Computer Science and Information Systems

Computer Science and Information Systems

Published by ComSIS Consortium

Special Issue on Advances on Mobile Collaborative Systems

Volume 10, Number 2
April 2013

Vol 10, No 2, April 2013

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia
Phone: +381 21 458 888; **Fax:** +381 21 6350 458
www.comsis.org; Email: comsis@uns.ac.rs

Volume 10, Number 2, 2013
Novi Sad

Computer Science and Information Systems

Special Issue on Advances on Mobile Collaborative Systems

ISSN: 1820-0214

ComSIS Journal is sponsored by:

Ministry of Education, Science and Technological Development of Republic of Serbia - http://www.mpn.gov.rs/

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. ComSIS also welcomes survey papers that contribute to the understanding of emerging and important fields of computer science. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2011 two-year impact factor 0.625,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Summon (Serials Solutions),
- EBSCO bibliographic databases,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (MS Word or LaTeX), or three hard-copies of the manuscript written in English, intended for publication and prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be sent to official Journal e-mail.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 25 (using the appropriate Word or LaTeX template).

Manuscripts will be refereed in the manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

Computer Science and Information Systems

Volume 10, Number 2, Special Issue, April 2013

CONTENTS

EDITORIAL

Mobile collaborative systems, which allow collaboration through wireless networks and mobile devices, have influenced and changed the quality of our lives over the past decade. Meanwhile, the explosive growth of data traffic for user services threatens the current mobile systems. Especially, mobility models, architectures, and application services have posed various challenges to those in academia and industry. In particular, the key challenges for improving efficiency, scalability, and reliability are the development of mobile collaborative systems and the measurement of precise performance of mobile collaborative systems. These challenges allow us to design and develop new models, architectures, and services for future mobile systems.

This special issue covers the following main topics:

- Mobile Internet architectures for collaborative systems
- Mobility models and performance evaluation
- Collaborative technologies for fast creation and deployment of new mobile services
- Collaborative location aware mobile systems
- Handoff, mobile networks and wireless web
- Mobile learning and groupware systems
- Mobile and ubiquitous applications exploiting semantics
- Context-aware environments for work and enterprises
- Synchronization algorithms for mobile collaborative systems
- Mobile social networking
- Evaluation of the usability of mobile collaborative applications
- Mobile technology and system for real time collaboration in enterprises.

These subjects, as well as some others, are the focus of this special issue of "Advances on Mobile Collaborative Systems". The special issue is organized as follows:

The first paper by Fei Song, Huachun Zhou, Sidong Zhang, Hongke Zhang, and Ilsun You, presents a throughput critical condition study for reliable multipath transport. Specific analytical mechanisms are proposed to analyze the potential problems which may lead to serious performance decrease. They investigate how to use multiple paths legitimately when network environments are fluctuating. In their simulation, the results have revealed some throughput critical conditions and could be helpful in designing scheduling schemes for multipath protocols.

The second paper by Chin-Ling Chen, Yu-Ting Tsai, Aniello Castiglione, and Francesco Palmieri, proposes a dynamic location-aware key management scheme based on the bivariate polynomial key pre-distribution, where the aggregation cluster nodes can easily find their best routing path to the base station, by containing the energy consumption, storage and computation demands in both the cluster nodes and the sensor nodes. This scheme is robust from the security point of view and able to work efficiently, despite the highly constrained nature of sensor nodes.

The third paper by Jinjing Zhao and Yan Wen, describes the relation between prefix hijacking and the Internet hierarchy. The Internet is classified into three tiers based on the power-law and commercial relations of autonomous systems. The relation between network topology and prefix hijacking influence is presented for all sorts of hijacking events in different layers. The results assert that the hierarchical nature of network influences the prefix hijacking greatly.

The fourth paper by Byungha Choi and Kyungsan Cho, proposes an improved detection scheme to protect a web server from detoured attacks, which disclose confidential/private information or disseminate malware codes through outbound traffic. Its scheme has a two-step hierarchy, whose detection methods are complementary to each other. The first step is a signature-based detector that uses Snort and detects the marks of disseminating malware, XSS, URL Spoofing and information leakage from the web server. The second step is an anomaly-based detector which detects attacks by using the probability evaluation in HMM, driven by both payload and traffic characteristics of outbound packets. Through the verification analysis under the attacked web server environment, they show that their proposed scheme improves the false positive rate and detection efficiency for detecting detoured attacks to a web server.

The fifth paper by Der-Chen Huang, Yi-Wei Lee, and Hsiang-Wei Wu, proposes a guarantee time slot mechanism to enhance the performance and utilization by using CFP. Their proposed method ensures each device has the authority to access the radio channel without any additional step. By comparing with the method of IEEE 802.15.4, the experimental results show that data average transmission delay and energy consumption can be reduced dramatically. In addition, the bandwidth and performance of network is improved since the pre-allocation mechanism can reduce the number of control packets.

The sixth paper by Jianfeng Wang, Hua Ma, Qiang Tang, Jin Li, Hui Zhu, Siqi Ma, and Xiaofeng Chen, describes a new verifiable fuzzy keyword search scheme based on the symbol-tree which not only supports the fuzzy keyword search, but also enjoys the verifiability of the searching result. Through rigorous security and efficiency analysis, they show that thier proposed scheme is secure under the proposed model, while correctly and efficiently

realizing the verifiable fuzzy keyword search. The extensive experimental results demonstrate the efficiency of the proposed scheme.

The seventh paper by Ho-Seok Kang, Sung-Ryul Kim, and Pankoo Kim, presents a method to efficiently defend against DoS attacks by modifying routing protocols in the WSN. This method uses a location based routing protocol that is simple and easy to implement. In the WSN environment where the location-based routing protocol is implemented, this method disperses the DoS attack concentration of traffic by using the traffic deflection technique and blocks it out before arriving at the target destinations. To find out the number of traffic redirection nodes proper for this method, they have performed a few experiments, through which the number of such nodes was optimized.

The eighth paper by Taerim Lee, Hun Kim, Kyung-Hyune Rhee, and Sang Uk Shin, describes a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS(E-Discovery as a Service) to make the best usage of its advantages and overcome the limitations of the existing E-Discovery solutions. EDaaS enables E-Discovery participants to smoothly collaborate by removing constraints on working places and minimizing the number of direct contact with target systems. What those who want to use the EDaaS need is only a network device for using the Internet. Moreover, EDaaS can help to reduce the waste of time and human resources because no specific software to install on every target system is needed and the relatively exact time of completion can be obtained from it according to the amount of data for the manpower control. As a result of it, EDaaS can solve the litigant's cost problem.

In the ninth paper by Jinhai Huo, Bowen Deng, Shuhang Wu, Jian Yuan, and Ilsun You, a topographic-awareness and situational-perception based mobility model with path optimization for tactical MANET is proposed. Firstly, a formalized process is constructed to generate a random acceleration on nodes as the disturbance caused by small-scale topographic factors in the battlefield. Secondly, a path optimization method with the artificial bee colony algorithm is introduced to mimic the trace planning when the nodes possess the terrain information of battlefield. Thirdly, a topographic-awareness based bypass strategy is proposed to simulate the action of nodes facing large-scale terrain factors in the case when the terrain information is lacking. Finally, a situational-perception based avoidance strategy is built to simulate the process of cognition and decision when there is an encounter with the enemies on the march. The mobility model consists of the four parts above and imitates the dynamic characteristics of tactical nodes in military environment.

The tenth paper by Hae-Duck J. Jeong, Jiyoung Lim, WooSeok Hyun, and Arisu An, proposes a new location-based SNS application for the disabled population (except those who are visually impaired or the disabled who are not able to use a smartphone) with three major characteristics of this

application to be considered as follows: (i) the person uses a Social Networking Service (SNS) by constructing a friend matching system such as Facebook or Twitter, which are the most widely used SNS in the world; (ii) the general population registers real-time information for a specific location on the map for the disabled population using SNS. This information with photos and messages is given and evaluated by users; and (iii) this system makes it easier to see that the menu in the GUI was implemented.

The eleventh paper by Myunggwon Hwang, Do-Heon Jeong, Jinhyung Kim, Sa-kwang Song, and Hanmin Jung, describes activity inference for constructing user intention model. User intention modeling is a key component for providing appropriate services within ubiquitous and pervasive computing environments. Intention modeling should be concentrated on inferring user activities based on the objects a user approaches or touches. In order to support this kind of modeling, they propose the creation of object–activity pairs based on relatedness in a general domain. They also show their method for achieving this and evaluate its effectiveness.

The twelfth paper by Hsing-Chung Chen, Marsha Anjanette Violetta, Chien-Erh Weng, and Tzu-Liang Kung, presents a novel Cognitive RBAC (Role-Based Access Control) scheme which can be applied to Mobile Heterogeneous Networks (MHNs). The MHNs consist of mobile communication systems and Wi-Fi systems. The required new definitions for the RBAC model are proposed in this paper. They can improve the ability of conventional RBAC model to meet new challenges. In their scheme, they assume that a Cognitive Server (CS) provides and manages the permissions of services, and Network Providers support and manage a variety CRs and CNs, individually. For more efficiently managing CR and CN and meeting the large scale heterogeneous networks, they let mobile user can perceive network candidate actively to access services, in which the permissions are depending to the contract made by CS with each Network Provider. In this paper, the new generalized cognitive RBAC model and their definitions are proposed, and could be applied to new applications in an MHNs environment.
The thirteenth paper by Yong-Hwan Lee, Bonam Kim, and Sang-Burm Rhee, proposes a new efficient and effective mobile image retrieval method that applies a weighted combination of color and texture utilizing spatial-color and second order statistics. The prototype system for mobile image searches runs in real-time on an iPhone and can easily be used to find a specific image. In order to evaluate the performance of the new method, they assessed the Xcode simulation's performance in terms of average precision and F-score using several image databases and compare the results with those obtained using existing methods such as MPEG-7. Experimental trials revealed that the proposed descriptor exhibited a significant improvement of over 13% in retrieval effectiveness compared to the best of the other descriptors.

