)=\varphi\left(\pi\left(c^j,o\right)\right)\bullet\pi\left(c^j,o\right). \tag{15}$$

In [10], the exponential weighted process is advanced. The kernel of this weighted process is lower the intrinsic information to a certain extent. So the linear weighted process, in which the weighted function is defined as

$$\varphi\left(\pi_{k,j}\left(c^j,o\right)\right)=a\cdot\pi_{k,j}\left(c^j,o\right) \tag{16}$$

where $a\in\left(0,1\right)$,is obtained.

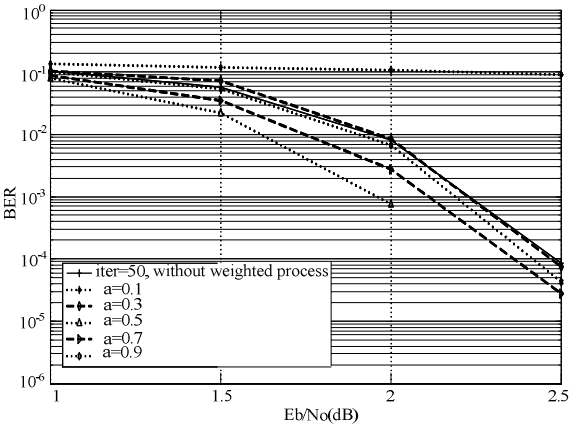


Fig. 4. The effect of different weighted coefficients

First, we set the inner iterative number $iter$ of SPA as 50, and choose different weighted coefficients. In Figure 4, we see that different weighted coefficients have different effect. Not all of the coefficients can elevate the performance, some even deteriorate the performance. The optimum coefficient is a=0.5 for $iter = 50$.

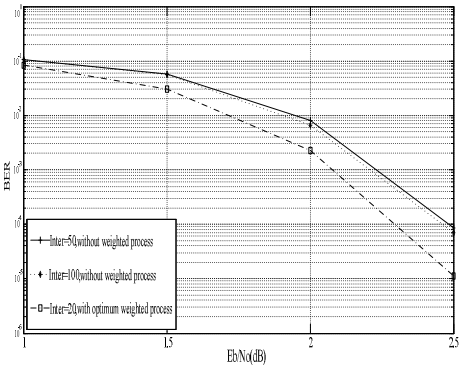


Fig. 5. The impact of the linear weighted process

Secondly, we change the inner iterative number $iter$ .In Figure 5, we see that, with the coefficient $a = 0.5$, the system in which $iter$ equal to 20 outperforms the system, which is without the weighted process with $iter = 100$. The former system performs 0.15 dB better than the latter at a BER of 10^{-4}. Given the same performance requirement, the linear weighted process can greatly decrease the inner iterative number, reduce the system latency and complexity, which makes concatenated system of FQPSK and LDPC code more feasible for the deep space communication application.

5 Conclusion

In this paper we have derived a new demodulation scheme for FQPSK, which named the MAP receiver based on the trellis diagram. In view of both the advantages of LDPC code and FQPSK, the combination of them is presented. When it comes to the phenomenon of positive feedback which is deleterious to the performance, a method named the linear weighted process is presented. Compared with the general concatenated system, the performance of the new system is improved in the presence of AWGN.

References

1. Skato, K.F.: XPSK: A new Cross-Correlated Phase-Shift-Keying Modulation Technique. IEEE Transactions on Communications 31(5), 701–707 (1983)
2. Simon, M.K., Yan, T.-Y.: Cross-Correlated Trellis Coded Quadrature Modulation, Patent Filed, October 5 (1999)
3. Simon, M.K., Divsalar, D.: A Reduced Complexity Highly Power/Bandwidth Efficient Coded FQPSK, Jet Propulsion Laboratory, Pasadena, California, May 15, pp. 1–17 (2001)
4. Simon, M.K.: Bandwidth-Efficient Digital Modulation With Application To Deep-Space Communications. Wiley, New Jersey (2003)
5. Lee, D., Simon, M.K., Yan, T.-Y.: Enhanced Performance of FQPSK-B Receiver Based on Trellis-Coded Viterbi Demodulatio. In: International Telemetering Conference, San Diego, California, pp. 23–26 (October 2000)
6. Simon, M.K., Yan, T.-Y.: Unfiltered Feher-patented quadrature phaseshift- keying (FQPSK): Another interpretation and further enhancements: Parts 1, 2. Applied Microwave & Wireless Magazine, 76–96, 100–105 (February/March 2000)
7. Bahl, L.R., Cocke, J., Jelinek, F., Raviv, J.: Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate. IEEE Trans. Inform., Theory IT-20, 284–287 (1974)
8. Gallager, R.G.: Low density parity check codes. IRE Trans. Info. Theory 8, 21–28 (1962)
9. Mackay, D.J.C.: Good Error-Correcting Codes Based on Very Sparse Matrices. IEEE Trans. Info. Theory 45(2), 399–431 (1999)
10. Kocarey, L., Tasev, Z.: A Vardy Improving Turbo Codes by Control Transient Chals in Turbo-decoding Algorithm. Electronic Letters 38(2), 1184–1186 (2002)

Design and Implementation of Efficient Reed-Solomon Decoder for Intelligent Home Networking

Ik Soo Jin

Dept. of Information & Communication Engineering, Kyungnam University,
449 Woryeong-dong, Masanhappo-gu, Changwon-si, Gyeongsangnam-do, 631-701, Korea
isjin@kyungnam.ac.kr

Abstract. Intelligent home concept is the integration of different services within a home by using a common communication system. Error correction coding is being used on an almost routine basis in most communication systems including intelligent home networking. In addition, critical protocol management information requires high fidelity forward error correction (FEC) coding to ensure that the protocol functions correctly in the worst case situations. A powerful technique in the worst case situations is a Reed-Solomon (RS) code. With this technique, one achieves a high level of performance by trading an increase in overall block length for a reduction in hardware complexity. This paper concerns the design and implementation of RS codec for intelligent home networking. The 3 symbol error correcting RS codec is developed on field programmable gate array (FPGA) chips using very high speed integrated circuit hardware description language (VHDL). The logic functions are confirmed by VHDL simulator tool. The circuits are generated by the synthesis tool. Two FPGA chips are required and the FPGA utilization is 70% and 40%, respectively.

Keywords: Reed-Solomon code, FEC, intelligent home networking, FPGA.

1 Introduction

Intelligent home concept is the integration of different services within a home by using a common communication system. The main advantages of an intelligent home system are created basically by using the synergies which arise when different functions are integrated within one system. It assures an economic, secure and comfortable operation of the home and includes a high degree of intelligent functionality and flexibility. The ongoing miniaturization and cost reduction in sector of electronic hardware has created ample opportunity for equipping private households with inexpensive intelligent devices for controlling and automating various tasks in our daily lives [1].

Communication and networking technologies have an important role in driving this development. The omnipresence of the internet via phone lines, TV cable, power lines, and wireless channels facilitates ubiquitous networks of intelligent devices that will significantly change the way we interact with home appliances. As in general, communication system aspects become more and more important in intelligent home.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 261–268, 2011.

It is obvious that such system will also influence significantly the domestic environment, including multimedia data services.

Error correction coding is being used on an almost routine basis in most communication systems including intelligent home networking. For example, to support data exchange with mobile devices, PLC (power line communication) networks will also need to cooperate with wireless network. A power line channel somewhat like a wireless channel: both supper from noise, fading, multipath, and interference. Power line noise is produced by the operating of electrical devices. In order to cope with thw wide variation in channel conditions, the PLC must be adaptive, intelligently using more robust modulation and coding schemes. In addition, critical protocol management information requires high fidelity forward error correction (FEC) coding to ensure that the protocol functions correctly in the worst case situations [2].

A powerful technique in the worst case situations is RS code [3]. RS code providing an efficient capability for correcting burst errors has been used in various communications and digital data storage systems, such as satellite and mobile communications, magnetic recording, and high definition television (HDTV) [4]-[5]. With this technique, one achieves a high level of performance by trading an increase in overall block length for a reduction in hardware complexity. The basic concept of the RS code is well understood, but the problem of design high bit rate RS encoder and decoder with a low complexity still remains an active area of research [6]-[7]. The codec design complexity increases considerably as the code length and the error correcting capability of the code increase. The efficient implementation plays an important role in the development of the codec with acceptable levels of circuit complexity and data speed to suit the particular application.

This paper describes the design and implementation of RS decoder for intelligent home networking. In the decoding process of RS code, the first step is to compute the syndromes. Syndromes are evaluated using Horner's rule [4]. Then, based on the syndromes, the error locator polynomial is found using the extended Euclidean algorithm proposed by Sugiyama et al. [8], a method for finding the greatest common divisor (gcd) of two polynomials. The next step is to find the roots of error locator polynomial. For this purpose, the Chien search method [9] is used. This method provides a simple and general algorithm for finding zeros of a polynomial over a finite field. And then the Forney's algorithm [10] is used to find the error values. Utilizing the above-mentioned algorithms, the (47, 41) RS codec is developed on FPGA chips using VHDL.

2 Mapping between Information Bits to M-Ary Channel Symbols

2.1 RS Encoder

RS codes are a sub-class of non-binary BCH codes. RS codes are optimum in that they are one of the few maximum distance separable (MDS) codes and their weight structure is known. The error correction capacity of a RS code is given in terms of the code symbol, thus it does not matter whether only one or all of the digits of a code

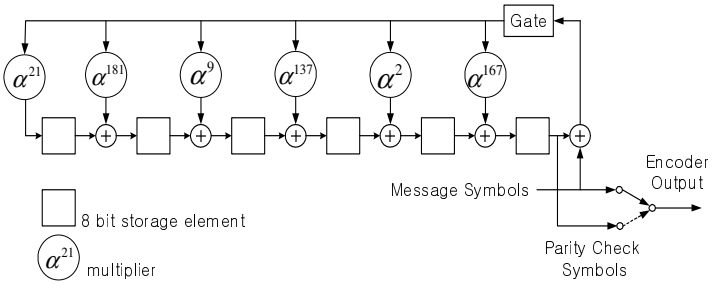


Fig. 1. (47, 41) RS encoder

symbol is in error. The RS encoder is shown in Fig. 1. The (47, 41) RS code which is capable of correcting up to 3 symbol errors is shortened from (255, 249) RS code. The generator polynomial $g(x)$ of (47, 41) RS code is

$$g(x) = \alpha^{21} + \alpha^{181}x + \alpha^{9}x^2 + \alpha^{137}x^3 + \alpha^{2}x^4 + \alpha^{167}x^5 + x^6 \tag{1}$$

where α is a primitive element in Galois field GF(2^8) and the primitive polynomial $p(x)$ is

$$p(x) = x^8 + x^4 + x^3 + x^2 + 1 . \tag{2}$$

In systematic form, the $2t$ parity check symbols are the coefficients of the remainder resulting from dividing the message polynomial $x^{2t}m(x)$ by $g(x)$. The state of the RS encoder, upon initialization, shall be the all-zero state. The 41 information symbols of 8 bits shall be output first, and the 6 parity symbols of 8 bits shall be output last.

2.2 RS Decoding Algorithm

RS decoding techniques can be classified into two categories, such as time domain techniques and frequency domain techniques. The time domain RS decoding estimates an error sequence directly based on the received data without a finite Fourier transform computations and can be easily implemented by the repeated use of simple basic cells. In contrast, the frequency domain RS decoding first estimates the Fourier transform of an error sequences, and then takes the corresponding inverse Fourier transform to obtain the sequence. Therefore the time domain RS decoding is more efficient for hardware implementation [5].

RS decoding algorithms are difficult because the decoding operation is basically solving a set of nonlinear simultaneous equations. In general, major procedures for both time domain and frequency domain RS decoding can be classified into four steps, such as syndrome calculation, evaluation of an error locator and error magnitude polynomial, error value searching, and error correction. The syndrome polynomial is used in the key equation solver for solving the key equation. The Euclidean algorithm, modified Euclidean algorithm, or Berlekamp-Massey algorithm

can be used to find error magnitude values corresponding to the error locations [11]. Since the decoding process which finds the error locator and error evaluator polynomial involves high computational complexity, it affects the speed and the hardware complexity of RS decoders.

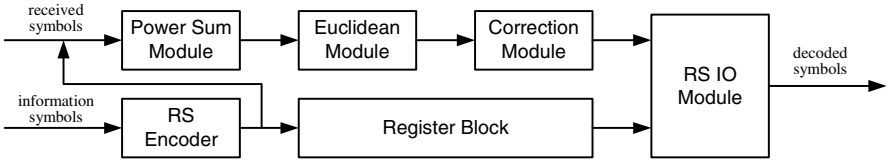


Fig. 2. The overall RS decoder architecture

Suppose a *t*-error correcting (*n, k*) RS code. Fig. 2 shows the overall block diagram of RS codec. The outline of decoding procedure is as follows:

Step 1) The first step is syndrome (S_j) computation. The 2*t* syndrome components are obtained by

$$S_j = y(\alpha^j) = \sum_{i=0}^{n-1} y_i \alpha^{ij}, \quad 1 \le j \le 2t \tag{3}$$

in which y_i represents received symbols. The power sums module (PSM) compute S_j using Horner's rule [4]:

$$S_j = (...((y_{n-1}\alpha^j + y_{n-2})\alpha^j + y_{n-3})\alpha^j + ...)\alpha^j + y_0 . \tag{4}$$

Step 2) The second step is to find the error locator polynomials $\sigma(x)$ and the error evaluator polynomials $\omega(x)$. $\sigma(x)$ and $\omega(x)$ satisfy the key equation

$$\sigma(x)S(x) = \omega(x) \bmod x^{2t} \tag{5}$$

where

$$S(x) = \sum_{j=1}^{2t} S_j x^{2t-j} . \tag{6}$$

For this purpose, the extended Euclidean algorithm proposed by Sugiyama et al. [8] can be used. The initial conditions of the algorithm are $\omega_T(x) \leftarrow x^{2t}$, $\sigma_T \leftarrow 1$, $\omega_B(x) \leftarrow S(x)$ and $\sigma_T \leftarrow 0$. See [11] for more details. The Euclidean module (EM) receives the syndrome digits S_j from PSM, runs the extended Euclidean algorithm, and outputs the coefficients of $\sigma(x)$ and $\omega(x)$.

Step 3) The third step of decoding is to find the roots of $\sigma(x)$. For this purpose, the Chien search method [9] can be used. This method is actually a trial and error approach for finding zeros of a polynomial over a finite field. Once the error locations (X_j) are known, the error values (Y_j) can be found using Forney's algorithm [10]. The error values are

$$Y_j = \frac{\omega(X_j)}{\sigma'(X_j)} X_j^{-1} = \frac{\omega(X_j)}{\sigma_{odd}(X_j)} X_j^{-1} \tag{7}$$

in which $\sigma_{odd}(X_j)$ is obtained from $\sigma(X_j)$ by taking the odd degree terms. The correction module (CM) evaluates $\sigma(x)$ and $\omega(x)$ at all the code locations, determines locations that are roots of $\sigma(x)$, and computes the corresponding error values. The output of the CM is a stream of error values.

3 Design and Simulation Results

Based on the above-mentioned algorithms, the (47, 41) RS codec is developed on FPGA chips using VHDL. For data service, information can be transmitted up to 128 *k*bps. The frame composed of 1536 bits is 10 *m*s in duration and the 10 *m*s frame is divided into four 384 bits sub-frames. The decoder accepts 376 bits excluding 8 tail bits in each sub-frame. The maximum data rate of RS codec is up to 512 kbps.

Fig. 3 shows the configuration of PCB board for test. Note that there are four commercial ROMs to store the received symbols of each sub-frame. Master clock

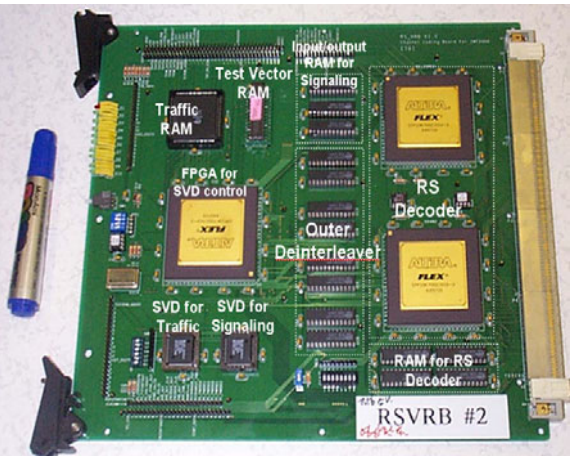
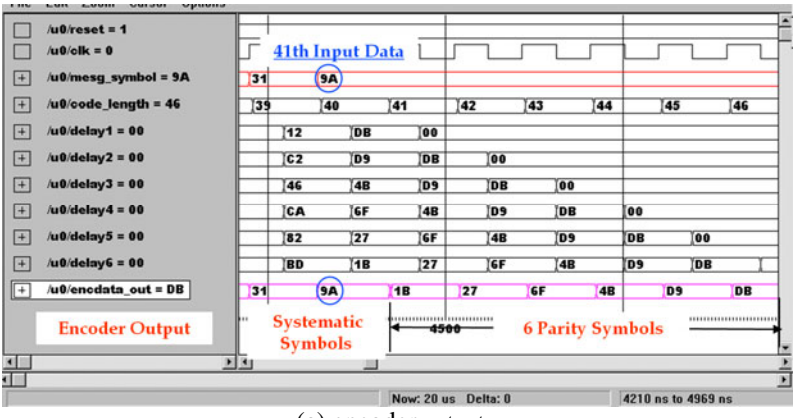
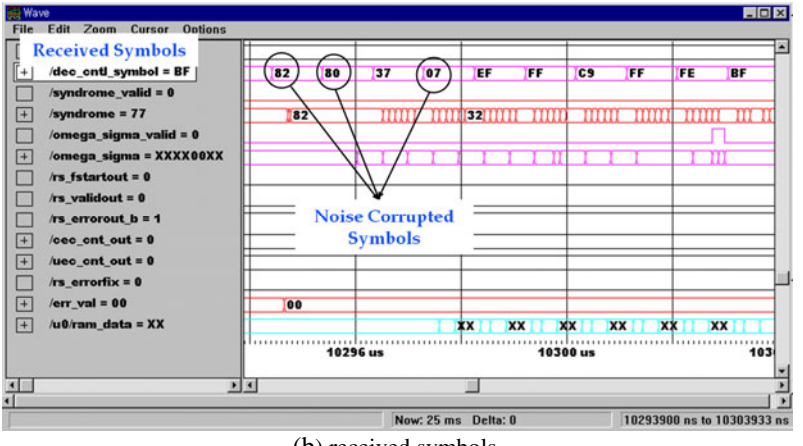


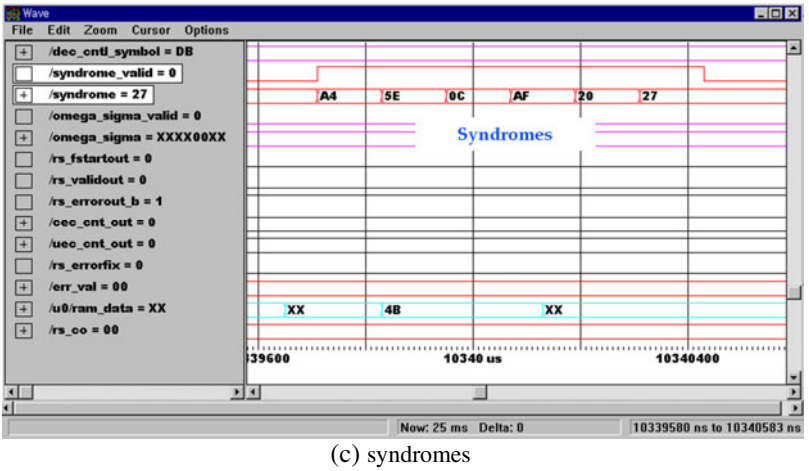
Fig. 3. Configuration of PCB board



(a) encoder outputs

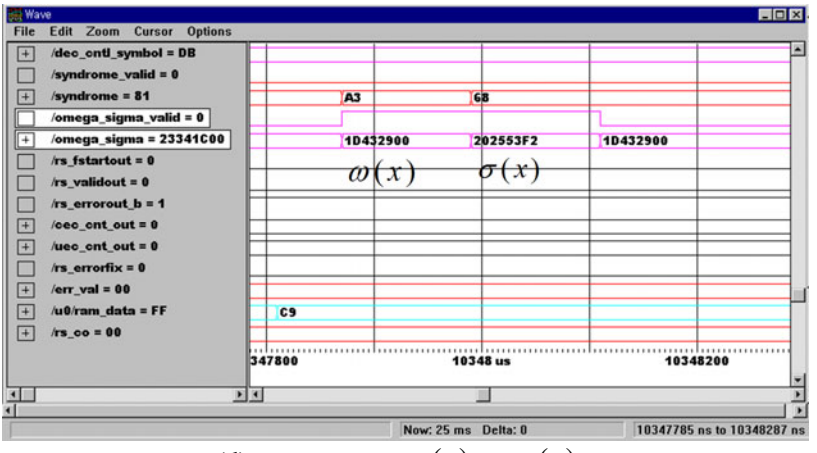


(b) received symbols

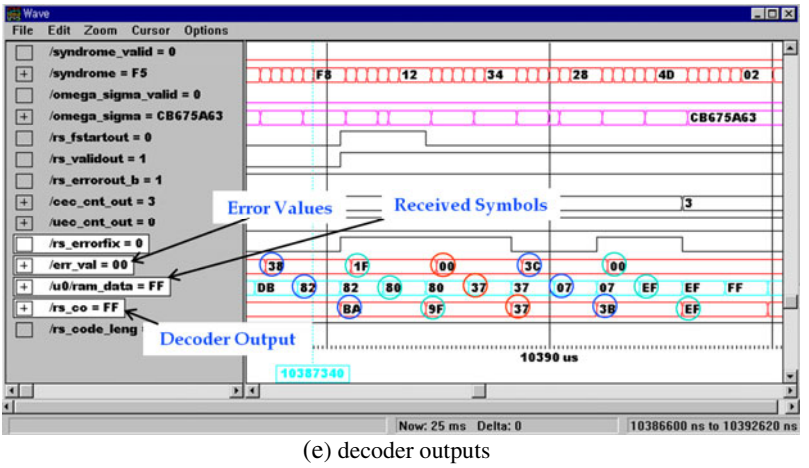


(c) syndromes

Fig. 4. VHDL post-simulation waveforms



(d) coefficients of $\omega(x)$ and $\sigma(x)$.