The fourteenth paper by Jen-Wei Hu, Chu-Sing Yang, and Te-Lung Liu, describes integrating management functions of virtual network in their testbed.

In this paper, they design and create a Future Internet testbed in Taiwan over TWAREN Research Network. This testbed evolves into an environment for programmable network and cloud computing. This paper also presents several finished and ongoing experiments on the testbed for multiple aspects including topology discovery, multimedia streaming, and virtual network integration.

The fifteenth paper by Inshil Doh, Jiyoung Lim, and Kijoon Chae, proposes an energy-efficient and secure channel group key establishment and rekeying management scheme for mobile open IPTV services. Their scheme provides the data authentication between an Evolved Node B (eNB) or a Base Station and the mobile devices for the security enhancement and efficiently rekeys the group key when the membership changes. Additionally, it proposes a pairwise key establishment mechanism for open IPTV services through eNBs. Their proposal can cope with the security vulnerability in mobile open IPTV services and guarantee the secure group key rekeying in addition to decreasing the storage and communication overhead.

The sixteenth paper by Bing Bai, Zhenqian Feng, Baokang Zhao, and Jinshu Su, presents a community-based single-copy forwarding protocol for DTNs routing, which efficiently utilizes the community structure to improve the forwarding efficiency. Simulation results are presented to support the effectiveness of their scheme.

In the seventeenth paper by Jyh-Horng Wen, Yung-Cheng Yao, and Ying-Chih Kuo, a Wiener-based successive interference cancellation (SIC) scheme is proposed to detect the OFDM signals. It provides good ICI cancellation performance; however, it suffers large computation complexity. Therefore, a modified Wiener-based SIC scheme is further proposed to reduce the computation complexity. Simulation results show the performance of the Wiener-based SIC scheme is better than those of zero forcing, zero forcing plus SIC and original Wiener-based schemes. Furthermore, with the modified Wiener-based SIC scheme, the performance is still better than the others. Although the performance of the modified Wiener-based SIC scheme suffers little degradation compared to Wiener-based SIC scheme, the computation complexity can be dramatically reduced.

The eighteenth paper by Satoshi Tanaka, Takashi Nishide, and Kouichi Sakurai, proposes an efficient implementation of computing multivariate polynomial systems for multivariate cryptography on GPU and evaluate efficiency of the proposal. GPU is considered to be a commodity parallel arithmetic unit. Moreover, they give an evaluation of their proposal. Their proposal parallelizes an algorithm of multivariate cryptography, and makes it efficient by optimizing the algorithm with GPU.

Finally, in the last paper by Tianhan Gao, Nan Guo, Kangbin Yim, they leverage the combination of PKI and certificate-based cryptography and

propose a hierarchical security architecture for the HMIPv6 roaming service. Under this architecture, they present a mutual authentication protocol based on a novel cross-certificate and certificate-based signature scheme. Mutual authentication is achieved locally during the mobile node's handover. In addition, they propose a key establishment scheme and integrate it into the authentication protocol which can be utilized to set up a secure channel for subsequent communications after authentication. their approach is the first addressing the security of HMIPv6 networks using such a hybrid approach. In comparison with PKI-based and IBC-based schemes, their solution has better overall performance in terms of authenticated handover latency.

We strongly believe that the papers presented in this special issue make significant contributions to the work and studies conducted by academic researchers, industry professionals, students, and everyone in the areas of advances on mobile collaborative systems.

We would like to thank all the authors for their valuable contributions. Our special thanks go to prof. Mirjana Ivanović, Editor in Chief of the Computer Science and Information Systems (ComSIS) Journal, for inviting us to prepare this special issue and for his productive comments and great support throughout the entire publication process.

Hae-Duck Joshua Jeong,
Fatos Xhafa,
Makoto Takizawa
Editors of the special issue

DOI: 10.2298/CSIS120725024S

The Throughput Critical Condition Study for Reliable Multipath Transport

Fei Song[1,2], Huachun Zhou[1,2], Sidong Zhang[1,2], Hongke Zhang[1,2], and Ilsun You[3]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University
100044 Beijing, P.R. China
{fsong, hchzhou, sdzhang, hkzhang}@bjtu.edu.cn
[2] National Engineering Lab for Next Generation Internet Interconnection Devices,
100044 Beijing, P.R. China
{fsong, hchzhou, sdzhang, hkzhang}@bjtu.edu.cn
[3] School of Information Science, Korean Bible University
Seoul, South Korea
isyou@bible.ac.kr

Abstract. Employing multiple paths for achieving high capability and robustness has some obvious benefits. New wireless technologies are giving more Internet access modes for notebooks and smart phones. However, most multipath protocols may not get acceptable throughput if reliable transmission is guaranteed. For some cases, the situation may even worse than using single path only. The main reason has strong relationship with the principal of multipath transmission. Motivated by these facts, specific analytical mechanisms are proposed to analyze the potential problems which may lead to serious performance decrease. Following that, we investigate how to use multiple paths legitimately when network environments are fluctuating. In our simulation, topologies for multiple paths and single path are set up for evaluating our analytical methods. Some distinguished scenarios are chosen from different perspectives. The results have revealed some throughput critical conditions and could be helpful in designing scheduling schemes for multipath protocols.

Keywords: reliable protocol, multipath transmission, critical conditions.

1. Introduction

The Internet terminal has been using one path for data transmission since the beginning. Expensive network interface and unitary access method lead to such circumstance. Single path protocols (like TCP and UDP in transport layer) are working very well even in complex network scenarios, such as Cloud computing [1], Optical networks [2], Wireless sensor network [3], etc. However, using one path may bring some insoluble problem. For example, when the path failure occurs, the users have to wait a long period for

recovery. Therefore, improving the transmission quality is always the significant target for protocol designers.

With the development of wireless technologies, people could use 802.11, 802.15, 802.16 and other methods to access the Internet, which provide the terminal a sense of possibility to adopt multiple paths for data transmission. As a promising solution for enhancing capability, reliability, security and mobility, multipath transport was widely discussed in both academic and industrial communities. New generation transport protocols, like SCTP [4] and DCCP [5], consider using multiple paths as backups. That means the data packets are still sent on one main path, only heartbeat packets are interchanging on other paths to keep them alive. This kind of mechanism could solve the problem suffered in single path transport (mentioned in previous paragraph). The users do not need to wait a long recovery time, but to use the pre-assigned path immediately. The throughput could be improved obviously. Although this scheme steps forward a little in multipath, the satisfaction is still far from enough. Concurrent multipath transfer is highly needed.

Any new technologies need an appropriate timing for expanding. The hypergrowth of notebooks and smart phones is giving multipath transport a golden opportunity to implement the academic ideas into industrialization. For example, the hardware platform and phone operation system (such as iOS or Android) could provide high speed computing and stable Internet access (via GPRS, 3G, WiFi, Bluetooth or Infrared). Many applications in smart phone are calling for high bandwidth capability based on multipath transport as well.

Using multiple paths simultaneously is quite helpful not only in aggregating the bandwidth, but also in obtaining the service from suitable Internet Service Providers (ISP) rapidly. Imaging a fancy service is located in the server of ISP A, the user who wants to get the service has two access authorities to both ISP A and ISP B. If the default setting is to connect with ISP B, all the data packets have to go through the Internet Exchange Point (IXP), which may cause evident increasing delay. Multipath transport could mitigate this by creating two connections from two ISPs, respectively. If only single path is mandatory for this service, the path created inside the same ISPs will be elected. If multiple paths are allowable, all the paths should be employed as soon as possible to boost the throughput.

The motivation of this paper is following current multipath research statues: Comparing with single path transmission, sending packets on multiple paths should have better throughput. However, that is not always true as expected in the realistic cases due to some potential problems [6]. Performance should be analyzed more carefully to figure out when the multipath should be adopted and when the single path could get higher performance. We only focus on the reliable multipath transmission here.

The main contributions of this paper are:

Proposed a specific mechanism for analyzing the potential problems which may lead to performance decline.

Studied some critical conditions by setting up suitable topologies and evaluating the analytical mechanism on both multiple paths and single path scenarios.

The ordinary principle of mainline multipath transport protocols will be discussed in Section 2. For the Internet terminals, the multipath protocols in transport layer have to consider many mechanisms to guarantee the reliable transmission. This leads to some potential problems which may influence the performance (detailed in Section 3). Some fundamental formulas are given in this Section as well. The evaluation part, given in Section 4, contains topologies description, parameters setting, performance analysis and comparison. In Section 5, related work is introduced and classified based on relevance. The conclusions and future works are explained in Section 6.

2. Principle of Multipath Transport

In multipath transport, data packets should be ejected and routed on more than one path. The end hosts may have several network interfaces and the multiple paths may have crossover point. Theoretically, no matter how complex the network scenario is, from the users' perspective, the terminals only need to care about the scheduling and reflection mechanism. In order to achieve reliable transmission, the sender needs to collect all the feedbacks giving by the network and receiver to adjust corresponding schemes.

As an indispensable part of reliable transport, schedule mechanism is responsible for controlling sending speed and sequences. For original TCP, there are some schemes to guarantee the performance of data transmission on single path. When multipath was introduced, all these schemes need to be investigated or modified.

Congestion Control: The multipath could be treated as independent individuals or a whole entity when facing the congestions. That means different congestion control schemes should be designed. The legacy of TCP could be used as reference. Some researchers deem that all the available paths should be optimal used to shift the congestion and obtain load balancing. The consensus of designing multipath protocols is to realize fairness and efficiency at the same time.

Flow Control: When TCP was just getting start running in the old days, the performance of receiver host was not very satisfactory. Especially if "high speed network" is connected, the arriving rate of data packets might be higher than the processing capability of receiver. Even for current modern equipments, establishing too many network connections (like P2P applications) may decrease the performance as well. So in multipath transport scenarios, flow control is also a critical problem. More than that, some recent works have expanded flow control to solve power consumption problems [7].

Sequence Control: For some software inside application layer, data packets are expected to arrive at the destination in order. Due to the parameter variation or restriction policies of ISP network, packets might be

halted inside some routers and lead to disorder situation. It is quite widespread in multipath transport. Assigning Transmission Sequence Number (TSN) to each packet and using them to distinguish out of order are always necessary for reliable multipath transmission.

Retransmission Control: The packet loss is inevitable in the Internet due to current switching and forwarding rules of network infrastructure. It is not hard to perceive which packet is lost based on the information carried by acknowledgments (ACK). The multipath protocols could set reasonable timers for calculation. Path selection in retransmission is also important and might affect the transmission fairness. The core idea in this part is to let the receiver host obtain the lost packet as soon as possible.

It seems that most controlling work are assigned to the sender side, and the receiver is just responsible for collecting and submitting the data packets into application layer, then send the acknowledgements back to the sender. In fact, receiver provides a lot of information in the process of data transmission. Both of them have to deal with some potential problems for enhancing the throughput when using multipath.

3. Potential Problems Analysis

For each Internet end host, there are send buffer and receive buffer in charge of storing unconfirmed output packets and trimming the disorderly input packets, respectively. Both of them may encounter a serious blocking event, which is able to decrease the performance according to the principle of current reliable transport on multiple paths and single path.

3.1. Send buffer blocking

There are two types of data packet inside send buffer: New data packets and retransmission data packets. Each output packets will be added at the tail of retransmission queue and removed when it is delivered to the upper layer by the receiver. If the space in the send buffer is too narrow for sending new packets and the sending speed has been seriously affected, the situation could be called send buffer blocking.

Reliable single path protocols in transport layer, such as TCP and SCTP, need to gather the status information from ACKs. Multipath inherits this method when exchanging data packets. Original ACKs could only indicate the earliest data packets the receiver needs. Although some packets have arrived and stored at receiver's buffer, the sender may still consider these packets have been lost. This will lead to unnecessary degradation of sending speed.

Selective Acknowledgment (SACK) is used to inform the sender how many packets have been received successfully in both ordered and disordered status. It is quite helpful for the sender to release the send buffer. The TCP throughput could be improved obviously [8] [9]. Nowadays, more and more

web servers support SACK. However, based on the RFC2018 [10], the receiver has the right to SACK some out-of-order packets and then discard them for some particular reasons (such as the operating system needs to reuse previously allocated memory for other processes). This action is called “data receiver reneging”. In order to tolerate the renege, the senders of reliable protocols have to store the copies of the SACKed data in its send buffer.

The receiver of TCP never deliver out-of-order packet to upper layer. However, In SCTP, multiple streaming is the default option which means the packets with continuous TSN may be assigned to different streams. Therefore, out-of-order packets with different TSN may belong to the same stream and have continuous Stream Sequence Number (SSN). These packets could be delivered to the application layer directly. If that happens, they become non-renegable. The problem is the sender could not distinguish the “real” disordering packets stored in the receive buffer from the “fake” disordering packets which has been delivered to the upper layer. That will lead to send buffer waste because only the cumulatively ACK could trigger the free up action for relevant packets. That means all the SACKed packets need to be reserved at send buffer even they have been delivered. Such situation could ascribe to the send buffer blocking.

Non-renegable selective acknowledgments (NR-SACKs) [11] is proposed to cope with above problems. The idea is to modify the original SACK packet by adding new field for reporting non-renegable packets. The performance of new method is tested both on multiple paths and single path scenarios [12]. Send buffer could be released efficiently by using NR-SACK based on the simulation results. This method is quite helpful for all the reliable protocol in transport layer that uses SACK and allow delivery of disordered data to upper layer. The authors deem that if data packet renege is rarely in the current Internet, NR-SACK would not bring too much burdens for both sender and receiver, which means that understanding the probability of renege is critical for deploying the NR-SACK.

3.2. Receive buffer blocking

The appearance of receive buffer blocking is quite common in reliable transmissions. In multipath transport scenario, all available paths have their own characteristics. Data packets sent on different paths may not keep in sequence when they arrive at the destination. These out-of-order packets have to be stored in the receive buffer. If there are a lot of these kinds of packets, the free space in the buffer will be quite small. Normally, the sending speed of reliable protocols is controlled by the size of local congestion window and free space of peer receive buffer, which will be detailed in the following. Therefore, the throughput of multipath transport might be limited by the receive buffer blocking.

For current Internet, there are three potential reasons may lead such phenomena. To facilitate the presentation, we employ two paths for illustration in this Section and below.

Buffer size related: In most implementations of transport layer protocols, there are a default size of receive buffer. For the single path transport, disordered packets are not quite prevalent (comparing with multiple paths environment). Even that, the relationship between receive buffer size and performance is also quite important. When sending the data packets via multipath, the setting of receive buffer size is hard to determine. The protocol could assign each path an independent receive buffer. However, the sum of buffer size on all paths is still finite and restrict packet deliver to the application layer. So this mechanism would not gain a lot. The general idea of the receive buffer size effect is that lager space could hold more out-of-order data packets. Meanwhile, larger space will bring more burdens for the operating system as well. The relevance between buffer size and transmission performance still needs more experiments.

Packets loss related: Basically, packets loss caused by intermediate routers or the links is quite common in the Internet. For the purpose of avoiding the heavy congestion, routers will drop some packet based on the different schemes (such as Drop Tail or RED). In wired network, link error loss is less than that in wireless network. Beside these two reasons, some loss event may appear when the handover occur in wireless environment. If there is a loss event on one path, packets sent on other paths could not reach the receiver side in order. For comparing the detail, we propose a mechanism for analyzing the whole process with “Non packets loss related” together in the following.

Non packets loss related: Receive buffer blocking may be triggered as well even if there is no packet loss in multipath transport. Different bottleneck bandwidth and transmission delay are able to generate out-of-order packets in the sending process. The value of bottleneck bandwidth determines forwarding speed at the intermediate routers. It will take long periods for a router which has low output bandwidth to send the data packets to the next hop. If the rate of ingress is higher than the rate of egress, some data packets have to wait in the queue of intermediate router.

Fig. 1 illustrates that multipath transport flows and other flows are sharing the same route. The packets of multipath were sent on two paths which have different bottleneck bandwidth. The value of bandwidth on path B is larger than another path. So for path A, there are more packets waiting inside the buffer of router. In Fig. 1, the number printed in each packet is TSN. Data packets sent from source are not able to arrive at destination in order. These packets will occupy much space in receive buffer.

The diversity of hops on available paths which are set up at the initial stage of connecting may bring different transmission delay in multipath transport. It is possible that the values of transmission time on different paths are not similar even if the paths have the same hops. In Fig. 2, there are two access modes (GPRS and 802.15x) for the sender to connect to the Internet. The receiver has two access modes (CDMA and 802.11x) as well. The data

packets which were sent from different interfaces may not be able to arrive at destination simultaneously.

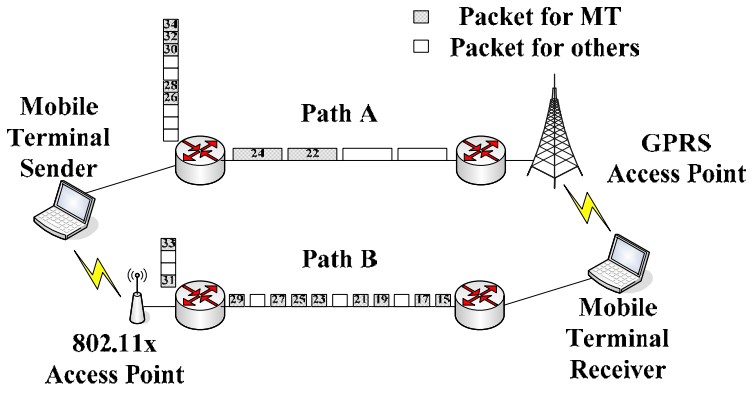


Fig. 1. Bottleneck bandwidth variety (with other flow)

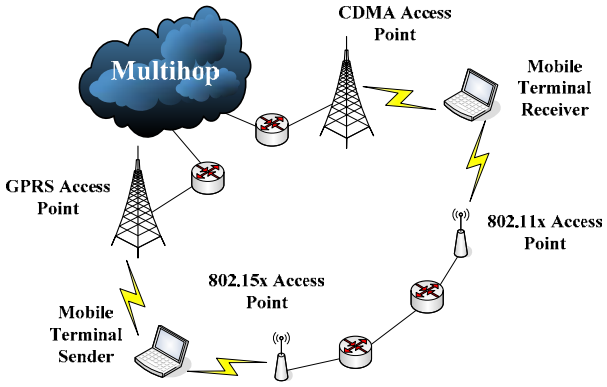


Fig. 2. Transmission delay variety

The reasons which may cause receive buffer blocking can be illustrated in Fig. 3 and Fig. 4. They are very useful for understanding the whole blocking process. CMT-SCTP proposed in [13] is used as an example in this Section.