(e) decoder outputs

Fig. 4. (*continued*)

with 32 MHz is used to drive the RS codec. Two Altera FPGA chips and EPROMs are required for the implementation of RS codec. The FPGA utilization is 70% and 40%, respectively. LEDs are used to display the difference between the original input data and the decoding output data for self-test.

The logic design is accomplished by top-down methodology. The logic functions are confirmed by VHDL simulator tool. The circuits are generated by the synthesis tool. VHDL post-simulation waveforms are shown in Fig. 4. The encoder outputs named as *encdata_out* are illustrated in Fig. 4(a). "9A" is the 41st information symbol ("9A") and the rest (from "1B" to "DB") are the 6 parity symbols. The errors generated by C language are added in the received waveforms named as *dec_cntl_symbol*. The symbols "82", "80" and "07" in *dec_cntl_symbol* are corrupted by noise in Fig. 4(b). The syndromes S_j are shown in Fig. 4(c). Fig. 4(d) shows the

coefficients of $\omega(x)$ and $\sigma(x)$. In Fig. 4(e), it is shown that the signal *rs_errorfix* is high if the errors are detected in the corresponding received symbol. The decoded outputs named as *rs_co* are obtained from taking exclusive-OR between *dec_cntl_symbol* and *err_val*. For examples, "80" ⊕ "1F" = "9F", and "37" ⊕ "00" = "37" and so on.

4 Conclusions

This paper concerns the design and implementation of RS decoder for intelligent home networking. The RS code is a shortened (47, 41) block code of 8-bit symbols which is capable of correcting up to 3 symbol errors. The RS encoder and decoder has been implemented using FPGA and commercial memories. The operation of RS codec is verified on the board with self-test module. Two Altera FPGA chips and EPROMs are required for the implementation of RS codec. The FPGA utilization is 70% and 40%, respectively.

Acknowledgments. This work was supported by Kyungnam University Foundation Grant, 2011.

References

1. Suh, C., Ko, Y.-B.: Design and implementation of intelligent home control systems based on active sensor networks. IEEE Trans. Consumer Electronics 54, 1177–1184 (2008)
2. Lin, Y.-J., Latchman, H.A., Lee, M.: A power line communication network infrastructure for the smart home. IEEE Wireless Communications 9, 104–111 (2002)
3. Xi, S., Ying, G.: The Application of RS code in CCK Modulation Technology. In: International Conference on Information Science and Engineering (ICISE), pp. 678–680 (2009)
4. Shao, H.M., Truong, T.K., Deutsch, L.J., Yuen, J.H., Reed, I.S.: A VLSI design of a pipeline Reed-Solomon decoder. IEEE Trans. Computer C-34, 393–403 (1985)
5. Shao, H.M., Reed, I.S.: On the VLSI design of a pipeline Reed-Solomon decoder using systolic-array structure. IEEE Trans. Computer 37, 1273–1280 (1988)
6. Zhu, J., Zhang, X.: Efficient Reed-Solomon decoder with adaptive error-correcting capability. In: The 19th International Conference on Wireless and Optical Communications Conference (WOCC), pp. 1–4 (2010)
7. Lee, H.: An ultra high-speed Reed-Solomon decoder. In: International Symposium on Circuits and Systems, pp. 1036–1039 (2005)
8. Sugiyama, Y., Kasahara, M., Hirasawa, S., Namekawa, T.: A method for solving key equation for decoding Goppa codes. IEEE Trans. Information and Control 27, 87–99 (1975)
9. Chien, R.T.: Cyclic decoding procedures for BCH codes. IEEE Trans. Information Theory IT-27, 254–256 (1965)
10. Forney, G.D.: On decoding BCH codes. IEEE Trans. Information Theory IT-11, 393–403 (1992)
11. Wicker, S.B., Bhargava, V.K.: Reed-Solomon codes and their applications. IEEE Press, New York (1994)

A Cost-Effective Multicasting Using an FP-LD Modulator in a WDM-PON

Hyuek Jae Lee

Division of Information & Communication Engineering, Kyungnam Univesity
449 Wolyoung-dong, Chanwon, 631-701, Korea
hyuek@kyungnam.ac.kr

Abstract. An external optical modulation method based on TE/TM-mode absorption nulls in a Multiple Quantum Well (MQW) Fabry-Perot laser diode (FP-LD), has been proposed and experimentally demonstrated for multicasting in a WDM-PON. The 32 channels multicasting system with the proposed modulator has been implemented, which shows power penalties of 1.53~4.15 dB at a bit error rate of 10^{-9} with more than extinction ratios of 14.5 dB at 622 Mbps.

Keywords: External optical modulator, Fabry-Perot laser diode, Passive optical network, multicasting.

1 Introduction

Recently, Wavelength Division Multiplexing (WDM)-Passive Optical Network (PON) for optical access networks has received a great deal of an attention. WDM-PON has a high security, a protocol transparency, and a wide bandwidth compared with Time Division Multiplexing (TDM)-PON. However, it requires an expensive wavelength specified optical source such as a distributed feedback laser to maintain an assigned wavelength [1]. To solve the problem, several optical sources [2, 3] with a color-free and low cost have been proposed. However, even though virtual point-to-point connections with such the optical sources can be made cost-effectively, it is inherently difficult to provide a multicasting (or broadcasting) function. Thus, optical sources for multicasting [4-6] or several multicasting overlay schemes [7,8] have been actively studied. In [4], a spectrum sliced amplified spontaneous emission (ASE) source has been proposed for multicasting but it requires an expensive external modulator and shows limited performances. The mutually injected Fabry-Perot Laser Diodes (FP-LDs) [5] and the side-mode suppressed multi-wavelength fiber laser [6] are good candidates. However, they require a sophisticated alignment method for a short cavity length, and still need an expensive optical modulator if not directly modulated. For the multicast overlay schemes [7,8], both point-to-point and multicast connections can be simultaneously constructed by using a single optical source. However, a complicated coding scheme and a synchronized controller are required. Moreover, expensive external modulators still make the system complex and costly.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 269–276, 2011.

In the previous work [9], an external optical modulator employing a low-cost multiple-quantum-well (MQW) FP-LD has been proposed. The work is based on the absorption null modulation and extended to wavelength converters in other previous works [10,11]. In this paper, further experiments are reported that it is possible for the proposed FP-LD modulator to have more than 622 Mbps modulation speed and to apply to multicasting transmission in a WDM-PON.

2 Principle

When a multiple-quantum-well (MQW) type FP-LD is driven over a threshold current, the FP-LD makes light by lasing on TE-mode. On the other hand, TM-mode in the FP-LD shows only absorption nulls not lasing owing to very small TM-gain inside the Fabry-Perot mirror (resonator). If no current is applied to the FP-LD, both modes (TE- and TM-mode) have absorption nulls. As the current increases from 0 mA to the threshold current, all the absorption nulls are moved to short-wavelength without any lasing on both modes. Even though light due to spontaneous emission occurs, the light at small current level can be ignored in comparison to externally incident light for optical modulation. The wavelength of the incident light is spectrally aligned with the center wavelength of one among the absorption nulls. Thus, optical modulation can be achieved by vibrating the center wavelength of the null due to current signal injection into a PF-LD.

The shift of the absorption null comes from the change of the refractive index in an MQW FP-LD waveguide, which is due to anomalous dispersion, plasma effect, and bandgap shrinkage [13]. However, the bandgap shrinkage can be neglected when an input light wavelength is longer than a bandgap wavelength of the waveguide in the MQW FP-LD [14]. In this case, the anomalous dispersion and the plasma effect become dominant for the refractive index change. The refractive index change due to the plasma effect in a bulky waveguide can be written as follow [13,16].

$$\Delta n = \frac{e^2 N}{2\omega^2 \varepsilon_0 \varepsilon_r m_c^*} n \tag{1}$$

where, N is a carrier density, n is the refractive index in the absence of the plasma effect, $\varepsilon_0 \varepsilon_r$ is the dielectric constant of the active region, and m_c^* is the effective mass of the carrier. The total refractive index change due to the plasma effect in a MQW waveguide can be easily obtained by summing refractive index changes in wells and barriers [13]. In (1), because $\omega = 2\pi / \lambda$, the $\sqrt{\Delta n}$ is proportional to an input light wavelength λ. On the contrary, a refractive index change due to the anomalous dispersion according to the wavelength λ is reduced. Therefore, as the input light wavelength increases, the plasma effect is more dominant than the anomalous dispersion. As a result, the total refractive index change becomes almost constant for a wide range of wavelengths (> 100nm) [13].

Let us consider a modulation speed for the proposed PF-LD modulator. To operate a PF-LD as an optical modulator, the injection current J with less than a threshold current J_{th} must be applied, i.e. carrier density N < threshold density N_{th}. As a

consequence, a photon density N_{ph} becomes almost zero, which means a stimulated emission can be neglected. Thus, the rate equation for the carrier density N can be expressed as [17]

$$\frac{dN}{dt} = \frac{J}{qd} - \frac{N}{\tau_e(N)} \tag{2}$$

where, q and d are the charge constant and the thickness of the active layer, respectively. And, $1/\tau_e(N)$ is the effective recombination rate, which $1/\tau_e(N) = A + BN + CN^2$, where AN is the nonradiative recombination rate, BN^2 is the radiative spontaneous rate, and CN^3 is the Auger recombination rate. Because of the carrier dependence term $\tau_e(N)$, Eq. (2) is nonlinear and can be solved numerically or approximately. Assuming τ_e is constant, i.e. $B = C = 0$, we can know easily that the carrier density $N(t)$ increases logarithmically and decays exponentially with the time constant τ_e. Thus, the speed for the proposed modulator strongly depends on τ_e. By the calculation from typical parameter values (A= 5×10^{8}, B= 1×10^{-10}, C= 3×10^{-29}, $N_0 = 1\times10^{18}$, and $N_{th} = 2.61\times10^{18}$) in [17], $\tau_e(N_0) \approx 1.58$ nsec and $\tau_e(N_{th}) \approx 1.04$ nsec, which mean roughly 630 Mbps ~ 1 Gbps modulation speed.

3 External Optical Modulation by Using a FP-LD

The proposed FP-LD modulator works under threshold current. Its basic principle is as followings: As current of FP-LD increases from 0 mA to threshold current, absorption nulls move to short-wavelength without any lasing on TM-mode. Meanwhile, the FP-LD on TE-mode makes very little light by the spontaneous emission. Only a small current can move the absorption nulls enough to modulate optical signals. If the current injection is stopped, the nulls moves back to the original position. Such a movement can make the FP-LD to act as an external optical modulator. A previous work [9] has already shown that the absorption nulls in a FP-LD is used for optical modulation on both TM- and TE-mode at the data rate of 155 Mbps. However, there are no results to report about what is the maximum speed of modulation.

The modulation by a FP-LD was tested. The FP-LD is an InGaAsP MQW type, and has a nominal wavelength of 1549.5 nm and longitudinal mode spacing of 1.16 nm. Also it has the threshold current of ~8 mA. To observe absorption nulls spectra, an amplified spontaneous emission (ASE) source with a bandwidth of ~26 nm is replaced at the position of the T-LD in Fig. 1. The broadband ASE light enters the FP-LD via the polarization controller1 (PC1) and the optical circulator (OC). By controlling the PC2, only a TM-mode beam can be set to pass through the optical polarizer (OP). As can be seen from the inlet figure of Fig. 1, the absorption nulls with the spacing of ~1.16 nm are observed, which are almost the same with the lasing mode spacing. According to the current injection, the TM-mode absorption nulls rapidly move to the short-wavelength. However, for over the threshold current

(~8 mA), the nulls gradually move back to the long-wavelength due to the thermal effect by lasing [12]. At the start for modulation, the incident light is spectrally aligned at the center wavelength of one among the absorption nulls with no bias current. The bias current makes the nulls to move to short-wavelength. As seen from the data in Fig. 1, only a current injection of ~1 mA can make ~0.26 nm wavelength shift for the nulls. Thus, by applying the current on/off, the nulls move back and forth to modulate optical signal. The wavelength shift of 0.26 nm is equal to ~32.5GHz, which is enough to modulate the optical signal at 622 Mps.

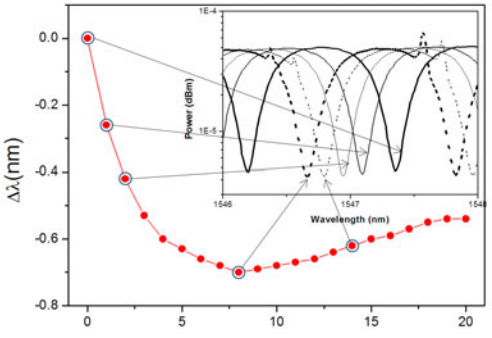


Fig. 1. Center wavelength displacement of the absorption null as a function of injected current (0-20 mA). The inlet is TM-mode absorption spectra of the FP-LD according to the current change.

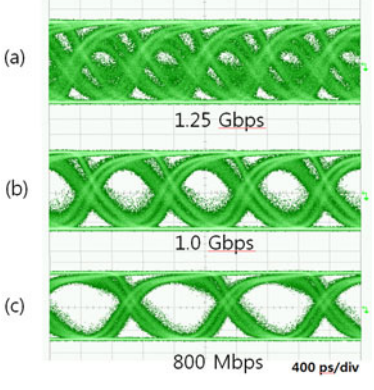


Fig. 2. Oscilloscope traces for three modulation speeds, (a) 1.25, (b) 1.0, and (c) 0.8 Gbps by using the proposed external optical modulator

The shift of the absorption nulls comes from the change of the refractive index in an MQW FP-LD waveguide, which is due to anomalous dispersion, plasma effect, and bandgap shrinkage [13]. However, the bandgap shrinkage can be neglected when an input light wavelength is longer than a bandgap wavelength of the waveguide in the MQW FP-LD [14]. By the calculation of the rate equation for the carrier density

N, the modulation speed can be roughly estimated to 630 Mbps ~ 1 Gbps [9]. To know the maximum speed of modulation for the FP-LD used in this experiment, 0.8, 1.0, and 1.25 Gbps modulations are tested and Fig. 2 shows the oscilloscope traces for the three data rate. It is seen that the maximum speed of modulation is about 800 Mbps. To get higher modulation speed (around 1.0 Gbps), careful design of electrode in the FP-LD and smaller current injection might be needed.

4 Multicasting by Using FP-LD Modulation in a WDM-PON

Fig. 3 shows the experimental setup for multicasting in a WDM-PON using the proposed FP-LD optical modulator. For cost-effectiveness of the multicasting, a multi-wavelength laser source [5,6] with the same periodicity of the longitudinal mode, should be used. However, due to be unavailable for the multi-wavelength source in the laboratory, the tunable-LDs are used as shown in Fig. 3. It seems reasonable to suppose that this paper focuses only on the possibility of an external optical modulator for multicasting using a FP-LD.

The gray colored blocks show the conceptual scheme for a WDM-PON system, which are not constructed in this experiment. The FP-LD generally has absorption nulls on the TM-mod. The absorption nulls are made by being kept and dissipated in a FP-LD cavity for the input light with wavelength such that the phase after each round trip has to be an integral multiple of 2π. In other words, $\lambda_p=2nL/p$, where λ_p is the wavelength of the pth cavity mode, L is the cavity length. By some manipulations, the spacing between two adjacent modes as this, $\lambda_p-\lambda_{p+1} \cong \lambda^2/2nL$. The null spacing for the InGaAsP MQW FP-LD used in this experiment is around 1.16 nm as shown in Fig. 2, which corresponds to cavity length of ~296μm (by using $n \cong 3.5$ for the substrate of InP and $\lambda_p= 1550nm$). In order to match with the channel spacing (0.8 nm) of arrayed waveguide grating (AWG) for multicasting, the cavity length need to be increased to ~429μm. Due to the periodicity of nulls of PF-LD, the proposed modulator combined with such a multi-wavelength source in [5,6] can be easily applied for multicasting in a WDM-PON.

In Fig. 3, the CW lights from T-LD1 and T-LD2 are sent to the FP-LD modulator, which are modulated at 622 Mbps (2^{31}-1 PRBS). Here, the polarization of the CW lights should be aligned on the TM-mode by adjusting the PC1 and PC2. The 622 Mbps modulated signal is sent to the optical coupler through OC, PC3, OP, EDFA, and optical attenuator (OA) as shown in Fig 3. After passing through 20 km single mode fiber (SMF), the signal is wavelength-demultiplexed to one of the optical network units (ONUs) by the 32 channel AWG located at the remote node (RN). Although only two sources, i.e. T-LD1 and T-LD2, are used in this multicasting experiment, the BERs for all the channels of the AWG can be measured by tuning the wavelength of the T-LD1 and T-LD2. Also, a thermoelectric cooler/heater (TEC) controller in the AWG module is adjusted to align to the center wavelength of the FP-LD's nulls. The insertion loss of the SMF is about 4.6 dB. The 32-channel AWG device has a channel spacing and a 3dB width of 100GHz and ~30 GHz, respectively. Fig. 4 shows BER measurements after 20 km transmission for 32-channels at 622

Mbps. The power penalties are observed from 1.53 to 4.15 dB compared to the back-to-back at a bit error rate of 10^{-9}. On the other hand, 1.0 Gbps modulation shows the error floor at a bit error rate of 10^{-6}.