To make it more simplify, we assume that delay ACK is not adopted. However, the following mechanism is also serviceable for delay ACK scenarios. When payload size of one packet is 1468 Bytes (It is according to the definition of CMT-SCTP. In TCP, the maximum size of payload is 1460 Bytes), the receiver is able to contain only 11 packets if the size of receive buffer is set to 16KB. Here are some definitions:

C*x*: the size of congestion window (*Cwnd*) on path *x*.

O*x*: the number of unacknowledged packets (i.e. Outstanding packets) which was sent on path *x*.

A*x* and B*x*: the interfaces of sender and receiver, respectively. Two interfaces are connected if the value of *x* is the same.

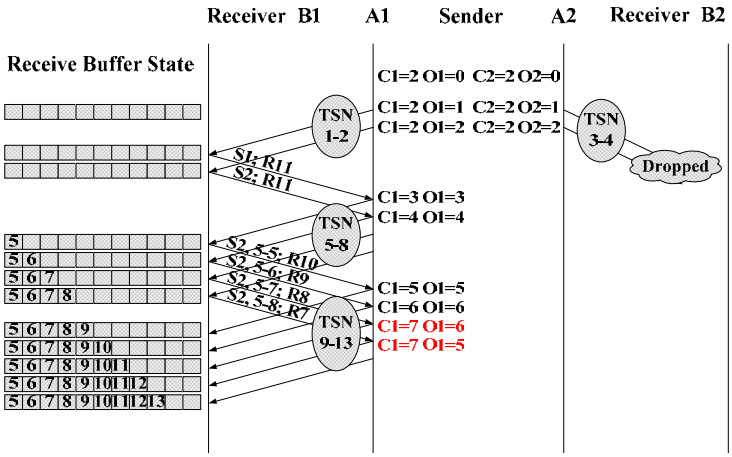


Fig. 3. Receive buffer blocking caused by packets loss

<S*a*, *b*-*c*; R*d*>: *a* indicates accumulative TSN. *b*-*c* means the packets with TSN between *b* and *c* have been acknowledged by gap. *d* means the size of free space in the receive buffer (*Rwnd*). The unit of all the parameters is packet.

Arrow lines indicate the process of sending and receiving packets.

The TSN of each packet is printed at the beginning of arrow lines.

There are 11 squares at the left side in the figures to show the state of receive buffer. It will be changed whenever the data packets arrive at receiver side. Only the latest state is shown when two packets which were sent on different paths reach the destination almost at the same time.

Inside receive buffer state squares, the transparent number indicate the TSN of packets which have been submitted to application layer. These squares with transparent number are just used as an identifier for submission process and will not take up any space in receive buffer. For instance, in Fig. 4, when the receive buffer state is showing: 13, 9 (written in transparent style), 14, there are only two packets in the buffer actually. And the sequence of them is 13 and 14.

Based on above preparations, we could describe the details about how the receive buffer blocking happens in packet loss situation (shown in Fig. 3). After hand shake stage, the initial congestion window on both paths is set to 2. The number of outstanding packets is 0. Then, the packets whose TSN is 1, 2 and 3, 4 will be sent from path 1 and 2, respectively. If the packets which were sent on path 2 are lost on the forward direction, the packets with TSN 1-2 will be submitted to application layer as soon as they arrive. The corresponding acknowledged packets will be generated and sent from receiver side. The accumulative TSN is 2 and the size of free space in the receive buffer is still 11.

The sender increases the value of *Cwnd* to 3 when the acknowledgement of packet with TSN 1 was received. Outstanding number decrease to 1(not

shown in Fig. 3). The number of packet which is permitted to send can be calculated by the formula (1) and (2):

$$SendWnd = \min\left(Cwnd,\ PeerRwnd\right) \quad (1)$$

$$PeerRwnd = Rwnd - \sum_{i=1}^{n} Outstanding_i \quad (2)$$

where *n* is the number of path which used for sending data packets. $Outstanding_i$ means the total number of unacknowledged packets on path *i*. In this case, the value of outstanding packets on path 1 and path 2 are 1 and 2, respectively. R11 indicates the free space in *Rwnd* is 11. Based on formula (2), the *PeerRwnd* is 8. The value of *Cwnd* on path 1 is 3. Therefore, the *SendWnd* is 3 according to the formula (1). Due to the packet with TSN 2 has not been acknowledged, the sender can only send 2 packets. After sending these packets, the number of outstanding packets is changed to 3. These figures only illustrate the final value of outstanding packets when all the permitted packets have been sent on one path. Transition states discussed here are not shown in the figures.

The *Cwnd* will be changed to 4 when the acknowledgment of TSN 2 was received. Outstanding value is changed to 2. According to the previous method, 2 packets are allowed to be sent. Then the value of outstanding packet is increased to 4.

When packets with TSN 5-8 arrive at the destination, the receiver can not submit them to the application layer at once because the TSN 3 and 4 have been lost. Therefore, the accumulative TSN at the receiver side is still 2 and packets with TSN 5-8 can only be acknowledged by gap in SACK. Meanwhile, the free space in receive buffer has been changed from 11 to 7.

When the acknowledgments of packets with TSN 5 and 6 have been received, the previous analysis method remains correct because *Cwnd* is smaller than *PeerRwnd*. However, the value of *PeerRwnd* will be 1 when the acknowledgment of packet with TSN 7 was received. That means only 1 packet can be sent. After sending this packet, the outstanding number will be changed to 6. We use red character to identify the sender has been influenced by receive buffer blocking. When the acknowledgment of packet with TSN 8 was received, the number of outstanding packets on path 1 is changed to 5. The *PeerRwnd* is 0 because both the *Rwnd* and outstanding packets on two paths are 7. Therefore, no packets can be sent. Only five packets were sent after the acknowledgments of packets with TSN 5-8 were received.

In non packet loss situation (shown in Fig. 4), the packets whose TSN is 1, 2 and 3, 4 will be sent from path 1 and 2, respectively. All these four packets can be submitted to application layer because they arrive at destination in order. The sender change the value of *Cwnd* and outstanding packets value when the acknowledgments of the packets with TSN 1-2 were received, then the sender will calculate *SendWnd* based on (1) and (2). After that, the packets with TSN 5-8 are sent via path 1.

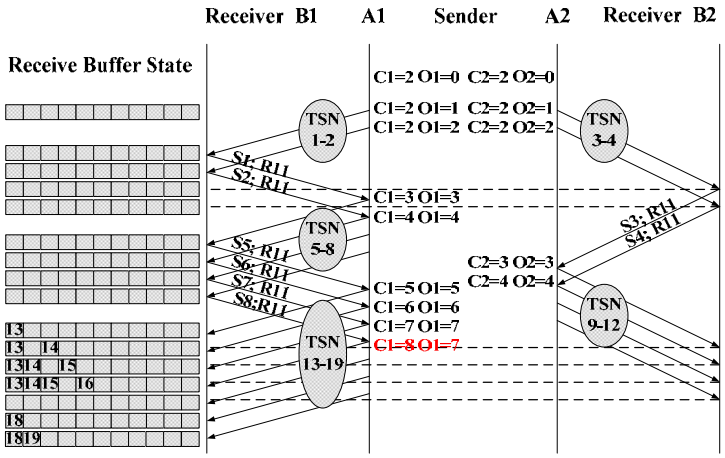


Fig. 4 Receive buffer blocking caused by non packets loss

For the path 2, when the SACKs of packets with TSN 3-4 are received, the sender will update the *Cwnd* and outstanding packets value again, and then transmit the packets with TSN 9-12 to the receiver. The C2 and O2 are equal to 4 when this transmission round is finished.

The *Cwnd* is always smaller than *PeerRwnd* when the SACKs of packets with TSN 5-7 are received. We can use similar method for analysis. However, when the acknowledgment of packet with TSN 8 arrives at the sender side, only 1 packet can be sent. Then the value of outstanding packet will not increase. So 7 packets were sent after the acknowledgements of packets with TSN 5-8 arrived. Receive buffer blocking happens again.