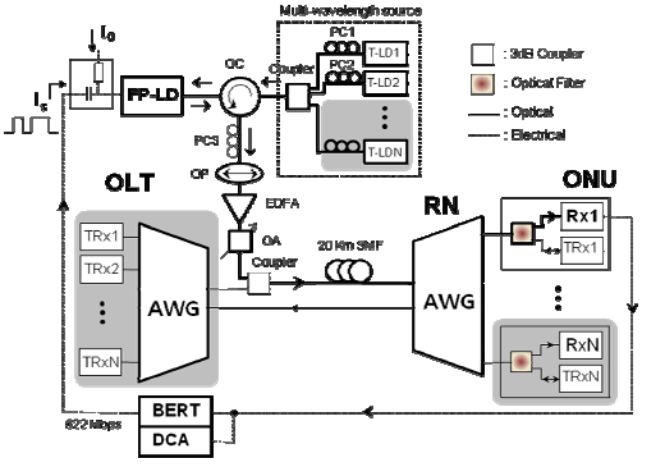


Fig. 3. Experimental setup for multicasting in a WDM-PON using the proposed external optical modulator

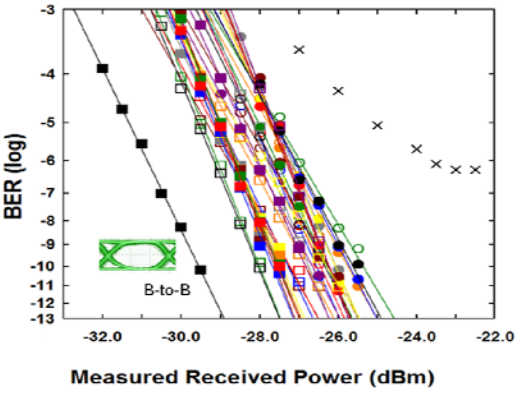


Fig. 4. Measured BER curves for back-to-back and 32 channels after 32 km transmission at 622 Mbps. For the comparison, the BER at 1.0 Gbps is measured.

If a 1300 nm FP-LD is used, both 1300 nm and 1500 nm wavelength light will be simultaneously modulated. This is because the FP-LD with gain in 1300 nm region is transparent to 1550 nm light [15]. Even though current injection leads to a change of effective refractive index in 1300 nm region, 1300 nm as well as 1550 nm light experiences the shift of absorption nulls. From the results, the proposed modulator is

likely to be a good candidate for ultra broadband multicasting with a low cost in a WDM-PON.

5 Conclusions

In this paper, a new optical modulation method for multicasting using a low-cost FP-LD has been proposed and experimentally demonstrated. For 32 multicasting channels, the maximum and minimum of measured power penalties at a bit error rate of 10^{-9} were 1.53 and 4.15 dB. Although the multicasting experiment in this paper has been conducted at the data speed of 622 Mbps, it may be possible to get around 1.0 Gbps modulation speed by concisely controlling injection current and designing the electrode of FP-LD. Moreover, if a 1300nm FP-LD is used as the proposed optical modulator, wideband operation will be working in the wavelength range of 1300nm to 1550nm.

References

1. Lee, C.H., Sorin, W., Kim, B.Y.: Fiber to the home using a PON infrastructure. J. Lightw. Technol. 24, 4568–4583 (2006)
2. Kim, H.D., Kang, S.G., Lee, C.H.: A low-cost WDM source with an ASE injected Fabry-Perot semiconductor laser. IEEE Photon. Technol. Lett. 12, 1067–1069 (2000)
3. Kim, D.C., Choi, B.S., Kim, H.S., Kim, K.S., Kwon, O.K., Oh, D.K.: 2.5 Gbps Operation of RSOA for Low Cost WDM-PON Sources. In: Proc. 35 th European Conference on Optical Communication (ECOC 2009), Vienna, Austria, Paper P2.14 (2009)
4. Cho, J., Kim, J., Gutierrez, D., Kazavsky, L.G.: Broadcast Transmission in WDM-PON using a Broadband Light Source. In: Proc. Conf. Optical Fiber Commun (OFC/NFOEC 2007), Anaheim, CA, Paper OWS7 (March 2007)
5. Yoo, S.H., Lee, H.K., Lim, D.S., Jin, J.H., Byun, L., Lee, C.H.: 2.5-Gb/s broadcast signal transmission in a WDM-PON by using a mutually injected Fabry-Perot laser diodes. In: Proc. Lasers and Electro-Optics(OSA/CLEO 2011), SanJose, CA, Paper CFH7 (May 2011)
6. Lee, K.I., Lee, S.B., Lee, J.H., Kim, C.H., Han, Y.G.: Side-mode Suppressed Multiwavelength Fiber Laser and Broadcast Transmission. In: Proc. Conf. Optical Fiber Commun (OFC/NFOEC 2008), San Diego, CA, Paper OThF1 (February 2008)
7. Xu, J., Zhang, Y., Chen, L.K., Chan, C.K.: A Delay-Based Multicast Overlay Scheme for WDM Passive Optical Networks With 10-Gb/s Symmetric Two-Way Traffics. J. Lightw. Technol. 28, 2660–2666 (2010)
8. Cai, L., Liu, Z., Xiao, S., Zhu, M., Li, R., Hu, W.: Video-Service- Overlaid Wavelength-Division-Multiplexed Passive Optical Network. IEEE Photon. Technol. Lett. 21, 990–992 (2009)
9. Lee, H.J., Won, Y.H.: External Optical Modulator Using a Low-cost Fabry-Perot Laser Diose for Optical Access Networks. J. Opt. Soc. Korea 8, 163–167 (2004)
10. Yoo, H., Lee, H.J., Jeong, Y.D., Won, Y.H.: All-optical Wavelength Conversion at 10 Gbit/s using Absorption modulation in a Fabry-Perot Laser Diode with a CW Holding Beam. Microwave Opt. Technol. Lett. 47, 508–511 (2005)

11. Tran, Q.H., Cho, J.S., Jeong, Y.D., Won, Y.H.: All-optical Multi-wavelength Conversion using Absorption Modulation of an Injection Locked FP-LD. IEICE Electron. Express 4, 612–616 (2007)
12. Higashi, T., Yamamoto, T., Ogita, S., Kobayashi, M.: Experimental analysis of temperature dependence of oscillation wavelength in quantum-well FP semiconductor lasers. IEEE J. Quantum Electron 34, 1680–1689 (1998)
13. Shim, J.I., Yamaguchi, M., Kitamura, M.: Refractive index and loss changes produced by current injection in InGaAs(p)-InGaAsP Multiple Quantum-Well (MQW) waveguides. IEEE J. Select. Topics Quantum Electron 1, 408–415 (1995)
14. Bennett, B.R., Soref, R.A., Del Alamo, J.A.: Carrier-induced change in refractive index of InP, GaAs, and InGaAsP. IEEE J. Quantum Electron 29, 113–122 (1990)
15. Lacey, J.P.R., Pendock, G.J., Tucker, R.S.: All-optical 1300-nm to 1550-nm Wavelength Conversion using Cross-Phase Modulation in a Semiconductor Optical Amplifier. IEEE Photon. Technol. Lett. 8, 885–887 (1996)
16. Iizuka, K.: Elements of Photonics, vol. II, ch. 14. Wiley & Sons, New York (2002)
17. Liu, M.M.-K.: Principles and applications of optical communications, ch. 12. IRWIN, Times Mirror Higher Education Group, Inc, (1996)

A Wireless CCTV Converter Based on Binary CDMA Technology

Yeong-Jin Baek[1] and Sang-Hoon Lee[2]

[1] Department of Information and Communication Engineering, Kyungnam University, Changwon, Korea
likedudwls@naver.com
[2] Department of Electronic Engineering, Kyungnam University, Changwon, Korea
sanghoon@kyungnam.ac.kr

Abstract. This paper describes the design and implementation of a wireless CCTV converter system using binary CDMA technology. The proposed system consists of a CPU module based on ARM11 (S3C6410) processor, a binary CDMA communication module based on Koinonia chip (KWPAN1200A), and interface module. All processing procedures of image coding/compression and decoding/decompression are carried out by H.264 algorithm. The experimental results show that the proposed system can transmit a high resolution image of VGA (720x480) up to 300 meters with a frame rate of 22 fps. And the system can control remotely functions such a pan, tilt and zoom.

Keywords: CCTV converter, binary CDMA, H.264.

1 Introduction

The CCTV system has been played an important role in not only the protection of property but prevention of crime. The advantage of using wired CCTV systems is that these systems are long range networking, but they are cost ineffective due to installation of complex links.

Moreover, the conventional wireless CCTV system has good picture quality based on analog technologies but has weak points such a short transmission distance (< 50 m), low noise immunity, weak security, and uncontrollable pan/tilt functions.

A wireless communication for image transmission can use such technologies as a wireless LAN, an UWB and a binary CDMA. The wireless LAN technology has weaknesses for security and for obstacles like iron materials, noise, magnetic field. An UWB technology can't be applied to CCTV converter because it can cover a short range, although it can transmit image information [1-3].

On the other hand, the binary CDMA technology has become very effectual in wireless communication because it has many strong points; small size, low power consumption, high security, strong to noise and interference, high and variable data rate, long transmission distance (~500 m), robust QoS (quality of service), encryption and compression functions, and transmittable moving picture under 80 km/hr [4-5].

In this paper, we focus on a system implementation of a wireless CCTV converter using binary CDMA technology. The proposed system consists of an ARM processor

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 277–282, 2011.

module, a binary CDMA communication module, and interface module between a processor module and a binary CDMA module.

2 Wireless CCTV Converter

2.1 Main Controller

Fig. 1 shows the system configuration for main controller based on ARM11 (S3C6410) processor. The S3C6410 core can connect multiple peripheral devices via multiple GPIO ports and support development of various solutions using these properties. The core operating at 533 MHz has abilities to extend functions and to handle multimedia data because it can support internal CODEC H/W (H.264/MPEG4/VC1). The S3C6410 core includes many types of sub-blocks; a display controller, a DMA controller, a TV scalar, a camera interface, and so on. We used the local bus interface to control the binary CDMA module. Moreover, 133 MHz DDR2 SDRAM and 128 MB NAND flash memory are used in external devices.

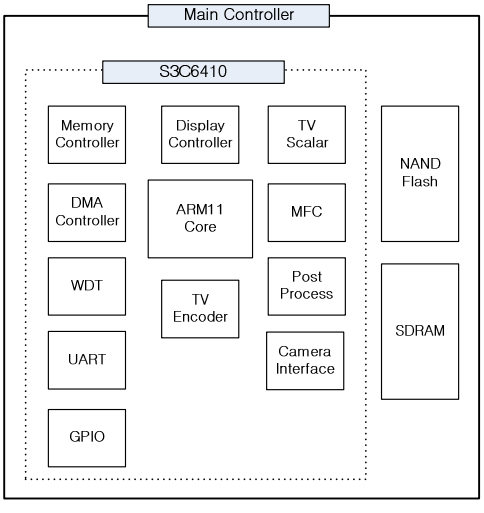


Fig. 1. Main controller configuration

2.2 Wireless CCTV Converter Master and Slave

Fig. 2 shows the functional block diagram of a CCTV converter master. The CCTV converter master consists of two functional blocks; a receiving block, and a transmitting block. The receiving block receives image data from CCTV converter slave through a binary CDMA module. After H.264 decoder decodes the data via stream buffer, it saves decoded data in a frame buffer. And then the stored data is converted into NTSC signal with a TV scalar and a TV encoder. Finally, they are to be generated into TV out port. The transmitting block sends the RS-485 data from DVR to binary CDMA module through UART device.

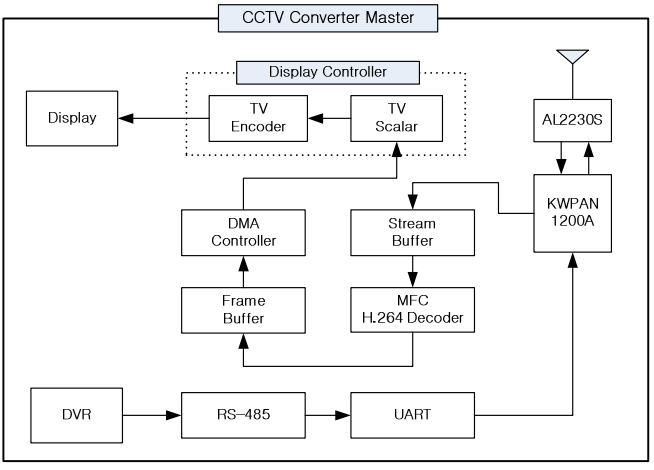


Fig. 2. Block diagram of a CCTV converter master

Fig. 3 shows the functional block diagram of a CCTV converter slave. The CCTV converter slave consists of two functional blocks; a receiving block, and a transmitting block. First, the transmitting block receives NTSC image data from analog camera. After TVP5150A video decoder chip convert NTSC image into digital format, it sends digital data to camera interface.

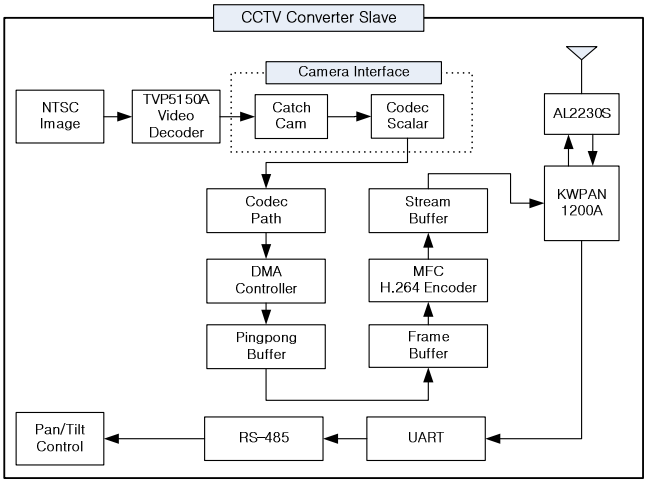


Fig. 3. Block diagram of a CCTV converter slave

And then H.264 encoder encodes the stored data in frame buffer and sends them to binary CDMA module. KWPAN1200A Koinonia chip in the binary CDMA module converts the encoded data into a type of spread binary CDMA packet.

The receiving block receives the RS-485 data from UART through the binary CDMA module. The RS-485 data are finally to be generated into output port. The RS-485 data controls the pan and tilt functions of camera.

2.3 Implementation of the Wireless CCTV Converter

Fig. 4 shows the prototype of the wireless CCTV converter. The system contains a main controller including a processor, a binary CDMA communication module including RF chips, and peripheral parts including various I/O ports.

The main controller (①) which is fabricated with a type of piggyback, consists of S3C6410 RISC processor, 133 MHz DDR2 SDRAM, and 128 MB NAND flash memory supported by Samsung Electronics. The binary CDMA module (②) which is fabricated with a size of 40x30 mm^2, consists of an AL2230S RF chip and a KWPAN1200A baseband chip of binary CDMA manufactured by Daewoo electronics. The AL2230S RF device can offer IEEE 802.11b/g compatibility. The peripheral parts include many types of I/O ports; NTSC image ports (③), power supply (④), serial output and connector for debugging (⑤), LED display (⑥), RS-485 port (⑦), and video encoder chip (⑧).

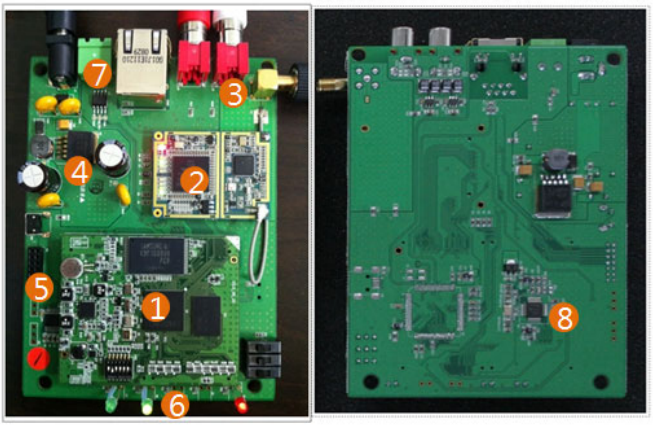


Fig. 4. Prototype of a wireless CCTV converter

3 Experimental Results

Fig. 5 shows the measured waveforms of the encoding and the decoding signals in the CCTV converter master and slave, respectively. The upper signal indicates to be transmitted frame data after H.264 encoding in slave side. The lower signal shows a frame data after H.264 decoding in the master side before TV out.

Fig. 6 shows the test results of image data which are transmitted and received through the wireless CCTV converter. The test is carried out in room of 10x10 m2 under operating wireless AP. Fig. 6(a) shows the regenerated output image with

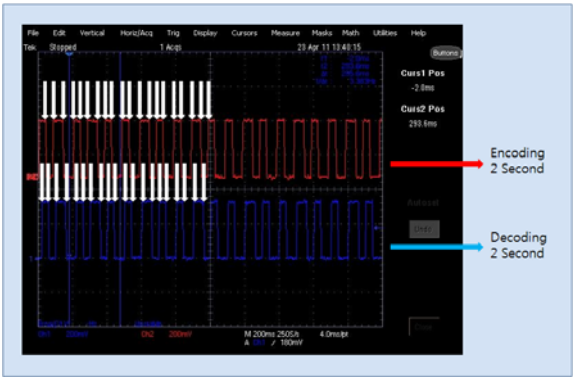


Fig. 5. Encoding/decoding frames in the CCTV converter master and slave

720x480 VGA resolutions after a series of procedures from the CCTV converter slave to master through binary CDMA module. The wireless CCTV converter slave received the original image captured by an analog camera and then transmitted them to wireless CCTV master. Fig. 6(b) shows comparative images with or without the image processing by the wireless CCTV converter. The picture of Fig. 6(c) is directly acquired from the analog camera without the CCTV converter, whereas the picture of Fig. 6(d) is regenerated from the wireless CCTV converter. We could get the clear image without damage through the wireless CCTV converter.

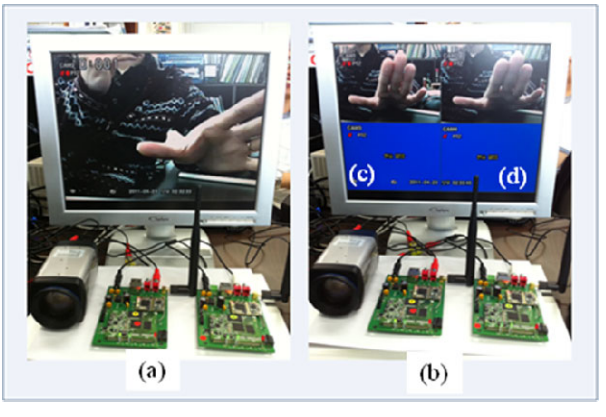
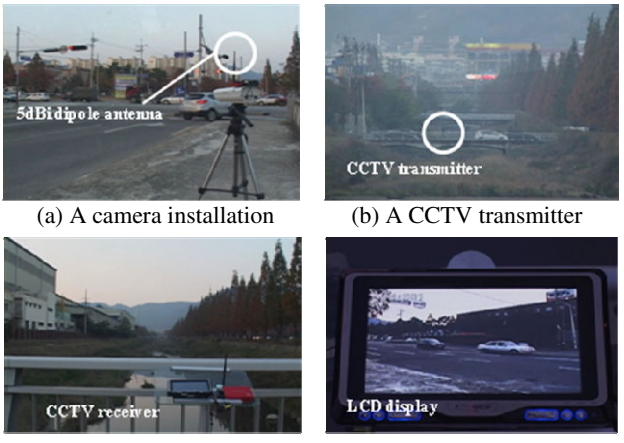


Fig. 6. Test results of the CCTV converter

Fig. 7 shows images measured at an outdoor in order to test the transmitting distance between a CCTV transmitter and a receiver. Fig. 7(a) shows a camera installation in the CCTV transmitter side with a 5dBi dipole antenna. Fig. 7(b) and Fig. 7(c) show a CCTV transmitter and a receiver installed on the bridge, respectively. The distance between the transmitter and the receiver is in the range of 300 ~ 500 m. Fig. 7(d) shows the image regenerated by the CCTV converter master in case the distance between the transmitter and the receiver is about 300 m.



(a) A camera installation (b) A CCTV transmitter

(c) A CCTV receiver (d) A received image

Fig. 7. Field tests at an outdoor

4 Conclusions

In this paper, we introduced a system implementation of a wireless CCTV converter using binary CDMA technology. The experimental results show that the proposed system can transmit a high resolution image of VGA (720x480) up to 300 meters with a frame rate of 22 fps. And the system can control remotely functions such a pan, tilt and zoom.

Acknowledgments. This work was supported by Kyungnam University Foundation Grant, 2011.

References

1. IEEE, Standards for Part 15.3: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High-Rate Wireless Personal Area Networks (WPANs) (Sepember 2003)
2. IEEE, Standards for Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs) (October 2003)
3. IEEE, Standards for Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs) (June 2002)
4. Koinonia Standardization ver. 1.1, KETI (August 2005)
5. Karaoguz, J.: High-rate Wireless Personal Area Networks. IEEE Communications Magazine 39(2), 96–102 (2001)

A Soft QoS Supporting Multi-path Routing Scheme for Mobile Nodes in MANETs

Tz-Heng Hsu[1], Yi-Da Li[1], and Meng-Shu Chiang[2]

[1] Department of Computer Science and Information Engineering,
Southern Taiwan University, Tainan, Taiwan, R.O.C.
[2] Department of Computer Science and Information Engineering,
Far East University, Tainan, Taiwan, R.O.C.
hsuth@mail.stut.edu.tw

Abstract. In this paper, a soft QoS supporting multi-path routing scheme for mobile nodes in MANETs is proposed. The proposed multi-path routing scheme can obtain network metrics of nodes during route discovery. A source node can then select the most suitable routing paths to transmit packets to the destination node according to the QoS requirement of its network applications. The experiments show that the proposed scheme can prolong the network lifetime by providing soft QoS support to mobile nodes.

Keywords: MANET, QoS, Multi-path Routing, Disjoint Path, Type of Service.

1 Introduction

A mobile ad hoc network (MANET) is a wireless network consisting of two or more mobile nodes that dynamically reconfigure wireless network with no fixed infrastructure. In a MANET, it doesn't require a centralized access point or a base station to coordinate the communications among mobile nodes; instead, every node can act as a router to forward packets to other nodes within its radio transmission range. Meanwhile, every node can move arbitrarily and communicate with each other by using multi-hop packet forwarding mechanism.

Several routing protocols for MANET have been proposed [1, 2, 3, 4, 5]. These routing protocols can be classified into proactive and reactive categories. The proactive protocols, e.g. Destination Sequenced Distance Vector (DSDV) [4] and Optimized Link State Routing (OLSR) [1], maintain routing information from nodes to nodes periodicity. The reactive protocols, e.g., Ad hoc On-Demand Vector (AODV) [5] and Dynamic Source Routing (DSR) [3], obtain routing information only when a route to a destination is required. Most routing protocols select routing paths according to shortest-path routing algorithm or maximum-bandwidth routing algorithm. MANETs have many characteristics like dynamic topologies, bandwidth-constrained, and energy-constrained [2]. Conventional shortest path routing and maximum-bandwidth routing algorithms are not suitable for routing packets in MANETs. How to select suitable paths for packet transmission according to the QoS requirement of applications becomes an important issue.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 283–292, 2011.

The multi-path routing has been used to achieve load balance and fault tolerance in the networking environments [6, 7]. Multi-path routing protocols in wireless networks provide the fault tolerance capability. When a node obtains information of routing paths, the node can select one of the best routing paths as the main transfer path and the other paths as backup paths. If a link broken or routing error occurs at the main transfer path, the node can select another path as new transfer path to forward packets. Multi-path routing protocols maintain more than one routing path for packet transmissions. As a result, the fault-tolerance and reliability of the MANET can be enhanced [8, 9].

In this paper, we propose a multi-path routing protocol named "DMP" which provides soft QoS support for mobile nodes in wireless ad-hoc networks. In the proposed multi-path routing scheme, nodes collect networking metrics, .e.g., transmission latency, energy status of mobile nodes, and link-bandwidth, to obtain information of routing paths. After the information of routing paths is obtained, each node will adaptively select suitable paths for forwarding packets to the destination according to the QoS requirements of network applications. In route discovery phase, a node-disjoint multi-path discovery mechanism is introduced. By using the multi-path routing scheme, data flows can be distributed to node-disjoint paths for packet transmission. In order to evaluate the performance of the proposed routing protocol, a comparison of the proposed scheme with DSR routing protocol is performed.

The rest of the paper is organized as follows. Session 2 introduces the related works. Session 3 describes the proposed multi-path routing scheme that supports soft QoS for mobile nodes in wireless ad-hoc networks. Session 4 depicts the soft QoS control policy of the proposed scheme. Session 5 shows the experimental result. Session 6 concludes the proposed routing scheme.

2 Related Works

This Section provides a brief overview of Quality of Service (QoS) support, and multi-path routing protocols in MANETs.

In MANETs, the Quality of Service (QoS) [10, 11] support can be classified into Hard QoS and Soft QoS categories [12]. The Hard QoS provides guaranteed services for low latency, low jitter, and assured bandwidth, which forcibly reserves network resources for nodes transmission [13, 14, 15]. In [16], Wen-Tsuen Chen et al. improve DiffServ framework for Expedited Forward class to provide absolute bandwidth guaranteed service. Irina Gerasimov et al. implement QoS support with AODV and TORA by using scheduling bandwidth queue to find QoS paths and provide guaranteed service [17].

In compare with forcibly resource reserved of Hard QoS, Soft QoS does not provide guaranteed resource for delivering data flows [18]. However, Soft QoS supports negotiated resource as much as possible for data flow required instead of forcibly reserve [19]. Jianfeng Wang et al. classify application class by use QoS index, and propose a target-orientated bandwidth allocation scheme [20]. Lei Chen et al. modify Hello packet header of AODV to measure bandwidth usage status and

provide QoS support [21]. Zhenyu Liu et al. propose EARA-QoS which classifies data flow by latency time and gives the highest priority to data flow which asks for low latency time [22]

3 The Adaptive Multi-path Routing Protocol with Soft QoS Support

The proposed routing protocol is based on the existing Dynamic Source Routing protocol (DSR) protocol [3]. In the proposed routing protocol, a multi-path discovering algorithm is introduced to discover suitable routing paths. In the standard DSR protocol, data packets are routed to the destination by the intermediate nodes using the entire path contained in the packet header. If a source node doesn't have routing information of its destination node, the source node will broadcast a Route Request (RREQ) packet to its neighboring nodes for finding the routing path. When an intermediate node receives the RREQ packet, the node checks whether this RREQ packet has been received before or not. If the RREQ packet has not been received before, the node will broadcast the RREQ packet to its neighboring nodes again. If the RREQ packet has been received before, i.e., a RREQ message with the same originator, this RREQ packet will be dropped.

Once the RREQ packet is reached at the destination node, the destination node will construct a Route Reply (RREP) packet and then send the RREP packet back to the source node according to the path information from the header fields of the received RREQ packet. When the RREP message arrives at the intermediate nodes and source node, the intermediate nodes and source node store the path information into its route cache (DSR_Cache). When the source node receives the RREP packet, the source node updates its routing table and starts using the newly found route.

3.1 Routing Request (RREQ)

In the original RREQ processing method of DSR protocol, route discovery is implemented by controlled flooding of the RREQ packet. Every intermediate node only forwards the first received RREQ packet to its neighboring nodes against packet flooding. Once a duplicate originator's RREQ packet is received at the intermediate node, the RREQ packet will be dropped. The method of broadcasting RREQ in DSR greatly decreases the possibility of finding multiple node-disjoint paths, because it limits the diversity of the multiple paths.

In order to discover node-disjoint paths as more as possible, duplicate originator's RREQ packets are be discarded by the intermediate node. However, route information in these duplicate RREQ packets are be cached and recorded if these duplicate RREQ packets contain different travel paths compared with that in DSR cache in the proposed protocol. We modified the method of processing RREQ packets in original DSR protocol, a cache table named DMP_Cache is added to help the processing of duplicate RREQ packets from the same originator that initiates the route discovery. A parameter named is_disjoint is also added into the optional field of RREP packet

header in the proposed protocol. This is_disjoint parameter will be used in the RREP packet processing stage to identify different node-disjoint path. The scope of flooded packets is constrained by a time-to-live (TTL) hop count. The RREQ value is set with a constant (Max TTL) value. During Route Discovery the source node sets a RREQ packet with a maximum TTL value. Each time the RREQ is rebroadcasted by the intermediate nodes this value is decremented. When the TTL reaches 0, the packet is discarded.

```
If (Node == destination_node) {
        // RREQ packet is reached at destination node.
        Add route to DSR_Cache.
        Check the first hop of RREQ with the first hop of routes in cache.
        If (Route path is disjoint)
                is_disjoint = true;
        Else
                is_disjoint = false;
        Return source route to requestor.
}
Else {
        // RREQ packet is reached at intermediate node.
        If (RREQ is first received) {
                Add route to DSR_Cache.
                Append Node_id to path of RREQ.
                Broadcast RREQ to neighboring node.
        }
        If (RREQ has already processed before) {
                If (Node_id is not in route path) {
                        Put route path into DMP_Cache.
                }
        }
           Drop this RREQ.
}
```

Fig. 1. The proposed algorithm for processing RREQ packets

When a source node requires a new route to a destination, the source node starts broadcast RREQ packets. When an intermediate node receives the RREQ packet, the node checks fields of RREQ packet to find who is the originator (source node) of the packet. If the originator was seen before at the intermediate node, i.e., a duplicate originator's RREQ packet but with different travel path, the traversed path information in the RREQ packet will be put into the DMP_Cache. Meanwhile, the duplicate RREQ packet will be drop to avoid flooding storm.

When the destination node receives a RREQ packet, the node checks the first hop node of the path included in this RREQ packet with all paths that stored in the

DSR_Cache. If the first hop node of the path included in this RREQ packet has not been recorded in the DSR_Cache, i.e., a new node-disjoint path, the extended parameter is_disjoint in RREP packet header is set to TRUE. If the first hop node of the path included in this RREQ packet has been recorded in the DSR_Cache, i.e., a recorded node-disjoint path, the extended parameter is_disjoint in RREP packet header is set to FALSE. After the parameter is_disjoint in RREP packet header is set, a RREP packet is sent to the source node. Figure 1 shows the algorithm of processing RREQ packets.

3.2 Routing Reply (RREP)

Once the destination node constructs a RREP packet, a reverse path included in the received RREQ packet is put into the RREP packet. Then the RREP packet is sent back to the source node.

In order to provide soft QoS support for mobile nodes in wireless ad-hoc networks, a metric table named DMP_Metrics is added in the proposed protocol to help the collecting of network metrics of paths. Every routing path that stored in the DMP_Cache has a corresponding network metric record in the metric table DMP_Metrics. The metric table DMP_Metrics stores status of routing paths, including transmission latency, energy status of mobile nodes, and link-bandwidth. In order to provide QoS Support, the RREP packet header is extended with optional data fields that contain information of measured network metrics. While a RREP packet is received at an intermediate node, the node will measure network metrics and update the metric information in the extended RREP packet header.

When an intermediate node receives a RREP packet, it checks is_disjoint field firstly. If the value of is_disjoint field is TRUE, i.e., the routing path included in the RREP packet is a node-disjoint path, the intermediate node forwards the packet to the next node in the routing path. If the value of is_disjoint field is FALSE, i.e., the routing path included in the RREP packet is a non-disjoint path, the intermediate node looks for another node-disjoint path from its DMP_Cache. When a first-hop node of shortest paths stored in DMP_Cache is differed from the first hop node of the routing path in the RREP packet, i.e., a new node-disjoint path from the intermediate node to the source node is founded, the new founded path will be put into the RREP packet to replace the old reversed routing path. Meanwhile, the is_disjoint field of RREP packet is set to TRUE to notify other nodes that a new node-disjoint path is founded. As a result, the proposed protocol can discover different node-disjoint path as many as possible.

When the source node receives a RREP packet, the routing path included in the RREP packet will be stored in the node's route cache which is as the same described in the DSR protocol. Meanwhile, the source node obtains information of network metrics of the routing path from DSR options header included in the packet. The source node stores the information into its metric table DMP_Metrics for providing QoS support.

3.3 Routing Error (RERR)

Because the topology of a MANET may change with time as the nodes move or adjust their transmission range, a routing path may fail. When an intermediate node detects that the next hop of a packet is not reachable, it unicasts a routing error (RERR) packet to the source node. When the source node receives the RERR packet, the source node will check its route cache to find the unreachable nodes. If there has unreachable nodes on the broken link, the source node removes all of unreachable nodes on the broken link and selects a new routing path to transmit data packets to the destination. If there has no path available in the route cache of the source node, the source node reissues a RREQ packet for route discovery again. Meanwhile, the source node will update related information of network metrics of routing paths in its metric table DMP_Metrics when it receives the RERR packet.

3.4 Packet Header Extending

In order to provide multi-path routing and QoS support, the options header in standard DSR protocol is extended by added a 32-bit Metric Timestamp field, a 16-bit Metric Energy field, and a 16-bit Metric Link-Bandwidth field. We also extend the Route Reply option by adding 1-bit is_disjoint field.

4 Soft QoS Support

After the route discovery is finished, the source node's route cache stores one or more node-disjoint routing paths. Every node-disjoint routing path has a corresponding network metric record in the metric table DMP_Metrics. The metric table DMP_Metrics stores status of routing paths, including transmission latency, energy status of mobile nodes, and link-bandwidth. In order to provide Soft QoS support for mobile nodes, the TOS (Type of Service) field of IP header is used to determine the QoS requirement of data flows in the proposed mechanism. The parameters of TOS field are depicted in Table 1. In the original QoS parameters 4th ~ 7th bits in IP header, we extend the 6th bit (Reliability) to represent the energy requirement for finding a routing path with maximum available energy of nodes.

When a node wants to send data to a destination node, the application can set the TOS field in the IP header for depicting its QoS requirement. Then the proposed routing procedure in the system kernel will check the TOS information in the IP packet firstly. After the QoS requirement of the data flow is obtained, the routing procedure will select a suitable routing path from its node-disjoint paths in the route cache to transmit data packets to the destination node.

4.1 QoS Control Policy

Assuming that all mobile nodes support multi-channel transmission, several QoS control policies for mobile nodes are defined in this paper. According to the TOS field, the QoS control policies could be extended as follows.

4th bit (Delay)
When the 4th bit of the TOS field is set, it denotes that the node needs low transmission latency for data delivery. For applications like Voice over IP (VoIP), data packets must arrive at the destination node as soon as possible. When the field is set to 1, the source node will select the path with lowest latency time from DSR_Cache to act as the routing path. The transmission latency of a path in DSR_Cache is obtained from the Route Discovery phase. First, the source node records the timestamp of broadcasting a RREQ packet to discover the destination node. While the source received the corresponding RREP packet, it compares the timestamp of the packet with the timestamp of received time and then records the transmission latency time of the path into DSR_Cache.

Table 1. Type of Service fields

Position	Usage
1st~3rd bits	Precedence
4th bit	Delay
5th bit	Throughput
6th bit	Reliability
7th bit	Cost
8th bit	Reserved

5th bit (Throughput)
When the 5th bit of the TOS field is set, it denotes that the node needs to transfer bulk data to a destination node, e.g., file transfer. When the field is set to 1, the source node will select the top two paths that have maximum Link-Bandwidth (LB) from DSR_Cache to work as the routing paths. The data packets are sent out from source node using multi-path routing with multi-channel transmission support. The value of Link-Bandwidth (LB) can be obtained from the processing of RREP packets. An intermediate node compares the LB value in the received RREP packet with its current link bandwidth, and then rewrites minimum value to LB value metric field of DSR Options Header if a minimum LB value is founded.

6th bit (Reliability)
When the 6th bit of the TOS field is set, it denotes that the node needs to transfer packets to a destination node with lowest energy consumption. Routing paths in MANETs may fail because of nodes run out their energy, which makes the source node need to originate a route discovery request to find new routing paths. When the field is set to 1, the source node selects the path with maximum available energy from DSR_Cache to act as the routing path. The value of maximum available energy of nodes can be obtained from the processing of RREP packets. An intermediate node compares the value of maximum available energy from the received RREP packets

with its available battery energy, and then rewrites minimum value to the energy metric field of the extended DSR Options Header.

7th bit (Cost)
When the 7th bit of the TOS field is set, it denotes that the node needs to transfer packets to a destination node with lowest cost. When the field is set to 1, the node selects a path with the minimum hops from DSR_Cache to act as the routing path.

5 Performance Evaluation

To evaluate the performance of the proposed DMP scheme, we use the NS2 network simulator to simulate the MANET environment. Comparisons with standard DSR protocol in NS2 simulator are performed. There are 50 mobile nodes in a 1500 meter by 500 meter grid in the simulation environment. We generated a scenario file with 15 pairs CBR data flows. Each CBR data flow sends out a 512 bytes packet per 250ms. The generated mobile nodes' mobility patterns are based on the random way-point mobility model with minimum moving speed of 5m/s and maximum moving speed of 10m/s. The random way-point model breaks the movement of a mobile node into motion and stay periods. A mobile node stays at a location for a certain time at first, and then it moves to a new randomly chosen destination at a speed drawn uniformly from a given interval. The parameters in the simulated IEEE802.11b WLAN environment are based on Orinoco 802.11b Card. Each node uses the EnergyModel in ns-2 simulator to simulate the energy consumption and remaining energy of nodes; each node has initial energy range from 10J ~ 60J randomly. In our experiments, the transmitting power txPower is set to 0.5 mW, receiving power rxPower is set to 0.25mW, and idle listening power is set to 0.05 mW. The QoS metric requirement is set to consider reaming energy of nodes firstly, the simulation runs with durations of 100 to 900 seconds.

The simulation results are shown in Table 2 and Table 3. In the tables, field Start Time represents the timestamp of the first packet that is send out, field End Time represents the timestamp of whole network stops to transfer packets. The results show that the DMP scheme can extend the life time of an ad hoc network because of it can select the suitable routing paths for delivering packets according to the QoS requirement. The standard DSR protocol uses the shortest path routing algorithm to transmit data packets. In the DSR experiment, the energies of intermediate nodes are consumed quickly because of DSR does provide QoS support of the energy consumptions of nodes. When nodes ran out the battery life, some routing paths may fail to work. Therefore, the source node needs to originate a route discovery request to find new routing paths in DSR protocol. On the other hand, the DMP scheme considers energy consumption of nodes, which can prolong the network lifetime. The node-disjoint routing paths stored in the DMP_Cache can used as backup paths for packet routing when the original routing path is broken, which can reduce route re-discovery time.

Table 2. Performance results of the proposed DMP scheme

Duration (sec)	100	200	300	400	500	600	700	800	900
Start time(sec)	2.56	2.56	2.56	2.56	2.56	2.56	2.56	2.56	2.56
Remain working nodes	50	38	29	21	19	14	10	3	3
End time(sec)	99.97	199.99	299.95	399.99	499.95	599.95	684.79	684.79	684.79

Table 3. Performance results of the DSR protocol

Duration (sec)	100	200	300	400	500	600	700	800	900
Start time(sec)	2.56	2.56	2.56	2.56	2.56	2.56	2.56	2.56	2.56
Remain working nodes	50	37	27	21	18	13	8	3	3
End time(sec)	99.99	199.92	299.93	399.94	499.95	599.91	655.19	655.19	655.19

6 Conclusion

In this paper, we propose a multi-path routing scheme with soft QoS support for mobile nodes in MANETs. In the proposed scheme, nodes can find out the most suitable routing paths according to its QoS requirement. The experiments show that the proposed scheme can prolong the network lifetime by providing soft QoS support to mobile nodes.