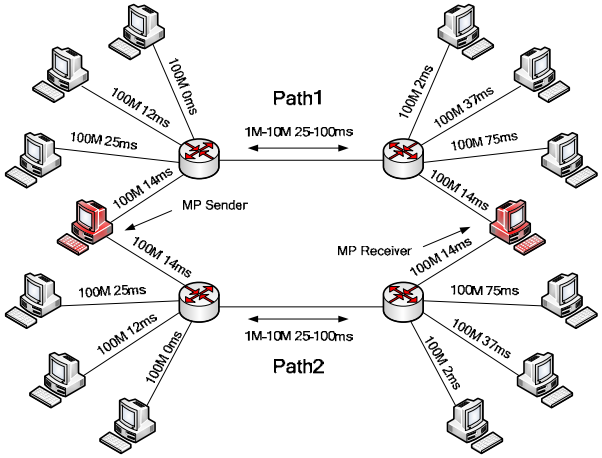


Fig. 5 Topology for multiple paths transport

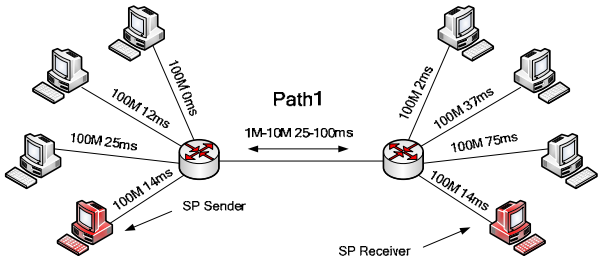


Fig. 6 Topology for single path transport

4. Performance Simulation

In Section 3, the main potential problems which may lead to performance decrease in reliable multipath transport were analyzed. We argue that if these problems emerge frequently, using single path might be a better choice to get higher throughput. To evaluate this idea, we set up two topologies following the guidelines proposed in [14] and adopt the CMT-SCTP and original SCTP as the object reliable multipath and single path protocols, respectively. All the simulations are running in NS2-2.35 [15]. The basic structures of multiple paths (MP) and single path (SP) topologies are showed in Fig. 5 and Fig. 6.

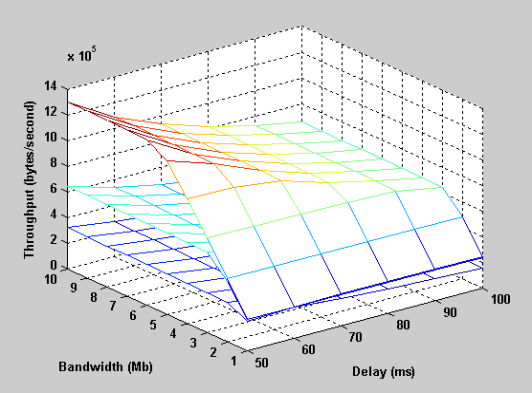


Fig. 7 Multipath throughput comparison for Scenario One

Two multipath transport nodes shown in red are adopted and other 12 nodes are used to generate application layer traffic. FTP, HTTP and UDP flows are sent randomly from left nodes to the right nodes. The value of bandwidth and delay between terminals and routers are fixed. The network parameters of bottleneck are set dynamically in different scenarios. The drop tail routers are used in our simulations. For the CMT-SCTP, three algorithms

(CUC, SFR and DAC) introduced in [16] are employed to ensure the transport quality. Chunk size is set to 1468 bytes. The send buffer of CMT-SCTP and SCTP terminals is set to the default size. The receive buffer size are changed in various experiments. Random seeds are used to run each experiment for 30 times.

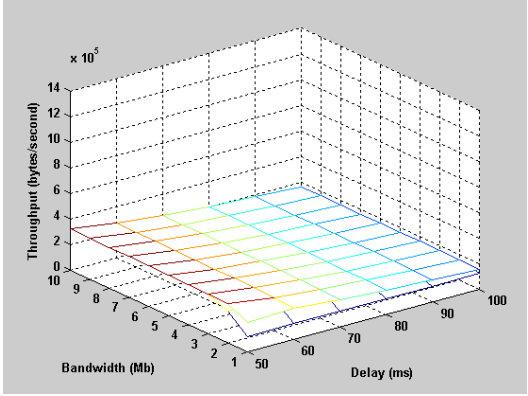


Fig. 8 Multipath vs Single path (buffer size = 32K) for Scenario One

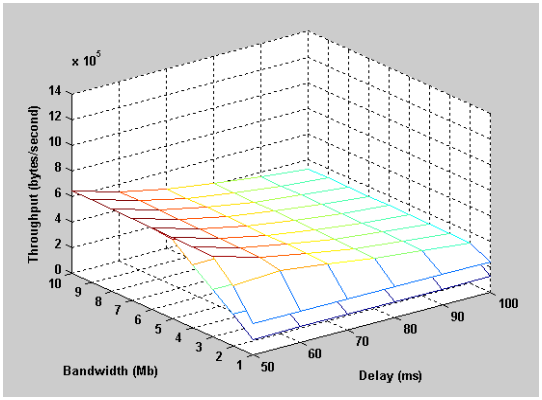


Fig. 9 Multipath vs Single path (buffer size = 64K) for Scenario One

In order to compare the transport performance between multiple paths and single path, several pretreatments are needed. Based on the previous analysis, NR-SACK could be used to enhance the utilization rate of send buffer. We enable the NR-SACK in all simulations. However, due to the restriction of receiving and delivering procedure, the influence of receive buffer blocking needs more investigation.

There are three possibilities which may aggravate receive buffer blocking and decrease the throughput. We set up three different scenarios for testing the performance.

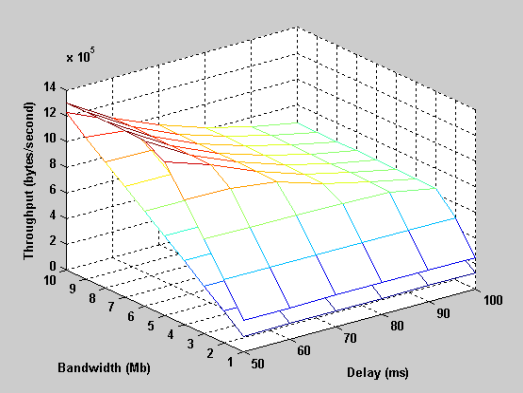


Fig. 10 Multipath vs Single path (buffer size = 128K) for Scenario One

4.1. Scenario One: influence of receive buffer size

The values of bandwidth and Round Trip Time (RTT) on the bottleneck links are modified. The value of bandwidth is changing from 1 M to 10 M (step size is 1M). The RTTs fluctuate from 50ms to 100ms (step size is 10ms). In order to clearly specify the relationship among bandwidth, RTTs and throughput, we did not add any packets loss in this scenario.

Three-dimensional figures could be used to illustrate the results. In Fig. 7, when the receive buffer size is set to 32K, 64K and 128K, the throughput is getting higher and higher. Based on analyzing the trace files, sending window is limited by the *PeerRwnd* if small buffer size is set. With the increasing of receive buffer, the limitation for the sender becomes weakening. There are not too much out-of-order data packets because the parameters of each bottleneck link are modified simultaneously.

We could find that the variation of RTT bring little changing in larger receive buffer size case if the bandwidth is low, which means the fluctuation of throughput is smaller than 1%. With the growth of bandwidth, the disadvantages brought by large RTT value are gradually obvious. In the 128K case, the throughput is reduced 49.8%. In other cases, the decrement rate is also around 50%.

However, the bandwidth plays an important role if larger receive buffer size are assigned. In 128K case, when the RTT is set to 50ms, the throughput will increase 424.2% if the bandwidth changed from 1M to 10M. The growth rate is getting low if smaller buffer size is set. In the case of 64K and 32K (RTTs are equal to 50ms as well) the performance will be increased 163.9% and 39.6%, respectively.

For analyzing the benefits given by larger receive buffer size, we fix the bandwidth and RTT to the best case and worst case to check the performance. When double and quadruple the 32K, the throughput increase

99.7% and 303.6% if 10M and 50ms are adopted. That means the growth rate is in proportion to the receive buffer size in the best case. However, the performance only increase 54.2% and 60.6% when amplifying buffer size to 64K and 128K in the worst case, i.e. bandwidth and RTT are equal to 1M and 100ms.

We compare the multiple paths and single path performance in Fig. 8, Fig. 9 and Fig. 10 scenario. The general view is that the disadvantage of multipath is not shown up. All results indicate single path could not get higher performance. We also find two main features: One is the throughput enhancement is quite obvious (86%, 96% and 98% in 32K, 64K and 128K, respectively) when low bandwidth and high RTT are set. The other feature is that marginal effect for multiple paths transport throughput is more serious than that in single path transmission when the bandwidth is varying from 1M to 10M.

4.2. Scenario Two: influence of packets loss factor

The packets drop will trigger the receive buffer blocking quickly. To simulate such process and test the performance, we run some experiments which fix RTTs (50ms on both paths) and change other parameters. The value of bandwidth is set from 1M to 10M and the loss rate is increased from 0.01 to 0.08. Receive buffer size are set to 32K, 64K and 128K.

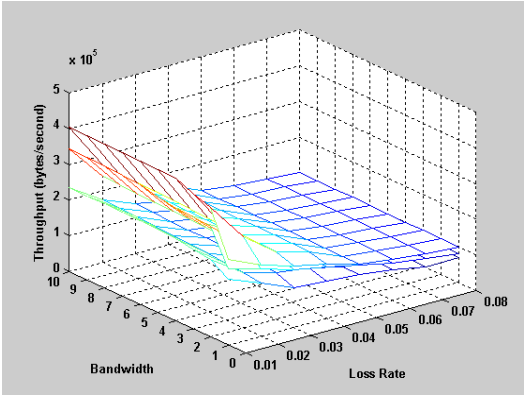


Fig. 11 Multipath throughput comparison for Scenario Two

In Fig. 11, three curved surfaces are used to illustrate and compare the performance of multipath transport. Generally, the larger receive buffer size is, the more throughput one could get. The difference among three curved surfaces is not significant. If the loss rate is getting higher (larger than 0.04), the variation tendency of them are quite similar. With the increase of bandwidth, the influence given by packet lost could be remitted, which enhance the multipath throughput. One interesting phenomenon, like in

scenario one, is that the marginal effect are highlighted when the bottleneck bandwidth is larger than 2M. We could find the point of inflection is smaller than that in scenario one.