Acknowledgments. This research is supported by the National Science Council of the Republic of China under the grant NSC 100-2221-E-218 -042 -.

References

1. Clausen, T., Jacquet, P.: Optimized Link State Routing Protocol (OLSR). RFC 3626, IETF Network Working Group (2003)
2. Corson, S., Macker, J.: Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations. Internet Informational RFC 2501 (1999)
3. Johnson, D.B., Maltz, D.A., Broch, J.: 'DSR: The dynamic source routing protocol for multi-hop wireless ad hoc networks. In: Perkins, C. (ed.) Ad Hoc Networking, pp. 139–172. Addison-Wesley (2001)
4. Perkins, C.E., Bhagwat, P.: Highly dynamic destination sequenced distance-vector routing (DSDV) for mobile computers. In: SIGCOMM 1994, pp. 234–244 (1994)

5. Perkins, C.E., Belding-Royer, E.M., Das, S.R.: Ad hoc on-demand distance vector (AODV) routing. IETF Internet Draft (2002)
6. Pham, P.P., Perreau, S.: Performance analysis of reactive shortest path and multi-path routing mechanism with load balance. Journal of Telecommunications and Information Technology 2, 38–47 (2003)
7. Tsirigos, A., Haas, Z.J.: Multipath routing in the presence of frequent topological changes. IEEE Communications Magazine 39(11), 132–138 (2001)
8. Chi, L., Hao, Z., Yao, C., Zhang, Y., Wang, K., Sun, Y.: A Simulation and Research of Routing Protocol for Ad hoc Mobile Networks. In: IEEE International Conference on Information Acquisition, pp. 16–21 (2006)
9. Lee, S.-J., Gerla, M., Toh, C.K.: A Simulation Study of Table-Driven and On-Demand Routing Protocols for Mobile Ad Hoc Networks. IEEE Network 13, 48–54 (1999)
10. Bouhouch, H., Guemara, S., Fatmi, E.: QoS Routing In Ad Hoc Networks by Integreting Activity in the OLSR Protocol. In: Second International Conference on Systems and Networks Communications, pp. 1–1 (2007)
11. Haghighat, A.T., Khoshrodi, O.J.: Stable QoS routing by AODV protocol. In: 15th International Conference on Software, Telecommunications and Computer Networks, pp. 1–5 (2007)
12. Chen, L., Heinzelman, W.B.: A Survey of Routing Protocols that Support QoS in Mobile Ad Hoc Networks. IEEE Network 21(6), 30–38 (2007)
13. Hsu, C.-S., Sheu, J.-P., Tung, S.-C.: An On-Demand Bandwidth Reservation QoS Routing Protocol for Mobile Ad Hoc Networks. In: IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, vol. 2, pp. 198–207 (2006)
14. Liao, W.-H., Tseng, Y.-C., Shih, K.-P.: A TDMA-based Bandwidth Reservation Protocol for QoS Routing in a Wireless Mobile Ad Hoc Network. In: IEEE International Conference on Communications, vol. 5, pp. 3186–3190 (2002)
15. Lin, C.R., Liu, J.-S.: QoS Routing in Ad Hoc Wireless Networks. IEEE Journal on Selected Areas in Communications 17(8), 1426–1438 (1999)
16. Chen, W.T., Lin, C.F., Tang, C.S.: A bandwidth reservation protocol for hard QoS guaranteed differentiated services. In: IEEE International Conference on Communications, vol. 6, pp. 1792–1796 (2001)
17. Gerasimov, I., Simon, R.: Performance analysis for ad hoc QoS routing protocols. In: Mobility and Wireless Access Workshop, pp. 87–94 (2002)
18. Ghosh, R., Das, A., Som, P., Venkateswaran, P., Sanyal, S.K., Nandi, R.: A Route Feedback Based Routing Model for Enhanced QoS in Real Time Ad-Hoc Wireless Networks. In: 2006 IEEE Region 10 Conference, pp. 1–4 (2006)
19. Kuo, W.H., Liao, W.: Utility-based Resource Allocation for Soft QoS Traffic in Wireless Networks. In: IEEE International Conference on Communications, pp. 5091–5096 (2006)
20. Wang, J., Cao, P., Yang, X.: Adaptive mobile multimedia QoS control and resource management, Networks. In: 9th IEEE International Conference on Networks, pp. 332–337 (2001)
21. Chen, L., Heinzelman, W.B.: QoS-aware routing based on bandwidth estimation for mobile ad hoc networks. IEEE Journal on Selected Areas in Communications 23(3), 561–572 (2005)
22. Liu, Z., Kwiatkowska, M.Z., Constantinou, C.: A biologically inspired QoS routing algorithm for mobile ad hoc networks. In: 19th International Conference on Advanced Information Networking and Applications, vol. 1, pp. 426–431 (2005)

An Adaptive Query Optimization in a Hierarchical Mediator System

Nam Hun Park[1] and Kil Hong Joo[2]

[1] Dept. of Computer Science, Anyang University, 102 Samsungli, Buleunmyun, Ganghwagun, Incheon, Korea, 417-833
nmhnpark@anyang.ac.kr
[2] Dept. of Computer Education, Gyeongin National University of Education, San 6-8 Seoksudong Manangu Anyangsi, Gyeonggi, Korea, 430-040
khjoo@ginue.ac.kr

Abstract. To hide the heterogeneous environment from a global user as much as possible while preserving the autonomy of each individual system to the highest degree, the study of a mediator has been actively carried out as an integration method for distributed heterogeneous information sources. To evaluate a global query on a view that integrates several distributed information sources, the query can be rewritten into a set of transformed queries based on the definition of the view. Each transformed query is targeted to one of the information sources. To speed up the execution of a global query, the previous results of frequently requested sub-queries are materialized in a mediator. This paper proposes a method of choosing the optimized set of materialized queries in each mediator such that available storage in each mediator can be highly utilized at any time. The integrating schema in a mediator can be incrementally modified and the evaluation frequency of a global query can also be continuously varied. In order to select the optimized set of materialized sub-queries with respect to their current evaluation frequencies, the proposed method for modeling the recent access behavior of each sub-query. As a result, it is possible to adjust the optimized set of materialized sub-queries adaptively according to the recent changes in the evaluation frequencies of sub-queries.

Keywords: mediator, data materialization, distributed query evaluation, heterogeneous information system, data integration, optimized query selection.

1 Introduction

To hide the heterogeneous environment from a global user as much as possible while preserving the autonomy of each individual system to the highest degree, the study of a mediator has been actively carried out as an integration method for distributed heterogeneous information sources [1]. A *mediator* is middleware that can provide a global user with the effect of accessing a single server by concealing the user from the actual heterogeneity of computing environment such as hardware, software, and databases in distributed information servers. To design a distributed database in general, a set of related local schemata are analyzed and integrated to a unique global

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 293–299, 2011.

schema. A mediator is built to handle the specific requirements of a global user, so that only the local information relevant to its function needs to be integrated. When a local schema is changed, only the view mappings of a mediator need to be modified since there is no global schema. Due to this reason, the definition of a view in a mediator can be changed or a new view can be added incrementally.

To evaluate a global query on a view that integrates several distributed information sources, the query can be rewritten into a set of transformed queries based on the definition of the view. Each transformed query is targeted to one of the information sources. In this paper, a term *sub-query* is used to denote a transformed query and each sub-query is a unit of evaluation in its target information server. In this paper, a sub-query evaluated by the modification method is called as a *modified sub-query* while a sub-query evaluated by the materialization method is called as a *materialized sub-query*. When a materialized sub-query is evaluated in a mediator, its materialized result stored in the mediator can be used if the result is valid. Consequently, the materialized method can reduce network traffic and provide the result of a sub-query quickly. However, it is inefficient when relevant base tables in the target information server of a materialized sub-query are updated frequently. Therefore, in order to use the materialization method effectively, it is important to model the evaluation frequency of a sub-query as well as the update frequency of its base tables precisely. These two frequencies are defined as the usage patterns of a sub-query. In order to describe how the sub-queries of a mediator are evaluated, a term '*implementation plan (IP)*' is used in this paper. It is denoted by a set of pairs (q_i, v_i). Each pair represents that a sub-query q_i is evaluated by a method $m_i \in$*{materialization or modification}*. Given a set of all sub-queries Q in a mediator, an implementation plan $IP(Q_{mat}, Q_{mod})$ denotes both the set of materialized sub-queries Q_{mat} and the set of modified sub-queries Q_{mod} $(Q_{mat} \cup Q_{mod} = Q,\ Q_{mat} \cap Q_{mod} = \varnothing)$.

This paper consists of five parts. Section 2 describes related works and Section 3 reviews how to find the optimized set of materialized queries in a hierarchical mediator system. In Section 4, the results of the proposed method are comparatively analyzed by a series of experiments to identify its various characteristics. Finally, Section 5 draws overall conclusions.

2 Related Work

View materialization is proven to be useful in data warehouse systems [2, 3]. A data warehouse schema itself can be regarded as a set of materialized views for operational databases or existing files [4,5]. Therefore, many researches have been concentrated on a view selection problem. In addition, a *supporting view* can be used to avoid the repetitive evaluation of a frequent query over base tables in a warehouse system [6]. In [2], a materialized view selection algorithm based on the data cube of a lattice structure is proposed.

HERMES [1], DISCO [7], TSIMMIS [8] and SIMS [9] are well-known mediator systems to integrate the heterogeneous databases in distributed environment. In these

systems, whenever a new information source is added to a mediator, a mediator manager defines the mapping between the mediator schema and the data domain of the new source. In addition, these systems provide a catalog of all accessible information sources and map a multi-source global query into a set of local queries that are evaluated by the wrappers of related information sources. While HERMES, DISCO and TSIMMIS use only the query modification method to evaluate a global query, SIMS [9] employs the materialization method additionally. An information source in SIMS is assumed to contain read-only data. Accordingly, it does not consider the update frequency of an information source. Based on the evaluation frequency of each user query, SIMS identifies frequently queried patterns whose target data elements are possible candidates for materialization. Given the evaluation frequencies of user queries, the CM (Cluster and Merge) algorithm is proposed to extract such patterns from user queries. Since SIMS only deal with read-only data sources, the priority of materialization among these patterns is determined by the evaluation frequency of each pattern. Therefore, a query with the highest priority is chosen one by one to be materialized as a query until available storage space is exhausted. As users' preferences may change over time, materialization need to be considered dynamically. [3] proposes a dynamic view management system with improved architecture. Data mining methods and probabilistic reasoning approaches are employed to predict incoming queries to select and materialize views.

3 Optimized Implementation Plan

The current usage patterns of a sub-query q_i indicate its recent rates of evaluation and update operations requested for the sub-query. A modified sub-query q_i is evaluated in its target information server as many as its evaluation frequency value, ef_i times recently. A materialized sub-query q_i is evaluated in its information server as many as its update frequency value, uf_i times recently. The evaluation cost of a modified/materialized sub-query q_i can be defined by Equation (1) and (2).

$$c_{mod}(q_i)=ef_i \times size(q_i) \tag{1}$$

$$c_{mat}(q_i)=uf_i \times size(q_i) \tag{2}$$

where $size(q_i)$ denotes the latest result size of the sub-query q_i by the amount of network transmission. The difference of these two evaluation costs for a sub-query is a possible gain when the sub-query is evaluated by the materialization method. As the difference becomes larger, the sub-query is more likely to be a materialized sub-query. Formally, the gain $gain(q_i)$ of a sub-query q_i is defined by the difference as follows:

$$gain(q_i)= c_{mod}(q_i)- c_{mat}(q_i) \tag{3}$$

Since the total gain $GA(Q)$ of all sub-queries Q can be defined by Equation (4), the total gain of a specific IP is equal to the total gain of its materialized sub-queries Q_{mat} as in Equation (5).

$$GA(Q) = \sum_{i=1}^{n} gain(q_i) \tag{4}$$

$$Gain(IP(Q_{mat}, Q_{mod})) = GA(Q_{mat}) \tag{5}$$

Once a mediator is in operation, it is not a good idea to suspend its normal operation for a long period whenever the current implementation plan is not efficient any longer. To generate an optimized plan in linear time complexity, a *essential implementation plan* can be considered as $EIP(Q_{mat}, Q_{mod})$. The implementation plan is obtained by the ordered list of sub-queries $Q=\{q_1, q_2,\ldots,q_i,\ldots,q_n\}$.

Given a set of sub-queries $Q_{set}^{k}=\{q_1^{k},q_2^{k},\ldots,q_l^{k}\}$ in a mediator m_k of a hierarchical mediator system, a term *extended practical implementation plan* $ext\_EIP_k(Q_{mat/mod}^{*})$ is used to denote the optimized implementation plan $IP(Q_{mat/mod})$ of a mediator m_k where $Q_{mat}^{*} \cup Q_{mod}^{*}=Q_{set}^{k}$ and $Q_{mat}^{*} \cup Q_{mod}^{*}= \varnothing$. It is generated in linear time complexity by the same method as in a single mediator system presented in Section 3. In order to generate the *ext_EIP* for a mediator m_k, all the sub-queries in Q_{set}^{k} are arranged by their normalized gain in decreasing order.

For the $ext\_EIP_k(Q_{mat/mod}^{*})$ of a mediator m_k in a hierarchical mediator system, its *materialized cost* $c_k(Q_{mat}^{*})$ is defined by the sum of the evaluation costs of all materialized sub-queries as defined in Equation (6). Likewise, its *modification cost* $c_k(Q_{mod}^{*})$ is defined by the sum of the evaluation costs of all modified sub-queries as defined in Equation (7). Accordingly, the *total evaluation cost* of the $ext\_EIP_k(Q_{mat/mod}^{*})$ is defined by the sum of these two costs as in Equation (8).

$$c_k(Q_{mat}^{*}) = \sum_{q \in Q_{mat}^{*}} c(q) \tag{6}$$

$$c_k(Q_{mod}^{*}) = \sum_{q \in Q_{mod}^{*}} c(q) \tag{7}$$

$$total\_evaluation\_cost_k = c_k(Q_{mat}^{*}) + c_k(Q_{mod}^{*}) \tag{8}$$

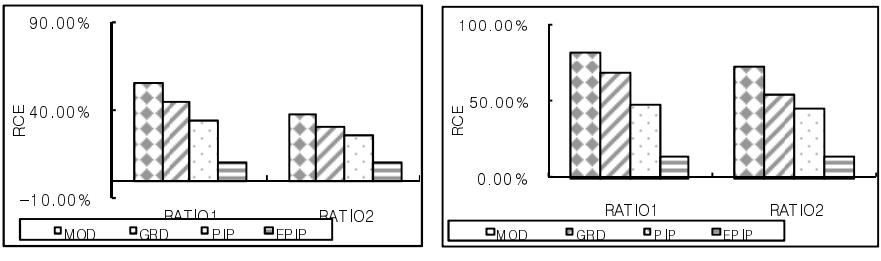
4 Experiments

The first log is denoted by RATIO1 that contains only evaluation operations without any update operation. The second log is denoted by RATIO2 whose ratio between evaluation and update operations is 50:50 respectively. Each log contains a sequence of about 3000 evaluation or update operations each of which is applied to a randomly chosen sub-query. Furthermore, the average number of sub-queries in each mediator is set to 100 and the initial result size of each sub-query is randomly chosen from 10MB to 200MB. The updated result size of each sub-query is also randomly chosen from the 10 percent of its previous size to 10 times of its previous size for each update operation. The operations in each log are uniformly distributed. The number of transformed sub-queries for a query imposed in a mediator is denoted by the fan-out of the query.

The performance of the globally optimized plan is compared with those of a general greedy (*GRD*) optimization method, the modification-only method (*MOD*) and the locally optimized method (*EIP*). A term *relative cost efficiency* (RCE) is introduced to measure the efficiency of each optimization method relatively to the true optimum plan. It is the ratio of the *total evaluation cost* (TC) of an *IP* generated by each optimization method over the *true optimum cost* (OC) of the true optimum *IP*.

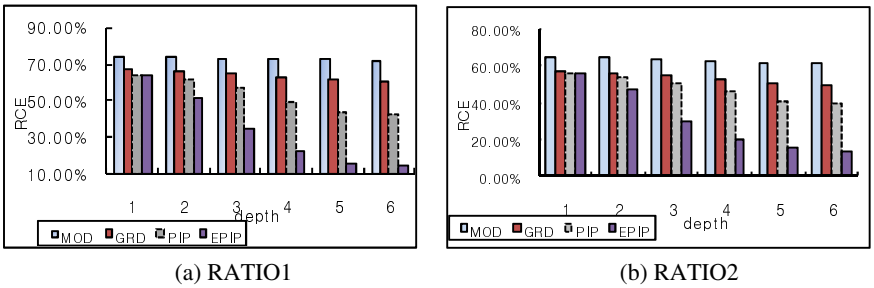
$$RCE = \left(\frac{TC - OC}{OC} \right) \times 100 \quad (\%)$$

In Figure 1, the performance of the proposed *ext_EIP* is compared with those of other optimization methods. The fan-outs of all queries are fixed to 4. As shown in this figure, the total evaluation costs of the proposed *ext_EIP* are the closest to the true optimum cost for the two types of request logs. As expected, the performance of the proposed *ext_EIP* on the RATIO1 log is better than that on the RATIO2 log. As the update ratio of a sub-query becomes smaller, the performance of the proposed *ext_EIP* becomes more efficient.



(a) depth : 3 (b) depth : 9

Fig. 1. *RCE* for different usage patterns



(a) RATIO1 (b) RATIO2

Fig. 2. *RCE* for the depth of a hierarchical mediator system

Figure 2 shows the effect of the depth of a hierarchical mediator system on the total evaluation cost. The fan-outs of all queries are fixed to 4 as well. The depth is varied from 1 to 6 in order to illustrate its effect. If the depth of a hierarchical mediator system is 1, all the sub-queries of each mediator are terminal sub-queries. As

the depth of a hierarchical mediator system becomes larger, the performance of the *ext_EIP* is more enhanced. This is because the number of intermediate sub-queries in each mediator of a hierarchical mediator system and the amount of network transmission in the modification method are increased proportionally to the depth.

5 Conclusion

In a hierarchical mediator system, the performance of query evaluation can be greatly enhanced when the materialization method is utilized to evaluate sub-queries in each mediator. This paper proposes a method of choosing the optimized set of materialized queries in each mediator of a hierarchical mediator system such that available storage in each mediator can be highly utilized at any time. By carefully monitoring the usage patterns of sub-queries requested in each mediator of a hierarchical mediator system, the proposed algorithm can determine when a new optimization process in each mediator should be invoked in order to change its current *IP*. As the number of sub-queries and the number of mediators are increased, the proposed algorithm can provide an efficient way of evaluating sub-queries in a hierarchical mediator system. Since the set of materialized sub-queries selected by the proposed algorithm is chosen based on the relative utilization efficiency of available storage space in each mediator of a hierarchical mediator system, it may not be optimum but practically efficient.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science Technology(2011-0025300).

References

1. Subrahmanian, V.S., Adali, S., Brink, A., Emery, R., Lu, J.J., Rajput, A., Rogers, T.J., Ross, R., Ward, C.: HERMES: A Heterogeneous Reasoning and Mediator System, http://www.cs.umd.edu/projects/hermes/overview/paper
2. Harinarayan, V., Rajaraman, A., Ulman, J.: Implementing data cubes efficiently. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Montreal, Quebec, Canada, pp. 205–216. ACM Press, New York (1996)
3. Daneshpour, N., Barfourosh, A.A.: Dynamic View Management System for Query Prediction to View materialization. International Journal of Data Warehousing and Mining 7(2), 67–96 (2011)
4. Cuzzocrea, A., Gunopulos, D.: Efficiently Computing and Querying Multidimensional OLAP Data Cubes over Probabilistic Relational Data. In: Catania, B., Ivanović, M., Thalheim, B. (eds.) ADBIS 2010. LNCS, vol. 6295, pp. 132–148. Springer, Heidelberg (2010)
5. Chaudburi, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. SIGMOD Record 26(1), 65–74 (1997)
6. Pardillo, J., Mazon, J.-N., Trujillo, J.: Extending OCL for OLAP querying on conceptual multidimensional models of data warehouses. Information Sciences 180(5), 584–601 (2010)

7. Kapitskaia, O., Tomasic, A., Valduriez, P.: Scaling Heterogeneous Databases and the Design of Disco, INRIA Technical Report, France (1997)
8. Hammer, J., Garcia-Molina, H., Nestorov, S., Yerneni, R., Breuning, M., Vassalos, V.: Template-based Wrappers in the TSIMMIS System. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, pp. 532–535. ACM Press, New York (1997)
9. Ashish, N., Knoblock, C.A., Shahabi, C.: Selectively Materializing Data in Mediators by Analyzing User Queries. International Journal of Cooperative Information Systems 11(1,2), 119–144 (2002)

On Implementation of a KVM IaaS with Monitoring System on Cloud Environments*

Chao-Tung Yang**, Bo-Han Chen, and Wei-Sheng Chen

Department of Computer Science, Tunghai University, Taichung 40704, Taiwan
ctyang@thu.edu.tw, adplus1986@gmail.com, adam.cws@gamil.com

Abstract. Cloud computing has been a popular topic in recent years. How the system used in large-scale cloud computing to build large-scale distributed storage cluster and the provision of services become very popular and provoke animated discussion. It is an inevitable trend to either the enterprise or the individual. The cloud of network services can be divided into three categories: Software as a Service (SaaS), Platform as a Service (PaaS), and infrastructure as a service (IaaS). Among these, the Software as a Service (SaaS) allows users to run applications remotely from the cloud. Platform as a Service (PaaS) includes the operating system and custom application-specific software. And Infrastructure as a Service (IaaS) is the computing resources as a service, including some virtual machines, hardware resource units. This research paper is about how to build the KVM environment in the cloud system and operation, and using the KVM to provide a virtual environment for users. In the user interface part, the work is able to reduce the complexity of cloud resources access. We use the web interface which is easy to understand and accessible for users in the operations as well. About the experimental results, the performance of physical machine and KVM virtual machine are compared and analyzed.

Keywords: Cloud computing, Virtualization, KVM, Virtual machines, IaaS.

1 Introduction

Cloud computing is currently a popular topic, but also all the main axis of development in recent years, the main points of infrastructure as a service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), cloud computing is not a new technology, it is a new concept [1, 10, 19, 23, 21, 22, 24, 28, 30]. The early stages of the laboratory started in the creation and development of gird computing cluster and other distributed computing technologies and related issues, for the vigorous development in recent years is also very interested in cloud computing [13, 15].

* This work is supported in part by the National Science Council, Taiwan R.O.C., under grants no. NSC 100-2218-E-029-001 and NSC 100-2218-E-029-004.
** Corresponding author.

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 300–309, 2011.

There are many companies currently offer a cloud of related services, like Google [30], Amazon [29], Yahoo! other companies, tens of thousands of servers used to construct a large-scale computing resources, and provides a variety of services previously not available such as: large storage space, a huge amount of computing power, no need to download the online features such as edit view, for their own local computing and storage resources are limited, users can access via the Internet computing resources they need.

This paper focused on the cloud computing infrastructure, particularly virtual machines and physical monitoring component [3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 18, 33, 34]. Goal is to achieve a system can provide users apply and use the virtual machine, and can monitor the physical system. The information can be monitored include CPU utilization, disk usage, virtual machine space, memory usage. This system also uses a mechanism for Migration, when a problem occurs, the administrator can shift the user's virtual machine to another physical machine operation, and the user will not feel any abnormalities. Meanwhile, this paper also carried on the system performance test using the KVM [2, 20, 26, 27].

In section 2, we describe the techniques used and some background knowledge. Section 3 describes the system architecture and key algorithms which this paper were used. Section 4 conducts some experiments for our proposed system. In final, section 5 states the conclusions and future work.

2 Virtualization

Virtualization technology is due to present a single host more and more powerful hardware performance, if only a single server implementation of the tasks seem too much idle time, so multiple hosts by the hardware virtualization technology, the original value Line by more than one virtual host, after the service, placed on a single powerful server is running, but also makes virtualization virtual machine after the machine easier to control than real checks and controls, more flexible configuration and can be anywhere in the world And can achieve real-time transfer of virtual machines to ensure uninterrupted service.

KVM (Kernel-based Virtual Machine) is a virtualization solution for Linux on x86 hardware containing virtualization extensions Technology (Intel VT or AMD-V) [2, 20, 26, 27]. It consists of a loadable kernel module “kvm.ko”, which provides the core virtualization infrastructure and a specific processor module, supports KVM intel.ko module or KVM amd.ko module. KVM on a machine can run multiple virtual machines. Each virtual machine has its own virtualized hardware, such as: network card, disk, video card etc.

Kernel-based Virtual Machine (KVM) is a Linux core, a part of the framework, the current structure of native virtualization support KVM hardware-assisted virtualization is supported by the CPU, Intel virtualization technology called VT or AMD's AMD-V Technology in Linux through the two CPU module to support two different KVM (Intel: kvm-intel.ko; AMD: kvm-amd.ko). KVM supports for the Para-virtualization, such as he now supports Linux and Windows, Para-virtual

network device drivers, and the balloon (on the memory technology VMM-virtual memory manager) has done for Linux Guest's CPU optimization.

3 Related Work

In recent years, performance improvement management process technology advances make it possible to try to use the virtual machine (VM) computing platform. Many studies have been implemented through the virtual network environment, reduce system costs. Data transmission between server nodes often appear in parallel and distributed computing systems, high cost of the network may cause significant loss of performance throughout the system.

Amazon EC2 (Elastic Compute Cloud) is a virtual machine allows users to perform the required lease operating system. EC2 way through the web service so that users can implement their own virtual machine at any time, users will be able to run any virtual machine that you want the software or application.

Binbin Zhang put forward for the KVM in the study of I/O optimization, how to simplify the client operating system by eliminating redundant operations in a virtual environment. Simplification of the client operating system will be an important direction for future research to optimize the performance of the VMM. They are studying how to make the guest operating system more effectively in a virtual environment, and how to make the guest operating system, VMM and the host operating system to better complement each other.

The user can always create, execute, terminate its own virtual machine, how much count how many times, and therefore the system is “flexible” use. This paper focuses VMs running on physical machines and use KVM technology to Implementation a virtualization environment for user to application and use.

4 System Implementation

As an Infrastructure as a Service (IaaS) provider, this paper applies an ideas of virtualizes in the cloud system to economize power, web interface and user friendly to manage the virtual machines. Therefore, there are some distinct on framework of cloud; our system architecture is shown in Figure 1. About user friendly, users simply connect to the site through the Internet, and then set their own needs, you can create a virtual machine, the user does not need to know what happened back may need to set any object, they can be consistent with their own Needs of virtual machines. The user diagram is shown on Figure 2.

According to the previous plan, this paper have established physical machine using KVM virtual machine system and provides a Web interface to manage the virtual machine. Our virtual system was built up with four homogeneous computers; the hardware of these computers is equipped with Intel Xeon CPU E5410 2.33 GHz, eight gigabytes memory, 500 gigabytes disk, Ubuntu 10.04 operating system, and the network connected to a gigabit switch.

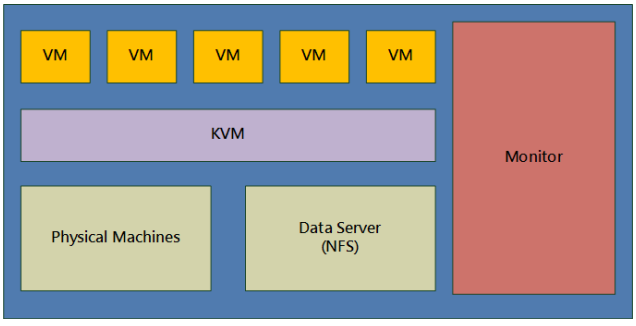


Fig. 1. Virtual Machine System Architecture

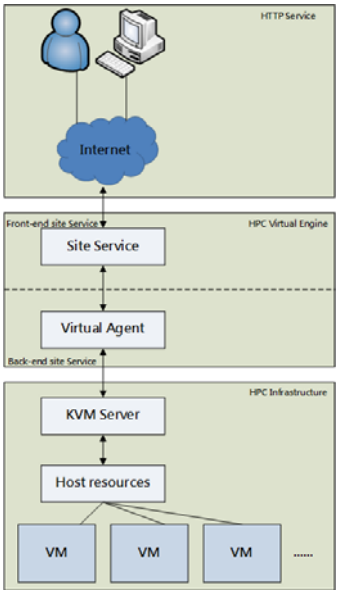


Fig. 2. The user diagram

5 Experimental Results

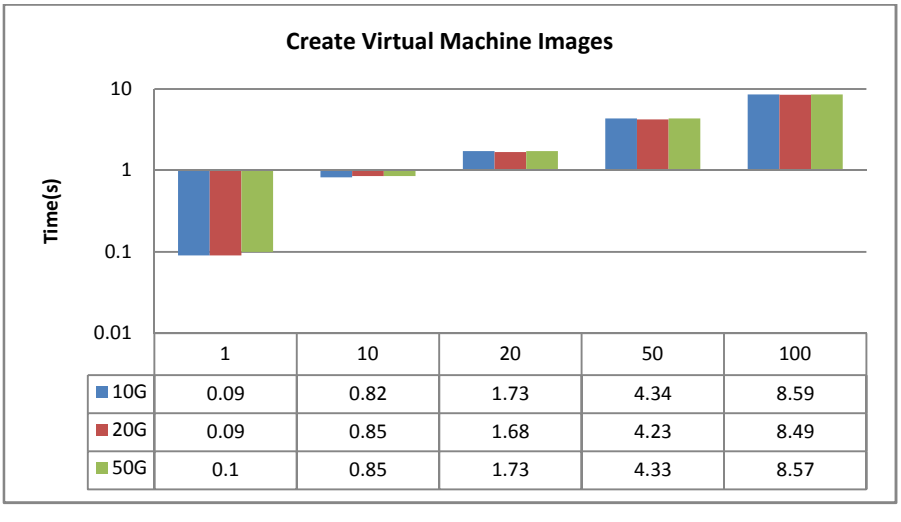
This paper focus on the efficiency of KVM virtual machines, include test the virtual machine build time, start time, migration time and computing performance while migration. The experiment programs are listed as below:

- **VM Build Time** - In order to enable users to obtain the application as quickly as the virtual machine, so this paper tested the speed of the virtual machine created in order to understand how their efficiency.
- **Matrix Multiplication** - Matrix multiplication is an efficient algorithm, commonly used in performance calculations, and seeking the path, is an application of strong algorithms. It is through the operation of multiplying a

matrix with either a scalar or another matrix by changing the size of the matrix of data can be different.

- **Jacobi** – An iterative method. In numerical linear algebra, Jacobi method is an algorithm to determine the solution of a system of linear equations of the absolute maximum in each row and column main diagonal elements [32]. Each diagonal element is solved, and a plug in the process of approximation, and then iterate until convergence.
- **HPCC** - HPCC (HPC Challenge Benchmark) [31] is a performance evaluation program, a number of different indices for the test can be used as high-performance computing capacity and determination of indicators. Content of the test include HPL, STREAM, Random-access, PTRANS, FFTE, DGEMM and b_eff Latency / Bandwidth, seven test items 34.
- **Computing Performance While Migration** - When the system is migrated and the user is using the resource calculation, the calculated result is not any migration and the result is the same.

This experiment, this paper tested the KVM image file in the creation of the efficiency, mainly on the number and size of construction compared as shown in Figure 3, when the construction is to quickly create, display image files in the establishment of KVM efficiency is good.

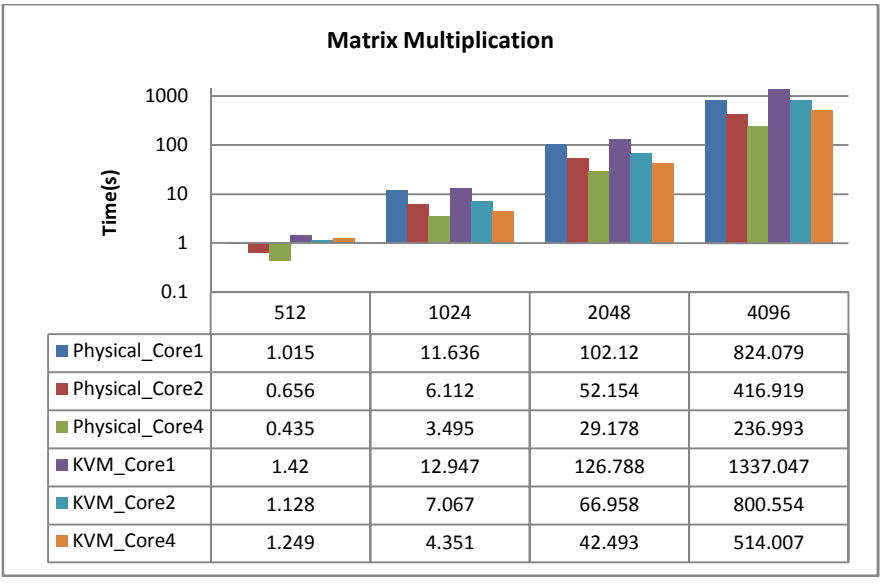


	1	10	20	50	100
10G	0.09	0.82	1.73	4.34	8.59
20G	0.09	0.85	1.68	4.23	8.49
50G	0.1	0.85	1.73	4.33	8.57

Fig. 3. Create virtual machine images test

We use Matrix Multiplication to show the KVM and Physical computing performance. About the efficiency, this paper tested the size of matrix 512,1024,2048,4096 in the Figure 4. This paper see the KVM, whether in single, dual and quad core in performance is not outstanding, the gap is about 1 to 1.5 times the computing speed. This paper knows that KVM is not prominent in the CPU support, so this paper tests another experimental.

In Figure 5, we use a Jacobi linear equation to record the KVM and physical machines performance. By changing the size of the equation, compare the effectiveness of change, can be seen from the execution time using Jacobi operator, the physical machine and KVM only a little difference. Instructions in the use of KVM CPU computing performance, the performance will be as affected by the type of calculation.

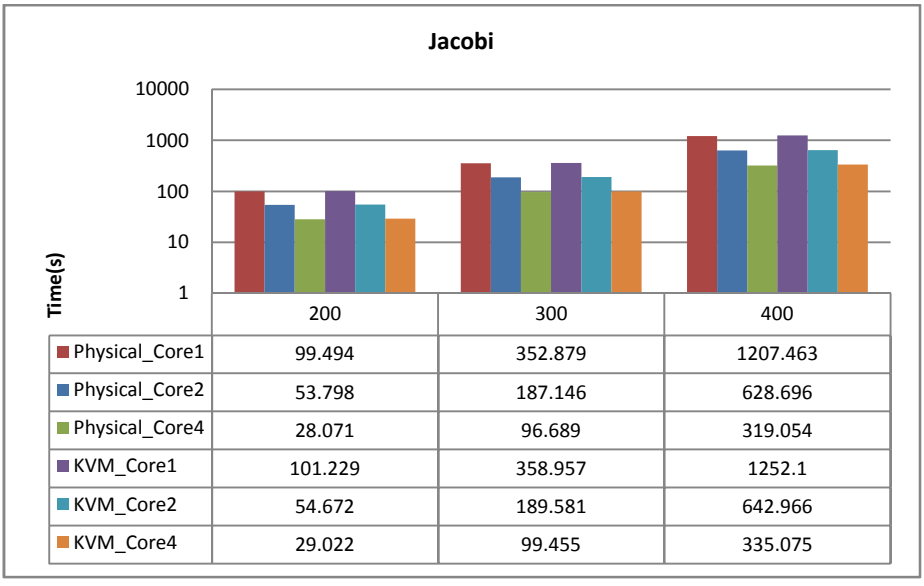


	512	1024	2048	4096
Physical_Core1	1.015	11.636	102.12	824.079
Physical_Core2	0.656	6.112	52.154	416.919
Physical_Core4	0.435	3.495	29.178	236.993
KVM_Core1	1.42	12.947	126.788	1337.047
KVM_Core2	1.128	7.067	66.958	800.554
KVM_Core4	1.249	4.351	42.493	514.007

Fig. 4. Matrix Multiplication results

In the case, we use HPCC [31] to measure the KVM and physical machine the overall performance. HPCC tests include HPL, STREAM, Random-access, PTRANS, FFTE, DGEMM and b_eff Latency/Bandwidth. In Figure 6, this paper can see that when the computation is not high, the physical machine's overall performance is better than KVM, but when computing the amount of increase, the KVM's performance with almost the same physical machine. Although the computing performance shown in KVM needs to be strengthened, but the overall performance, the place still has its expectations. KVM from the execution time can be found in the large amount of computing time, the efficiency has been with almost the same physical machine, or even slightly beyond. This is very surprising.

In the last experiment, if user using computing resources, and then the machine starts migration. Experiments using performance measurement program HPCC. Start migration when the HPCC program running. In Figure 7, the case will measure the results and the previous data of HPCC for comparison, this paper can see that when the Migration is in progress, significantly reducing the efficiency, but is maintaining a certain performance.



	200	300	400
Physical_Core1	99.494	352.879	1207.463
Physical_Core2	53.798	187.146	628.696
Physical_Core4	28.071	96.689	319.054
KVM_Core1	101.229	358.957	1252.1
KVM_Core2	54.672	189.581	642.966
KVM_Core4	29.022	99.455	335.075

Fig. 5. Jacobi results

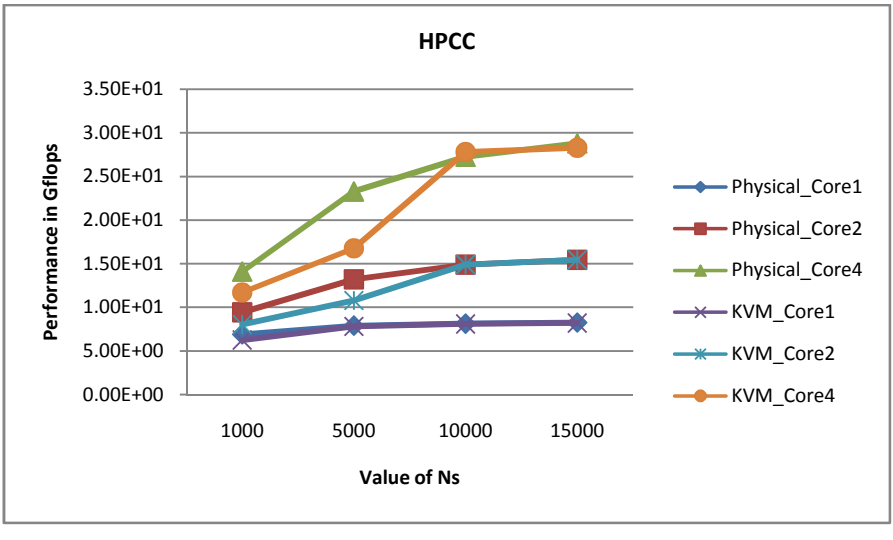


Fig. 6. HPCC of performance results

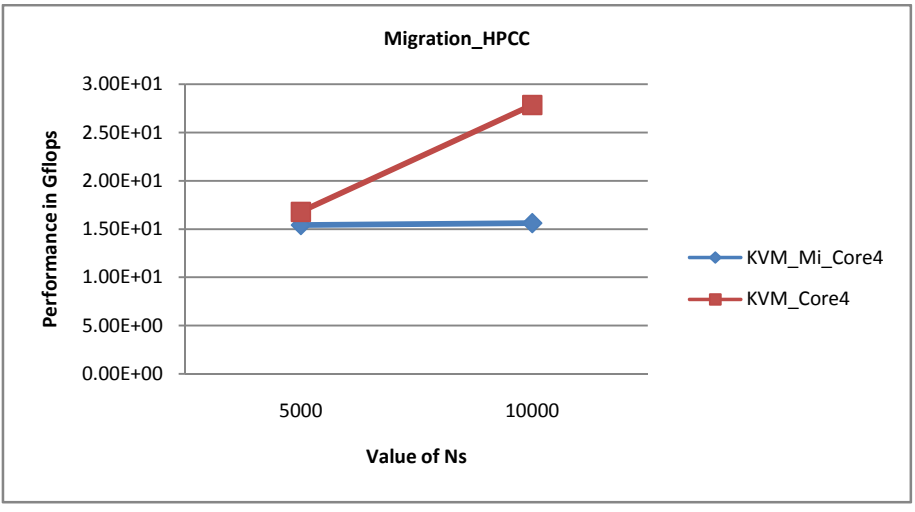


Fig. 7. Migration performance with HPCC test

6 Conclusions

This paper implemented a cloud of KVM infrastructure and monitoring website, which offers users to apply for the use and monitoring of VM state, and the main page with easy to understand the type, the user in the application and monitoring, can Obtained through the needs of the most simple steps in order to user friendly. Unlike the past, the usage of Xen as the virtualization technology, this paper tries to use KVM virtualization technology as a major. In addition, this paper also tested a number of VM creations and implementation of the efficiency of the test, although there is still a gap from the best performance, but the final results were very satisfactory.