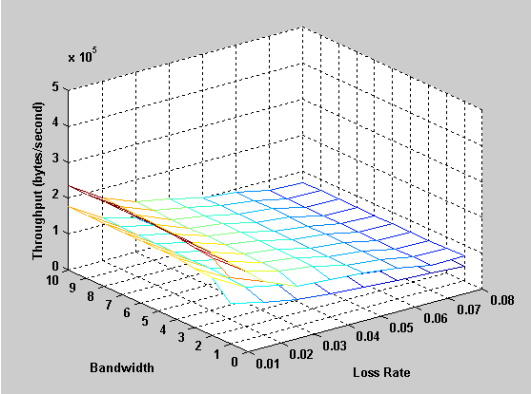


Fig. 12 Multipath vs Single path (buffer size = 32K) for Scenario Two

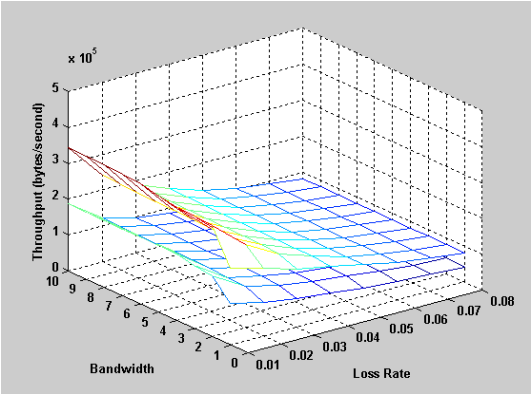


Fig. 13 Multipath vs Single path (buffer size = 64K) for Scenario Two

In Fig. 12, Fig. 13 and Fig. 14, single path transport results are added for comparison when the buffer size is set to 32K, 64K and 128K. Although, the bandwidth and loss rate are assigned to different value in the experiments, which may aggravate the receive buffer blocking, the throughput of multipath is always higher than the values of single path transport. The results show the performance is enhanced 36.3%, 84.1% and 114.5% in the best case.

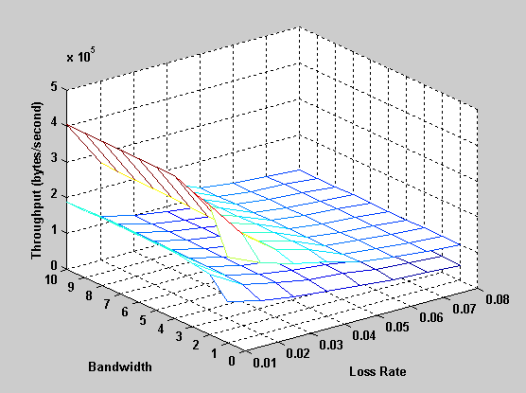


Fig. 14 Multipath vs Single path (buffer size = 128K) for Scenario Two

4.3. Scenario Three: influence of non packets loss factor

In this scenario, we fix the bandwidth of bottleneck link to 10M and vary the receive buffer size and RTT. Based on the analysis in Section 3, different RTT on two paths may block the receive buffer. The first experiment, shown in Fig. 15, is to change the RTT (from 50ms to 100ms) of the alternate path and leave main path RTT (50ms) untouched. 32K, 64K and 128K receive buffer size are still used here. When two paths have same RTT, the throughput of multiple paths and single path are quite similar if the buffer size is set to 32K or 64K, which means the performance is restricted mainly by the size of buffer. Such situation is released when the buffer size is 128K. With the increase of RTT on alternate path, the throughput of multipath is getting down. The decrease rate in 32K, 64K and 128K are 47.3%, 41.5% and 48.7%, respectively. The multipath (shown in solid line) is defeated by single path (shown in dash line) when the RTT is higher than 55ms.

The second experiment shown in Fig. 16 explores the performance variation when changing two paths' RTT simultaneously and maintaining the sum RTT of two paths (equals to 100ms). Due to the RTT of main path is getting smaller, the throughput of single path will increase. Although the curve of multipath has fluctuation, it is lower than the curve of single path in most cases. When the receive buffer size is set to 64K, the decrease rate will reach 64.1%. We could find the variation tendency of multipath and single path are quite different with the changing of RTTs.

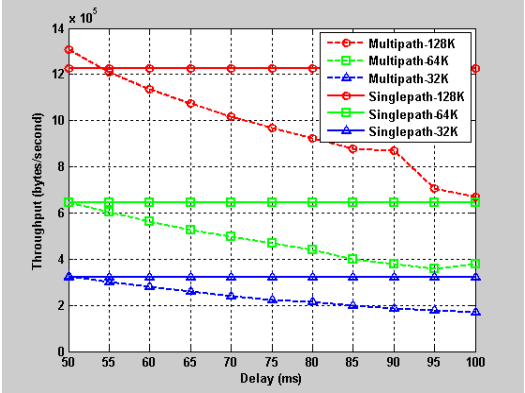


Fig. 15 Multipath vs Single path (adjust RTT in one path) for Scenario three

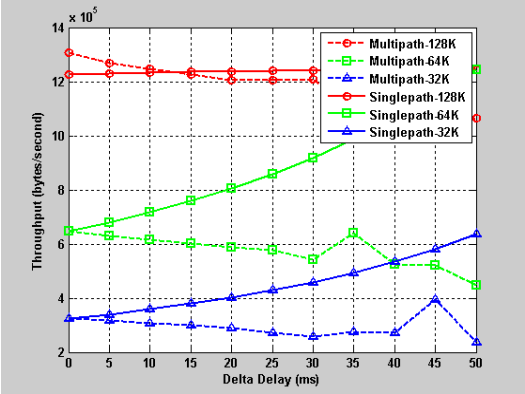


Fig. 16 Multipath vs Single path (adjust RTT in two paths) for Scenario Three

5. Related Work

Researches related with reliable transport on multiple paths have different emphases. Multipath TCP or SCTP should be discussed firstly due to correlation of our work.

Based on the principle of resource pooling [17], some new mechanisms were introduced and evaluated. The main goal of designing MPTCP [18] [19] is to be deployable and usable without significant changes to existing Internet infrastructure. In [20], an effective proposal for large-scale data centers is presented to enhance the throughput and fairness. Raiciu et al. [21] proposed an opportunistic-based mobility method by using multipath TCP. In order to settle the congestion problem in MPTCP, Wischik et al. [22] designed a multipath congestion control algorithm and implemented it into Linux system. The adoption of MPTCP in current network was analyzed from business

perspective in [23], which might be quite helpful in understanding the critical issues in the deployment process of multipath TCP.

In send buffer blocking part, although NR-SACK or other renege solutions are quite helpful in mitigating the blocking problem, investigations for occurrence frequency of data renege in current Internet are highly needed. A reasonable way is trying to process the dataset gathering from the real network environment. In [24], TCP traces obtained from Cooperative Association for Internet Data Analysis (CAIDA) [25] was analyzed to infer the state of receive buffer. The first step results show that data reneging appears constantly. However, the further work [26] point out that the generation of SACKs in many TCP implementations was unreasonable comparing with the RFC2018. Some necessary SACKs were wrongly sent or even not sent. Seven misbehaviors were demonstrated and discussed to show the risks. A series of extensions mechanisms which could be added into TBIT [27] were proposed for detecting these misbehaviors.

For the research of receive buffer blocking, Iyengar et al. [28] compare different retransmission schemes in finite receive buffer size to see which one is more suitable for reducing receive buffer blocking in concurrent multipath transfer (CMT). More details can be found in [16]. Natarajan et al. [29] added a new state called “Potentially-failed” into original CMT mechanism to reduce the receive buffer blocking. It can provide acceptable throughput when the path failure occurs. However, these works are not able to improve the performance in non packet loss related situation. Increasing the size of receive buffer can relieve this problem in all scenarios obviously, but it makes no sense for these hosts which do not have enough resource. The receive buffer blocking is still an open issue.

6. Conclusions and Future Work

The multipath transport has been attracting the research attention for some time. It is a reasonable solution for improving the capability, reliability, security, mobility and other property in the Internet. Normally, comparing with single path transport, people consider that sending data packets on multipath should have better performance. We argue such an opinion would not always hold water for the reliable multipath transport. In this paper, we proposed a specific mechanism for analyzing the potential problems which may lead to performance reduction. Then some suitable topologies for evaluating the analytical mechanism on both multiple paths and single path scenarios were set up.

The evaluations are divided into three scenarios. For the first scenario, when changing the size of receive buffer, all experiments are affected. Using multipath shows better throughput than adopting single path. An interesting finding is the marginal effect for multipath is more significant than single path. It is also confirmed via scenario two. In addition, the performance contrast of the second scenario is shown when the bandwidth and drop rate are adjusted

simultaneously. In scenarios three, the insufficient of multipath are revealed when fix the bandwidth and adjust the receive buffer size and RTTs, which explain the transmission on single path could beat multipath in some specific environments. More details, including increasing and decreasing rate for each experiments, could be found in the performance simulation part.

This paper is focusing on the end to end perspective. A more challenging question is trying to enable multipath in the whole network and validate the performance in different scenarios, which should be studied in the future work.

Acknowledgment. This work was supported in part by the SRFDP under Grant No. 20120009120005, in part by the Natural Science Foundation of China under Grant No. 61271202, in part by the MIIT of China under Grant No. 2012ZX03005003-04, in part by the Beijing Natural Science Foundation under Grant No. 4122060.

References

1. T. Lee, H. Kim, K. H. Rhee, and S. U. Shin. A Study on Design and Implementation of E-Discovery Service based on Cloud Computing, Journal of Internet Services and Information Security, 2(4), pp. 65-76, Nov. 2012.
2. V. Q. Bien, R. V. Prasad, and I. Niemegeers, Handoff in Radio over Fiber Indoor Networks at 60 GHz, Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 1(3), pp. 71-82, Sep. 2010.
3. Ž. Živanov, P. Rakić, and M. Hajduković, Wireless Sensor Network Application Programming and Simulation System. Computer Science and Information Systems, Vol. 5, No. 1, 109-126. 2008.
4. R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. Schwarzbauer, T. Taylor, I. Rytina, M. Kalla, L. Zhang, and V. Paxson, Stream Control Transmission Protocol, RFC 2960, IETF, Oct. 2000.
5. E. Kohler, M. Handley, and S. Floyd, Datagram Congestion Control Protocol (DCCP), RFC 4340, IETF, Mar. 2006.
6. F. Song, H. Zhou, S. Zhang, H. Zhang, and I. You, Performance Analysis of Reliable Transmission on Multiple Paths and Single Path, IMIS Innovative Mobile and Internet Services in Ubiquitous Computing, 2012.
7. T. Enokido, A. Aikebaier, and M. Takizawa, Computation and Transmission Rate Based Algorithm for Reducing the Total Power Consumption, Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 2(2), pp. 1-18, Jun 2011.
8. R. Bruyeron, B. Hemon, and L. Zhang, Experimentations with TCP Selective Acknowledgment, ACM Computer Communication Review, 28(2), pp. 54-77, 1998.
9. M. Allman, C. Hayes, H. Kruse, and S. Ostermann, TCP Performance over Satellite Links, 5th International Conference on Telecommunications Systems, 1997.
10. M. Mathis, J. Mahdavi, and S. Floyd, A. Romanow, TCP Selective Acknowledgment Options, RFC2018, IETF, Oct. 1996.

11. P. Natarajan, N. Ekiz, E. Yilmaz, P. Amer, J. Iyengar, and R. Stewart, Nonrenegable Selective Acks (NR-SACKs) for SCTP, International Conference on Network Protocols (ICNP), Orlando, 2008.
12. E. Yilmaz, N. Ekiz, P. Natarajan, P. Amer, J. T. Leighton, F. Baker, and R. Stewart, Throughput Analysis of Non-renegable Selective Acknowledgments (NR-SACKs) for SCTP, Computer Communications, 33(16), 2010.
13. J. Iyengar, P. Amer, and R. Stewart. Concurrent Multipath Transfer Using SCTP Multihoming over Independent End-to-end Paths. IEEE/ACM Transactions on Networking, 14(5):951-964, 2006.
14. L. Andrew, C Marcondes, S. Floyd, L. Dunn, R. Guillier, W. Gang, L. Eggert, S. Ha, and I. Rhee, Towards a Common TCP Evaluation Suite. In PFLDnet, 2008.
15. The Network Simulator NS-2, version 2.35. www.isi.edu/nsnam/ns/.
16. J. Iyengar, P. Amer, and R. Stewart, Performance Implications of a Bounded Receive Buffer in Concurrent Multipath Transfer, Computer Communications 30 (2007) 818-829.
17. D. Wischik, M. Handley and M.B. Braun, The Resource Pooling Principle, ACM SIGCOMM Computer Communication Review, 38(5), Oct. 2008.
18. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, Architectural Guidelines for Multipath TCP Development, RFC 6182, IETF, Mar. 2011
19. M. Bagnulo, Threat Analysis for TCP Extensions for Multipath Operation with Multiple Addresses, RFC 6181, IETF, Mar. 2011.
20. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, Improving Datacenter Performance and Robustness with Multipath TCP, ACM SIGCOMM, 2011.
21. Raiciu, D. Niculescu, M. Bagnulo, and M. Handley, Opportunistic Mobility with Multipath TCP, ACM MobiArch, 2011.
22. Wischik, C. Raiciu, A. Greenhalgh, and M. Handley, Design, Implementation and Evaluation of Congestion Control for Multipath TCP, Usenix NSDI, 2010.
23. T. Levä, H. Warma, A. Ford, A. Kostopoulos, B. Heinrich, R. Widera, and P. Eardley, Business Aspects of Multipath TCP Adoption, In Towards the Future Internet-Emerging Trends from European Research, IOS Press, 2010.
24. N. Ekiz, P. Amer, A Model for Detecting Transport Layer Data Reneging, PFLDNeT, Lancaster, PA, 2010.
25. http://www.caida.org/data/passive/
26. N. Ekiz, A. Rahman, and P. Amer, Misbehaviors in TCP SACK Generation, ACM Computer Communications Review, 41(2), 2011.
27. The TCP Behavior Inference Tool, www.icir.org/tbit/
28. J. Iyengar, P. Amer, and R. Stewart, Receive Buffer Blocking in Concurrent Multipath Transport, IEEE GLOBECOM, St. Louis, MO, Nov. 2005.
29. P. Natarajan, J. Iyengar, P. Amer, and R. Stewart, Concurrent Multipath Transfer using SCTP Multihoming: Introducing Potentially-failed Destination State, IFIP International Conference on Networking, Singapore, May. 2008.

Fei Song received his Ph.D. degree from Beijing Jiaotong University. He is now a Lecturer in National Engineering Laboratory for Next Generation Internet Interconnection Devices, School of Electronic and Information Engineering, Beijing Jiaotong University. His current research interests are Protocols Optimization, Wireless Communications and Cloud Computing.

Huachun Zhou received his B.S. degree from People's Police Officer University of China in 1986. He also received his M.S. and Ph.D degree from Beijing Jiaotong University, Beijing, China in 1989 and 2008, respectively. He is now a professor in the National Engineering Laboratory for Next Generation Internet Interconnection Devices at Beijing Jiaotong University. His research interests include mobility management, mobile and secure computing, routing protocols, network management technologies and database applications.

Sidong Zhang received his M.S. degree in communication and information engineering from Beijing Jiaotong University in 1982. He is now a professor in Beijing Jiaotong University. His research interests are computer networks, wireless communications technology, ad-hoc networks, sensor networks and information theory.

Hongke Zhang received his M.S. and Ph.D. degrees in Electrical and Communication Systems from University of Electronic Science and Technology of China in 1988 and 1992, respectively. From Sep. 1992 to June 1994, he was a post-doc research associate at Beijing Jiaotong University. In July 1994, he joined the School of Electronic and Information Engineering, Beijing Jiaotong University, where he is a professor. He has published more than 100 research papers in the areas of communications, computer networks and information theory. He is the holder of more than 50 Chinese patents and is the Chief Scientist of a National Basic Research Program of China ("973 Program"). Dr. Zhang is the recipient of various awards such as the Zhan Tianyou Technical Innovation Award and the Mao Yisheng Technical Innovation Award.

Ilsun You received his M.S. and Ph.D. degrees in Computer Science from Dankook University, Seoul, Korea in 1997 and 2002, respectively. From 1997 to 2004, he worked for the THINmultimedia Inc., Internet Security Co., Ltd. and Hanjo Engineering Co., Ltd. as a Research Engineer. Since March 2005, he has been an Assistant Professor in the School of Information Science at the Korean Bible University, South Korea. Prof. You has served or is currently serving on the organizing or program committees of international conferences and workshops such as IMIS, MobiWorld, MIST 2009, MUE, UASS, TRUST, PCASS-06, WPS-08, FGCN-07, IPC-07, SAMNet-08 and so forth. He is in the editorial board for International Journal of Ad Hoc and Ubiquitous Computing (IJAHUC), Computing and Informatics (CAI), International Journal of Smart Home (IJSH) and Journal of Korean Society for Internet Information (KSII). Also, he has served as a guest editor of several journals such as CAI, MIS, AutoSoft and WCMC. His main research interests include?internet security, authentication, access control, MIPv6 and ubiquitous computing. He is a member of the IEICE, KIISC, KSII, KIPS, and IEEK.

Received: July 25, 2012; Accepted: January 05, 2013

DOI: 10.2298/CSIS120907025C

Using Bivariate Polynomial to Design a Dynamic Key Management Scheme for Wireless Sensor Networks

Chin-Ling Chen[1], Yu-Ting Tsai[1], Aniello Castiglione[2] and Francesco Palmieri[3]

[1] Department of Computer Science and Information Engineering
Chaoyang University of Technology,
Taichung, 41349, Taiwan
{clc, s10027612}@mail.cyut.edu.tw
[2] Department of Computer Science, University of Salerno
Via Ponte don Melillo, I-84084 Fisciano (SA), Italy
castiglione@ieee.org
[3] Department of Industrial and Information Engineering, Second University of Naples, Via Roma, I-81031 Aversa (CE), Italy
francesco.palmieri@unina.it

Abstract. Wireless sensor networks (WSN) have become increasingly popular in monitoring environments such as: disaster relief operations, seismic data collection, monitoring wildlife and military intelligence. The sensor typically consists of small, inexpensive, battery-powered sensing devices fitted with wireless transmitters, which can be spatially scattered to form an ad hoc hierarchically structured network. Recently, the global positioning system (GPS) facilities were embedded into the sensor node architecture to identify its location within the operating environment. This mechanism may be exploited to extend the WSN's applications. To face with the security requirements and challenges in hierarchical WSNs, we propose a dynamic location-aware key management scheme based on the bivariate polynomial key pre-distribution, where the aggregation cluster nodes can easily find their best routing path to the base station, by containing the energy consumption, storage and computation demands in both the cluster nodes and the sensor nodes. This scheme is robust from the security point of view and able to work efficiently, despite the highly constrained nature of sensor nodes.