References

1. Nagarajan, A.B., Mueller, F., Engelmann, C., Scott, S.L.: Proactive fault tolerance for HPC with Xen virtualization. In: Proceedings of the 21st Annual International Conference on Supercomputing, Seattle, Washington, June 17-21, pp. 23–32 (2007)
2. Zhang, B., Wang, X., Lai, R., Yang, L., Wang, Z., Luo, Y., Li, X.: Evaluating and Optimizing I/O Virtualization in Kernel-based Virtual Machine (KVM). In: NPC 2010 Proceedings of the 2010 IFIP International Conference on Network and Parallel Computing, September 13-15, pp. 220–231 (2010)
3. Matthews, C., Coady, Y.: Virtualized Recomposition: Cloudy or Clear? In: ICSE Workshop on Software Engineering Challenges of Cloud Computing, May 23, pp. 38–44 (2009)

4. Tseng, C.-H., Yang, C.-T., Chou, K.-Y., Tsaur, S.-C.: Design and Implementation of a Virtualized Cluster Computing Environment on Xen. In: Presented at the The second International Conference on High Performance Computing and Applications, HPCA (2009)
5. Waldspurger, C.A.: Memory Resource Management in VMware ESX Server. SIGOPS Oper. Rev. 36(SI), 181–194 (2002)
6. Bellard, F.: Qemu, a fast and portable dynamic translator. In: Proceedings of the USENIX 2005 Annual Technical Conference, FREENIX Track, pp. 41–41 (2005)
7. Kecskemeti, G., Terstyanszky, G., Kacsuk, P., Nemetha, Z.: An approach for virtual appliance distribution for service deployment. Future Generation Computer Systems 27(3), 280–289 (2011)
8. Raj, H., Schwan, K.: High Performance and Scalable I/O Virtualization via Self-Virtualized Devices. In: The Proceedings of HPDC 2007, pp. 179–188 (2007)
9. Van Nguyen, H., Tran, F.D., Menaud, J.-M.: Autonomic virtual resource management for service hosting platforms. In: ICSE Workshop on Software Engineering Challenges of Cloud Computing, May 23, pp. 1–8 (2009)
10. Oi, H., Nakajima, F.: Performance Analysis of Large Receive Offload in a Xen Virtualized System. In: Proceedings of 2009 International Conference on Computer Engineering and Technology (ICCET 2009), Singapore, vol. 1, pp. 475–480 (January 2009)
11. Smith, J.E., Nair, R.: The Architecture of Virtual Machines. Computer 38(5), 32–38 (2005)
12. Paul Willmann, J.S., Carr, D., Menon, A., Rixner, S., Cox, A.L., Zwaenepoel, W.: Concurrent Direct Network Access for Virtual Machine Monitors. In: The Second International Conference on High Performance Computing and Applications, HPCA, pp. 306–317 (2007)
13. Kertesz, A., Kacsuk, P.: Grid Interoperability Solutions in Grid Resource Management. Systems Journal 3(1), 131–141 (2009)
14. Adams, K., Agesen, O.: A Comparison of Software and Hardware Techniques for x86 Virtualization. In: ASPLOS-XII: Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems, pp. 2–13. ACM Press, New York (2006)
15. Rodero-Merino, L., Vaquero, L.M., Gil, V., Galán, F., Fontán, J., Montero, R.S., Llorente, I.M.: From infrastructure delivery to service management in clouds. Future Generation Computer Systems 26(8), 1226–1240 (2010)
16. Milojičić, D., Llorente, I.M., Montero, R.S.: OpenNebula: A Cloud Management Tool. IEEE Internet Computing 15(2), 11–14 (2011)
17. Luszczek, P., et al.: Introduction to the HPC Challenge Benchmark Suite. LBNL-57493 (2005)
18. Endo, P.T., Gonçalves, G.E., Kelner, J., Sadok, D.: A Survey on Open-source Cloud Computing Solutions. In: VIII Workshop Em Clouds, Grids e Aplicações, pp. 3–16 (2011)
19. Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A.: Xen and the Art of Virtualization. In: SOSP 2003: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, pp. 164–177. ACM Press, New York (2003)
20. Qumranet, White Paper: KVM Kernel-based Virtualization Driver. Qumranet, Tech. Rep (2006)
21. Borja Sotomayor, R.S.M., Llorente, I.M., Foster, I.: Virtual Infrastructure Management in Private and Hybrid Clouds. IEEE Internet Computing 13, 16–23 (2009)

22. Soltesz, S., Potzl, H., Fiuczynski, M.E., Bavier, A., Peterson, L.: Container-based Operating System Virtualization: A Scalable, High-performance Alternative to Hypervisors. In: EuroSys 2007, pp. 275–287 (2007)
23. Hagen, W.V.: Professional Xen Virtualization. Wrox Press Ltd., Birmingham (2008)
24. Emeneker, W., Stanzione, D.: HPC Cluster Readiness of Xen and User Mode Linux. In: 2006 IEEE International Conference on Cluster Computing, pp. 1–8 (2006)
25. Li, Y., Yang, Y., Ma, N., Zhou, L.: A hybrid load balancing strategy of sequential tasks for grid computing environments. Future Generation Computer Systems, 819–828 (2009)
26. Zhang, X., Dong, Y.: Optimizing Xen VMM Based on Intel Virtualization Technology. In: 2008 International Conference on Internet Computing in Science and Engineering (ICICSE 2008), pp. 367–374 (2008)
27. Dong, Y., Li, S., Mallick, A., Nakajima, J., Tian, K., Xu, X., Yang, F., Yu, W.: Extending Xen with Intel Virtualization Technology. Journal, ISSN, Core Software Division, Intel Corporation, 1–14, August 10 (2006)
28. Hai, Z., et al.: An Approach to Optimized Resource Scheduling Algorithm for Open-Source Cloud Systems. In: 2010 Fifth Annual, ChinaGrid Conference (ChinaGrid), pp. 124–129 (2010)
29. Amazon, http://aws.amazon.com/ec2/
30. Cloud computing, http://en.wikipedia.org/wiki/Cloud_computing
31. HPCC, http://icl.cs.utk.edu/hpcc/
32. Jacobi, http://en.wikipedia.org/wiki/Jacobi_method
33. KVM, http://www.linux-kvm.org/page/Main_Page
34. Xen, http://www.xen.org/

RFID and Supply Chain Management: Generic and Military Applications

Tae Hwan Oh[1], Young B. Choi[2], and Rajath Chouta[1]

[1] Department of Networking,
Security and Systems Administration,
Golisano College of Computing and Information Sciences,
Rochester Institute of Technology,
52 Lomb Memorial Drive,
Rochester, NY 14623
{tom.oh,rsc5726}@rit.edu
[2] Information Systems Technology
Department of Natural Science, Mathematics & Technology
School of Undergraduate Studies, Robertson Hall 464
Regent University
Virginia Beach, Virginia 23464-9800
ychoi@regent.edu

1 Introduction to RFID

This paper explains two main aspects of RFID supply chain management – the generic supply chain management and the specific military application for the purpose of asset viability. The generic supply chain management is studied by examining how to improve supply chain management through accurately determining inventory levels in real time without human interaction. For the military application, the current legacy systems used by the military for Supply Chain Management and Asset Visibility are analyzed and suggestions are made for improvements in the form of Total Asset Visibility through Radio Frequency Identification tags, and Geographic Information Systems.

1.1 What Is RFID?

"RFID technology is classified as a wireless automatic identification and data capture technology" [2]. This technology is composed of three parts: an identification tag, a reader and a computer. The identification tag can take one of three forms, active passive, semi-passive; though the two major forms are active and passive.

"Active RFID allows extremely low-level RF signals to be received by the tag (since the reader/interrogator does not power the tag), and the tag (powered by its internal source) can generate high-level signals back to the reader/interrogator. Active RFID tags are continuously powered, and are normally used when a longer tag read distance is desired.'

T.-h. Kim et al. (Eds.): FGCN 2011, Part I, CCIS 265, pp. 310–323, 2011.

'Passive RFID tags reflect energy from the reader/interrogator or receive and temporarily store a small amount of energy from the reader/interrogator signal in order to generate the tag response. Passive RFID requires strong RF signals from the reader/interrogator, and the RF signal strength returned from the tag is constrained to very low levels by the limited energy" [2].

Semi Passive tags are similar to active tags, but the battery is used to run the microchip's circuitry but not to broadcast a signal to the reader. Some of these tags conserve battery life by sleeping until they are woken up by the reader's signal. Semi-Passive tags may also be referred to as battery-assisted tags[24].

Active tags are more expensive than passive tags because they send information using an internal battery source and store more data, while passive tags rely on the reader. Also "tags have a discrete memory capacity that varies from a small license plate to thousands of records. Data within a tag can provide any level of identification for an item during manufacture, in-transit, in storage, or in use" [4].

"In addition to tags, an RFID system requires a means for reading or "interrogating" the tags to obtain the stored data, and then a way of communicating this tag" [4]. Attached to readers are antennas which transmit data read from a passive or active tag to the computer. This information can be used to track the asset through the supply chain.

Using RFID technology "could allow supply chain members to automate manual tasks, reduce human errors, and improve the traceability and availability of items (products, boxes, pallets, etc.), generating savings for all the supply chain members" [2]. "This can be done because RFID is a totally non-intrusive methodology for data capture" [4]. A non-intrusive methodology is a process that does not require human intervention. This is done by using automatic readers which gather information from the RFID tag attached to the item. It is also a "non-line-of-sight technology, and may possess both read and write options within the same equipment item" [4]. Radio waves are used to make this process capable of non-line-of-sight technology, which means the tags can be read even if they are not visible to the reader.

RFID tags have also proven to be incredibly durable. They have been subjected to physical, environmental and real world tests. The physical tests included stacking, dropping, and vibration. Even after being subjected to these rugged tests an overwhelming majority of the tags were still functionally operational. Tests concerning variations and extremes in temperature and humidity were used during the environmental testing. It was found that some degradation to the read rates did occur when tested at the extremes. It is also interesting to note that using "sensor enhanced RFID devices enable them to monitor their physical context (awareness) such as temperature or moisture" [9]. Real world tests included damage to a few tags such as scratches and depressions. These defects did not have a significant reduction in tag performance [3].

The government requires that in the coming years the military will need to have total asset visibility. To achieve this goal RFID technology will need to be used, along with effective databases.

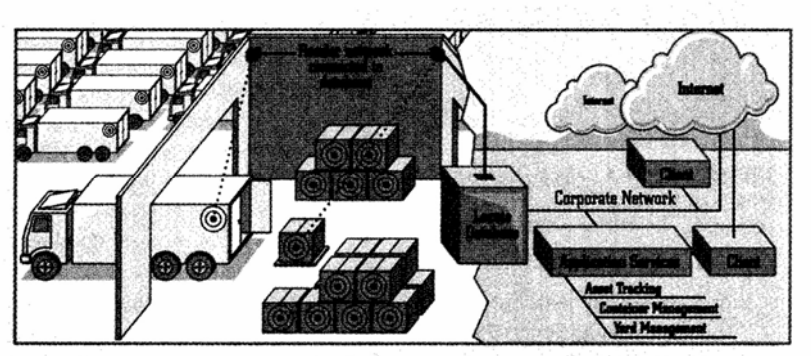


Fig. 1. A potential setup of the distribution and network constituting a RFID SCM system [11]

1.2 What Is Supply Chain Management?

Supply chain management (SCM) has always been one of the most essential elements in business. In the past, the focus has been more so on improving the distribution process and refining inventory models instead of technological components. However, it still lacks great efficiency. That being said, how can we improve supply chain management using technology? While it will never be perfect, utilizing advancements in technology, such as RFID can drastically improve the system. Systems that implement automated processes to collect, maintain, organize, and analyze information in real-time empower businesses to exponentially increase efficiency, resulting in increased profits. By looking at the current problems and limitations facing supply chain management, this paper examines RFID's potential in the supply chain and provides solutions to some of these problems.

2 History of Supply Chain Management

2.1 Supply Chain Management

Supply chain management (SCM) is a broadened management focus that considers the combined impact of all the companies involved in the production of goods and services, from suppliers to manufacturers to wholesalers to retailers to final consumers and beyond to disposal and recycling. (Helms, 2006)

The goal of SCM is to seamlessly link all of the activities responsible for bringing goods and services to the market and efficiently manage the flow from start to finish. Basically, no matter how many companies are involved, this process should take place as though there were only one, very well managed company performing the activities. All companies in the supply chain can benefit from sharing information and working closely with the other companies that are within their supply chain.

2.2 Supply Chain Management in the Military

Currently, it is evident that the military has not adopted a very efficient Supply Chain Management (SCM) system. Previous errors in package handling, frivolous spending on legacy systems and dirty data that is used in decision making are all reasons that the military needs to adopt a more efficient SCM system that has as its vision is of Total Asset Visibility (TAV). Also, "With the end of the cold war, there has been a shift in paradigm from a large heavy force to a light, reactionary mobile force" [5]. This effectively means that the equipment needs to be as responsive and mobile as the soldiers. The equipment is only as good as the soldiers that use it, and the soldiers are only as good if they have the proper equipment to do their jobs. To achieve the goal of a mobile logistic system good asset management should be used on the supply chain. This would include managing assets individually, allowing locating the right assets, providing information about the current physical status (quality) of an asset, and to keep an information history of an asset. [9]

The major need for an integrated supply chain management network arose from the problems the soldiers in the field experienced during the first Gulf War. United States Air Force Gen. Walter Kross explains concisely that "during the Gulf War, we simply did not have good information on anything. We did not have good tracking; we had no real asset visibility. Materiel would enter the logistics pipeline based on murky requirements, and then it could not really be tracked in the system.... We lacked the necessary priority flows to understand where and when things were moving. It was all done on the fly, on a daily basis... It truly was brute force...We had too much, and, worse yet, we did not know what was where" [10].

These problems led to wasted equipment and ill-equipped soldiers. It was also not cost efficient; because of the way the military had set up its shipping and receiving, many of the pallets sent over to the theater went to waste. The Department of Defense's (DOD) logistics strategy was scattered and running on out of date technology. "During the 1991 Gulf War, the DOD's biggest logistic nightmare involved huge shipping containers sitting unopened in the desert, full of unidentified equipment and supplies. Consequently, additional equipment and supplies had to be shipped, resulting in costly duplication and waste" [11]. When a shipment was received its arrival was not put into a database or any form of computer system; instead carbon copies were used. This method further increased the number of duplicated shipments and also the confusion of what exactly was in each container. It became clear that the military needed to create an organized supply chain in which not only would the data be easily accessible but that the equipment would also be able to be quickly located and allocated; the military needed total asset visibility (TAV).

"That is, they need to have actionable information available to them at all times about the location, quantity and state of their material assets and personnel" [3]. This system would help to solve General Kross' problem of not knowing what was what or where it was. Eric Wagner believes that this can be done with radio frequency identification (RFID) tags. It [RFID technology] integrates the digital and the physical world by seamlessly connecting objects in the physical world with their representations in information systems [9]. Wagner says that by using these tags "we

are tracking [the soldiers] movements, their actions and their interactions with each other" [12]. If these tags can be used to accurately collect and store data on the complexities of human movement it can also be applied to the objects in the supply chain. This new technology has the power to completely restructure the militaries supply chain in a way that is cost effective and reliable.

2.3 Problems and Limitations within the Supply Chain System

SCM affects all aspects of business. Accounting, inventory control, marketing and ultimately a company's revenues are all intimately linked to and dependent upon efficient management of their supply chain. Supply chains are systems with a flow of goods from supplier to consumer, made up of many facilities where communication and intensive coordination are a necessity. As with any system, one malfunctioning node can have major effects on all congruent and subsequent actions along the chain. This being said, supply chain management is perhaps the one area of business where creativity and flexibility are most necessary. Improved processes in a company's supply chain can yield a plethora of cost cutting opportunities. Digital technologies and growing trend toward globalization are influencing supply chain management in very profound ways.

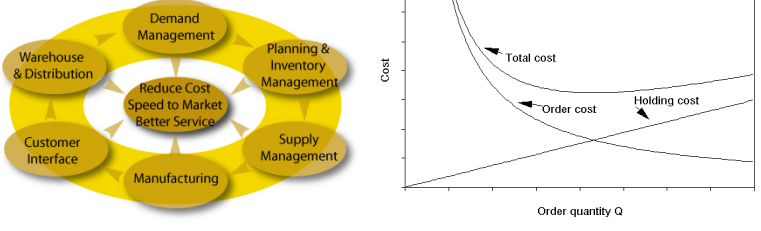


Fig. 2. & 3. The integrated operating activities of a typical supply chain (left) and Cost V Order Quantity *(Image obtained from: josephdamiano.com/Articles/JIT.htm, www.pwc.com/extweb/service.nsf/docid/E4D3338C2AC0BB7C80256E6D0048CE0B)*

Inefficient supply chains can hinder businesses in a number of ways. Faulty distribution strategies can put the right product in the wrong place or the wrong amount of product in places where they do not sell well. Companies lose revenue from missed sales opportunities and likely have to pay to correct the malfunction. Inefficient management of cash flow between departments can disrupt accounting processes and affect the bottom line. Inventory control has been one of the most visible problems facing supply chain management. Inventory management affects the way companies handle their manufacturing overhead. Companies lose money by storing inventory items that are not selling. Whether to rely on mathematic equations, computer inventory records or a mixture of the two to determine the optimal inventory level is an issue facing companies today. Recently there has been a shift in

the way businesses think about their inventory needs. Businesses are moving away from the industrial age model of "Iron Mountains of just-in-case inventory" into the 21st century information age model of "demand driven… just-in- time inventory" [26]. Efficient supply chain processes that abide by just-in-time inventory principles can greatly reduce their inventory holding costs. By not buying and or producing, and ultimately not warehousing products that are not in high demand and not selling companies can save money.

While Just-in-time inventory has the potential to solve many of the inventory problems facing current companies supply chains, just-in-time inventory practices are not without their limitations. The principles of just-in-time inventory may seem obvious; companies only hold as much as they can sell. Actually implementing just-in-time principles and achieving them in practice can prove to be extremely difficult. Unforeseen fluctuations in customer demand can create a situation where stock outs force customers to shop elsewhere, costing the company in lost revenue and more importantly; customer good will. Controlling inventory in real-time should be the goal of any supply chain.

Traditionally, companies relied on large advertising departments to develop and maintain marketing plans to provide potential customers with information about their products. These processes are exhaustive and require extensive financial and human resources to wage an effective campaign to attract customers. The emerging demands caused by globalization, and changing consumer trends require a more efficient system for gathering, maintaining and utilizing the information to effectively market products to intended customers. Product information Management or PIM is a emerging as a solution to these issues. PIM is the processes and technologies corporations use to gather and manage information about products with the aim of improving its marketing and increasing its overall selling power.

Changing of traditional business processes and experimentation with new technologies is the key to survival in today's competitive business environment. New technologies are providing answers to questions that most never thought to ask. The danger many businesses face is becoming comfortable with processes they have implemented, today's business culture is fast paced and ever changing. Alexander Drobik and Jeff Woods in their article *Development of Chaos-Tolerant Processes Is Key to Supply Chain Optimization* say "the next generation of supply chain excellence will not come from lean-influenced or demand-driven-only planning". This comes only decades removed from the renaissance of just-in-time centric thinking. Due to these types of realizations businesses are increasingly embracing the idea of Chaos Tolerant processes. The theory behind this thinking is that "the elimination of complexity and uncertainty solely through fail-safe process design is futile." Many argue that implementing rigid policies of supply chain principles "across an unstable, performance-based network" cannot provide "coherent management span of control" and doing so can lead to diminishing returns. Designing "fault-tolerant rather than fail-safe processes" [20] will become increasingly necessary in the future.

Businesses, managers and executives must be flexible and open to embracing new ideas in order to remain relevant in the ever changing technology driven global market place. Information is power; therefore it follows that Information Technology

is power. Nearly all business problems can be solved by proper manipulation and implementation of information. Alexander Drobik notes; “supply chain operations [today] continue to accept a level of inefficiency… to overcome this, further redesign of the supply chain and leverage of new enabling IT is needed” [20]. New technologies are always evolving tackle problems with the hopes of increasing productivity, lowering waste and increasing profit margins. Radio Frequency Identification is a new technology whose applications are limited only by one’s imagination. RFID will change the way companies conduct business and revolutionize all aspects of supply chain management systems.

3 Solutions

3.1 Hardware

In order to successfully implement RFID applications in the supply chain, there will need to be several components installed both within items and in the store.The EPC itself will be embedded into the RFID tag in order to be transmitted. To successfully translate the tag, a tag reader must be used. The tag reader must be able to handle readings both fast and effectively, allowing for greater increase in speed through the supply chain.

3.2 Software

Enterprise software must be created that performs a number of major tasks supporting the network of readers, namely, data smoothing, reader coordination, data forwarding, data storage, and task management. The software must be efficient in that it can accurately read and record product information at high speeds and can remove duplicates in real time. In addition, the software must realize what information can be terminated at a certain point and what information must be forwarded up and down the supply chain. The resulting databases must be designed to store mass amounts of information as well as the ability to quickly retrieve information of a specific item. Along with enterprise software, companies will need to utilize an object name service, or ONS. An ONS is similar to a DNS in that it will be used to associate bits of information together. The ONS must take the EPC from the enterprise software and quickly find the location of the detailed information of the product associated with the code.

3.3 Language

In addition to the necessary hardware and software arises the need for a new programming language called physical markup language, or PML, which is derived from XML. “PML is intended as the global standard to be used across industries for

describing physical objects, processes, and environments using a hierarchical basis of taxonomy" [17].

4 Achieving TAV

A combination of RFID technology and a well maintained supply chain management system can result in TAV system. It is important to the military to be able to track equipment, vehicles, ammunition, soldiers, and even military civilians in and out of theaters (war-zone). The information will be stored in a database environment, and allow access to military personnel to query the information for logistical purposes. The query will return all the relevant information that was held on the RFID. Things such as the Item Code, Status, Location, and other information will be stored on the tag, and then transferred to the database. "It is expected that RFID will provide better inventory management and control. This can also translate into better released support for the troops in the battlefield" [8i]. When the military is able to closely track these items operations will run more smoothly due to the fact of less variability and more reliable information of the localities of equipment and personal necessary for the war effort. Having the power to know where all supplies, and personnel are in a hazardous and fast paced environment allows planners to make good well informed decisions, which leads to a more powerful and responsive military.

"Moreover, when coupled with a wireless network, an RFID system allows access to continuous real-time information on "smart items," anytime, anywhere in the supply chain, thereby enabling end-to-end supply chain visibility" [2]. "In addition, if there is a crucial requirement to locate a specific item at any of the nodes, the logistics operator can query the TAV system, [database], to find all the locations where the item is current[ly] located" [10]. Since active tags are more expensive than passive ones the Department of Defense has decided to use passive tags for a high-volume rollout [16]. This means that a majority of the equipment that the military uses will already have the RFID tags on them, allowing a TAV system to be easily established.

While a TAV system is a step in the right direction, it is not exactly what is needed for the military to have complete visibility of their assets. All the information will be obtained but it will be stored into a database. Only trained soldiers will be able to access the data and even if the data is retrieved it takes a long time processing all the information retrieved. The next ultimate maneuver is to integrate the TAV system to a Geographic Information System (GIS). This adds a visual perspective of the data that a TAV system cannot provide. A GIS system can be used to show the location of assets using global positioning system (GPS) technology, as well as show the terrain that the military has to encounter. It is important to integrate RFID with a "geolocation sensor since a RFID tag cannot provide x-y coordinates, but can only indicate its location within the field of its reader, or provide environmental observations" [8].

When an RFID tag is joined with a geolocation sensor “they will supply simple identifying information, including location coordinates. When queried, the RFID tag will return such location information as latitude and longitude, effectively serving as a digital survey stake” [6].

5 Improvements

While the above section discusses in detail about how the solutions can be incorporated with RFID, there is still room for achieving more. Although the solutions by themselves are sufficient to make big changes to supply chain management, it is possible to push the boundaries and get more in the process. In the sections below, we will look at some of the ways in which improvements can be included in order to ensure that the RFID solution will be an effective one that is capable of expansion and can provide more than just basic supply chain management.

6 Introduction to GIS

“Internet-based geographic information systems (GISs) are tremendous tools in providing constituent access to government information. The ability of a GIS to aggregate and easily display layered themes of data on maps can greatly enhance constituent access and streamline operations by avoiding requests for printed map production. GIS’s can, and are, playing a supportive role in improving government's homeland security posture” [1]. GISs can be used for:

1. Visually locating and mapping critical infrastructure
2. Help locating public-safety facilities based on population and proximity to at-risk critical infrastructure
3. Identifying the best evacuation routes based on population, traffic patterns and road capacities
4. Routing emergency vehicles efficiently in real time, using Global Positioning System and Automatic Vehicle Location
5. Supporting field workers who collect data and report status in real-time (for example, reporting on the spread of fire or performing triage in multiple affected areas to help prioritize rescue and recovery operations) [1]

Military application of the above uses:

1. Allows for a more detailed perspective of the battlefield that can be disseminated to the troops so that they can be better prepared for possible conflict.
2. This could be used so that the military will always know where safe spots and potential medic tents can be set up.
3. With the increasing danger of roadside bombs and RPG attacks finding an appropriate convoy route can be created.

4. If a "hot spot" (a place where conflict has broken out) has occurred then the other vehicles can be routed to this location to provide support, or if the vehicle is not capable of fighting, it can be immediately re-routed around this conflict zone. "Spatially oriented information systems can improve road safety engineering by more accurately representing roads and intersections, as well as by providing more in-depth analysis of the factors that contribute to traffic accidents" [15].
5. This point will help the infantry soldier, if they are in trouble, the logistical information officer will be able to tell on the map, and send help to the location requested.

All this information is provided on the map, which could be a lot of data to take in at one time, but because it is done visually the information is able to be seen and understood in less time than if it were done with text. "Information visualization techniques amplify understanding by increasing human mental resources, reducing search times, improving recognition of patterns, increasing inference making, and increasing monitoring scope" [13].

This GIS is a great tool for military strategic planners. It gives them the information in a form that is easy to understand and process. "The inclusion of these types of technologies provide more accurate, up-to-date situational awareness and help make both training and tactical missions successful while improving force protection" [5]

6.1 Expanding TAV to Include GIS

To improve on the GIS capability would be to include the usage of RFID tags with the TAV system. Since the RFID tags would be set up with specific information to store such as their unique code, location, and timestamp, the GIS map system can be set up to include time as a third coordinate. Outdated mapping technology is only able to map in 2D, this dimension being space called "tracks, which were used by the Air Force and Navy command and control systems to show on a map surface the trails of moving entities" [13].

This only allows information to be plotted on x, y coordinates. This means that looking at the data the user does not know how long a piece of equipment or personnel have been at the same location, or where they are supposed to be going. The maps if you would are just snap shots in time and do not incorporate the past or present into the picture. "Time can extend above the map in the third dimension proving time does not need to be a "hidden dimension" [13].

There are two major ways that this could be done. First, from the point where the object is located a line can go either up to indicate in the future, or below the graph to indicate in the past. Second, from the objects location on the x, y coordinate a line will go upwards to a time scale, where the scale will determine whether it will be in the past, present, or future. Examples of these two ways of mapping time onto the third dimension are showed below.

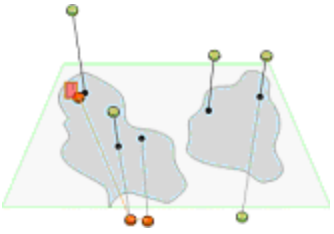


Fig. 4. Line extension and passage through terrain to indicate time. [13]

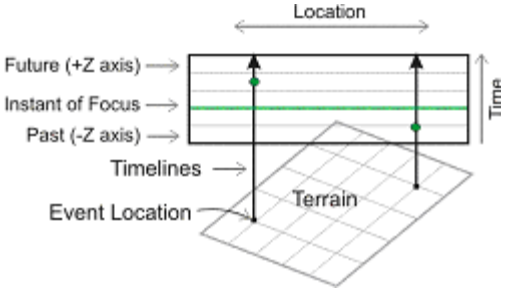


Fig. 5. Time-Chart timelines depicting terrain location connections on the map. [13]

The GIS now gives the military command and control the ability to have complete TAV. With the advent of including time to maps creates the potential for well informed decisions, "Not only can they see their units, status and movements, they can track live fire, soldiers, vehicles, other entities and assets in real time" [5]. In addition, they will be able to track all of this important information. This information will allow the user to be able to see the past and prospective future for the movements of the tracked items.

The good thing that the map offers is the visualization that a command and control officer wants to process information quickly when it is needed, but the GIS allows for even more detail through the use of drill-down options. "Mouse over drill down, allows additional information, such as text or images, to be displayed in the visualization" [14]. The GIS system will be in sync with the database that holds all the RFID tags information, and because "the memory of smart assets can be used to store history information that can be requested later" [9] when the option to drill-down on a specific object on the map the GIS will run a query to populate information in a pop-up window. Information will be presented to the user in a textual format. Status, Unique Code, Name, Location, and Timestamp are all possibilities of what data will be returned to the user.

6.2 Improvement to JIT

The main problem with JIT inventory is that stock levels are determined by historic demand and do not account for sudden spikes in demand resulting in unhappy customers. The ability of RFID enabled systems to maintain an accurate count of

inventory both on hand and in the warehouse can resolve this issue. As demand rises the system can automatically order new stock resulting in increase profits for the firm. In addition to the increased profits, the firm will see an increase in customer morale as the products they wish to buy will always be available.

6.3 Improvement to PIM

PIM's biggest goal is to place more detailed product information out to customers and promote the development of e-commerce. Currently, in order to get information into a PIM system, there must be constant processing of products and data resulting in high labor cost. RFID will be able to streamline the process and provide the PIM with extensive information. Increasing product information will give more power to the consumers and will drive sales. Since the process will be automated and the tags will contain all the information, labor costs of PIM is decrease again resulting in increased profit.

7 Benefits

The benefits from implementing RFID in supply chain management are limitless. The main benefit is the improvement in inventory tracking from when the product is manufactured all the way until the product leaves the store. RFID technology provides companies the ability to control inventory levels at precise levels and automates the process. This will save the time, cost, and energy of workers who will no longer need to physically scan and log every item. An additional benefit is the ability to manage returns and recalls effectively. Products can be tracked to exactly where they came from which allows items that are inferior or dangerous to be quickly removed from the supply chain. Other benefits include: increased speed throughout the supply chain, enhanced inventory throughput, improved inventory management, reduced labor costs, reduced shrinkage, advance shipping notices, and anti-theft and aid in recall and returns.

8 Conclusion

Upon examining the problems associated with supply chain management and also indentifying some of the solutions RFID technology can offer, it is clear that the integration of RFID technology into SCM will have a positive impact. While the technology is still relatively new and the initial investment associated with the transition to RFID based systems is substantial, it is clear that the benefits of RFID technology far outweigh the costs in the long run. As more companies adopt the usage of this technology, its cost will continue to decline and will lead toward standardization within the industry – much like the barcode system. Companies who use this technology will see several positive improvements, such as: reduced inventory shrinkage, increased efficiency maintaining inventory levels leading to reduced holding costs, decreased labor costs associated with inventory management, increased

speed and information sharing throughout the supply chain, and increased proficiency in recalls and returns. These are just examples of some of the tangible benefits that can be seen right now. In time, numerous intangible benefits will be realized.

If the military wants to stay ahead of the swords, it needs to adopt a RFID embedded SCM system. With this system in place the military will be able to easily track all of their assets and be able to manage the data in an organized fashion. This fusion of an RFID and SCM system would also lead to the opportunities to incorporate a GIS mapping function that would allow the military's command and control officers to enjoy the benefits of easy to understand information in real time. This visual representation of data speeds cognitive understanding and will also be useful in emergency situations as well as day to day work. Together the two technologies, RFID and GIS, will offer the military total asset visibility. Not only would these benefits improve asset visibility, the system would also end up saving the military large sums of money that could be used across the board for various programs.

The military needs to update its legacy systems. Its current systems are dangerous not only the military's budget but also to the soldiers in the field who are being denied expedient delivery of equipment. Conversely it is also dangerous to the Military's inventory in that soldiers or military personnel can easily walk off with an item of equipment and have it not be accounted missing for days or even weeks. The research indicates that the most efficient and effective way of updating the legacy systems would be to implement an RFID embedded SCM system that integrates GIS mapping technology.

References

[1] Ahlawat, R., Allan, A., et al.: Hype Cycle for Government. Garner, 1–54, July 3 (2006)
[2] Bendavid, Y., Lefebvre, L.A., Wamba, S.F.: Proof of concept of an RFID-enabled supply chain in a B2B e-commerce environment. In: ICEC, pp. 564–568 (August 2006)
[3] Buckner, M.: MICLOG RFID Tag Program Enables Total Asset Visibility, vol. 2, pp. 1422–1426. IEEE (2002)
[4] Campbell, A., Das, A., et al.: RFID Security, Syngress (2006)
[5] Copley, R., Wagner, E.: Improved Situation Awareness through GIS and RFID in Military Exercises. ESRI, http://gis2.esri.com/library/userconf/proc06/papers/papers/pap_2350.pdf (retrieved March 27, 2007)
[6] Erle, S., Gibson, R., Walsh, J.: Mapping Hacks. O'Reilly (2005)
[7] Kreizman, G.: GIS: Public Access for and Against Homeland Security. Gartner, 1–4, July 30 (2002)
[8] Lahiri, S.: RFID SourceBook. IBM Press (2005)
[9] Lampe, M., Strassner, M.: The Potential of RFID for Moveable Asset Management. RFID Journal (September 2007)
[10] Manish, B., Moradpour, S.: RFID Field Guide: Deploying Radio Frequency Identification Systems. Prentice Hall (2005)
[11] Navas, D.: DoD Logistics New Rules of Engagement. Industry Insights, 28–32 (August 2001)
[12] O'Conner, M.C.: U.S. Army Uses UWB to Track Trainees. RFID Journal (2005), http://www1.rfidjournal.com/article/view/1987/ (retrieved April 01, 2007)

[13] Shuping, D., Wright, W.: Geo-Temporal Visualization of RFID Providing Global Visibility of the DoD Supply Chain (Part 1 of 2). Directions Magazine (2005), http://www.directionsmag.com/article.php?article_id=951&trv=1 (retrieved April 02, 2007)
[14] Shuping, D., Wright, W, Geo-Temporal Visualization of RFID Providing Global Visibility of the DoD Supply Chain (Part 2 of 2). Directions Magazine (2005), http://www.directionsmag.com/article.php?article_id=1954 (retrieved April 02, 2007)
[15] Vining, J.: Case Study: Gwinnet County, Georgia, Uses GIS to Reduce Traffic Collisions. Gartner, 1–8, August 30 (2006)
[16] Woods, J.: DoD Will Nurture Transofrmative RFID to Help Warfighters. Gartner, 1–5, January 8 (2004)
[17] Angeles, R.: Rfid Technologies: Supply-Chain Applications and Implementation Issues. Information Systems Management 22(1), 51–65 (2005)
[18] Blosch, M., Sampler, J.L.: Executive Summary: Strategy as Synchronization: Leading in Real-Time. Gartner, ID Number: G-10-9114, pp. 1-14 (2002) (retrieved April 4, 2008) from Gartner Database
[19] Davison, J., et al.: Transforming Hype Cycle for Supply Chain Management and Procurement. Gartner, ID Number: G001449865, pp. 1-6 (2007) (retrieved April 4, 2008) from Gartner Database
[20] Drobik, A., Woods, J.: Development of Chaos-Tolerant Processes Is Key to Supply Chain Optimization. Gartner, ID Number: G00138252, pp. 1-8 (2006) (retrieved April 4, 2008) from Gartner Database
[21] Helms, M.M., Inman, R.A.: Supply Chain Management. In: Helms, M.M. (ed.) Encyclopedia of Management, 5th edn., pp. 870–873. Gale, Detroit (2006), from Gale Virtual Reference Library via Gale, http://go.galegroup.com/ps/start.do?p=GVRL&u=viva_jmu (retrieved April 10, 2008)
[22] Kiranoudis, C.T., et al.: Dynamic Modeling and Control of Supply Chain Systems: A Review. Computers &Operations Research 35(10), 4–9 (2007) (retrieved April 4, 2008) from Business Source Premier (EBSCOhost) Database
[23] Radio Frequency Identification. Encyclopedia of Emerging Industries, 5th edn. Gale, Detroit, from Gale Virtual Reference Library via Gale, http://go.galegroup.com/ps/start.do?p=GVRL&u=viva_jmu (retrieved April 10, 2008)
[24] RFID Journal (n.d.). Glossary of Terms, http://www.rfidjournal.com/article/glossary/3 (retrieved April 10, 2008)
[25] Simpson, R.: Glossary of Mobile and Wireless Terminology. Garner, 30 (2007)
[26] Strauss, H.: Transforming Defense Supply Chains: What Every Defense CIO Should Know About Defense Supply Chain Management. Gartner, ID Number: G00145941, pp. 1-7 (2007) (retrieved April 4, 2008) from Gartner Database
[27] Technology Development Centre (2006), RFID/ePC for Supply Chain Management, http://www.rp.sg/tdc/scm/RFIDePC%20for%20Supply%20Chain.gif (retrieved April 9, 2008)
[28] White, A.: Frequently Asked Questions: Product Information Management, Business Applications and Business Intelligence. Gartner, ID Number: G00144152, pp. 4 (2006) (retrieved March 15, 2008) from Gartner Database

Author Index