Keywords: Sensor Networks, Key Management, Authentication, Bivariate Polynomial Key Distribution.

1. Introduction

WSNs have been deployed in different environments, including disaster relief operations, seismic data collection, monitoring wildlife and battlefield management/military intelligence. Sensors can be installed in a variety of

environments and usually establish a wireless network infrastructure to communicate and exchange information into their operating area. The sensor node is characterized by limited computing power and hence has a low price. Due to their small size, sensors can be spatially scattered to form an ad hoc network. Therefore, WSNs require an appropriate cryptosystem to ensure secure communication and mutual trust between their component nodes. In this scenario, key management becomes an issue of paramount importance since most of the encryption-related primitives require the use and distribution of keys in their operations. The high computational cost of the strongest available techniques (e.g., Diffie-Hellman key management [1] or Rivest Shamir Adleman encryption [2]) make most of them not suitable for use in a WSN, characterized by "hardware-constrained" devices, so that the use of "plain" symmetric cryptography becomes an unavoidable choice. Furthermore, also the key dimension and the number of potentially pre-storable keys may become a significant obstacle to the deployment of strong cryptographic techniques on these tiny devices due to their limited amount of available memory. The last important issue is energy consumption, which is widely known to increase proportionally to the computing efforts [12], such as the ones required by strong cryptosystems. In 2011, Xia et al. [3] focused on addressing the energy efficiency problem in sensor networks.

Accordingly, the use of efficient, lightweight and robust symmetric encryption schemes, together with the associated management protocols, needed to establish and distribute the corresponding keys among the network nodes, assumes a fundamental importance in WSN security. The simplest WSN architectures are based on a strongly meshed, flat interconnection scheme, which is known to exhibit a limited scalability when the number of nodes grows. However, in recent years, starting from the consideration that, in most applications, the connectivity between all the sensors is not necessary, more cost and performance-effective hierarchical schemes [4] are emerging. These schemes, structured according to a multi-tier hierarchical model, allow some "cluster" nodes, characterized by a more powerful hardware equipment (storage and computing capacity, antenna power, battery duration, etc.) to aggregate and pre-process the data incoming from the inexpensive sensor nodes, by reducing the traffic load, the energy consumption as long as the number of hops needed to communicate with the base station (BS), that is in charge for the overall WSN's operations. The most common hierarchical model is two-tier, providing two classes of sensors (basic sensors and cluster aggregators), apart from the BSs.

To cope with this scenario several hierarchical key management and distribution reference schemes have been developed. In detail, Chan, Perrig and Song proposed a Random Key Pre-distribution scheme (RKP) [5], where each node randomly picks *m* keys from a large key pool, such that any two-sensor nodes will share at least one common key with a certain probability. The PIKE scheme [6] addressed the problem of high-density deployment requirements in RKP. Cheng and Agrawal proposed an improved key distribution mechanism known as IKDM [7] using the polynomial keys and easily generating the session keys between the sensor and the cluster nodes.

The cluster nodes use the authentication key to authenticate the sensor nodes. The concept of the large-scale network model underlying IKDM is helpful for us to design our key management scheme. However, the IKDM incurs in high computation cost between the cluster nodes. In this work, we leveraged on the dynamic key management mechanism (KDM) proposed in [8] and on the bivariate polynomial key scheme presented in [7] to establish the session keys and achieve mutual authentication between nodes. The resulting WSN security framework not only can solve the RKP defects (a small number of compromised nodes may expose a large fraction of common keys between the non-compromised nodes [5]), but also successfully copes with the PIKE need to use a lot of sensor memory, and reduces the computation cost of the IKDM. In addition, we also introduced in the resulting key management scheme several location-based considerations and mechanisms [9,10] involving the GPS (Global Positioning system) to identify, in presence of GPS-equipped nodes, the nearest sensors and hence the best key path to the BS on each distribution step.
The rest of this paper is organized as follows. Section 2 reviews some background prerequisites. Section 3 presents our scheme, analyzed in Section 4. Finally, conclusions are presented in Section 5.

2. Backgrounds

2.1. Polynomial Key pre-Distribution Schemes

Polynomial key pre-distribution schemes use the polynomial mathematics in order to generate key pool and perform key assignment among the involved parties. A key distribution server (KDS), performs off-line distribution of several polynomial shares of degree *k* to a set of nodes so that any *k* users are able to calculate a common key that can be used in their communications without any kind of interaction. By evaluating its own stored polynomials with the identifiers (ID) of the other (*k − 1*) parties, each node can determine a common key, independently shared with the other nodes. Blundo et al. [11] proposed a bivariate polynomial $f(x,y)$ that can be used to compute the key; the parameters (x,y) were defined as the unique ID between the sensors x and y respectively. The polynomial is defined as:

$$f(x,y) = \sum_{i,j=0}^{k} a_{ij} x^i y^j \qquad (1)$$

where the coefficients a_{ij} (0 ≤ *i* , *j* ≤ *k*) are randomly chosen from a finite Galois field *GF(Q)*; *Q* is a prime number that is large enough to accommodate a cryptographic key. The bivariate polynomial has a symmetric property like:

$$f(x,y) = f(y,x) \tag{2}$$

In our specific WSN environment each sensor has a unique ID and, as the first step of network deployment, the KDS first initializes the sensors by giving to each sensor *p* a polynomial share $g_p(y)$, which is obtained by evaluating $f(x,y)$ with $x = p$,

$$g_p(y) = f(p,y) \tag{3}$$

In other words, each sensor node *p* stores a number of *k* coefficients g_j, (0 ≤ $j \le k$) in its memory,

$$g_j = \sum_{i=0}^{k} a_{ij}(p)^i, (0 \le j \le k) \tag{4}$$

where *p* is the node ID of the sensor, and g_j is the coefficient of y^j in the polynomial $f(p,y)$.

2.2. Properties of Polynomial Key pre-Distribution Scheme

The main strength of the bivariate polynomial key pre-distribution scheme is that there is no overhead during the node-to-node pairwise key establishment activity. The main known drawback, on the other hand, is the "*K*-security" property: a *k*-degree scheme is only robust against coalitions of up to *k* compromised nodes [7]. Until the number of compromised nodes is kept lower than *k*, even if all the compromised nodes share their secret data, the unknown coefficients of the polynomial cannot be calculated. However, when more than *k* nodes are compromised, the coefficients can be determined from the combination of all the available data.

3. The Proposed Scheme

We combine the effectiveness of bivariate polynomial key management schemes with the location based ones. By assuming that the sensor nodes' position can be dynamically determined via GPS, our scheme is also able to leverage on the sensor location information to improve its overall performance, with respect to traditional location-oblivious schemes. It allows data aggregation in a fewer number of places, located on better paths to the BS, by simultaneously achieving the same connectivity and security degree, with a lower number of keys to be stored in each sensor node. The operating phases of the integrated framework are described in the following.

3.1. Notation

The following notation is used in the following.

BS : the base station

CN_i, SN_i : the *i*-th cluster node and the *i*-th sensor node, respectively

ID_{SN_i}, ID_{CN_i}, ID_{BS} : the identity of the *i*-th sensor, cluster and BS

$h(.)$: a one-way hash function

$h_{key}(.)$: a one-way hash function with key

$f(x, y)$: a bivariate polynomial, where (x, y) are defined as the unique ID between the sensors x and y

$E_k(M)$: the symmetric encryption making use of key *k* to encrypt *M*

$D_k(M)$: the symmetric decryption making use of key *k* to decrypt *M*

$X \stackrel{?}{=} Y$: determines if *X* equal to *Y*

$Seed$: seed for updating the finish message key pre-deployed in nodes

$a_{i\ i}^{h}, a_i^{h-1}$: two parameters pre-deployed in the *i*-th sensor node for the generation session key (where *h* is an integer of the hash operation)

$b_{i\ i}^{h}, b_i^{h-1}$: two parameters pre-deployed in the *i*-th cluster node for the generation session key (where *h* is an integer of the hash operation)

$SNID_{CN_i}$: the identity list of the sensors served by the cluster node CN_i

$SNKEY_{list}$: the key list of sensor dynamic keys stored to the BS

$SNKEY_{CN_i}$: the key list of the sensors' dynamic keys, generated by the *BS*

K_{SN_i} : the dynamic key of the SN_i, $K_{SN_i} = h(a_i^h, a_i^{h-1})$

K_{CN_i} : the dynamic key of the CN_i, $K_{CN_i} = h(b_i^h, b_i^{h-1})$

$K_{CN_i - CN_j}$: the polynomial session key of the CN_i and CN_j

$K_{CN_i - BS}$: the polynomial session key of the CN_i and BS

N_{CN_i} : the nonce generated by the BS for the CN_i

$MAC_{CN_i - BS}$: the message authentication code (*MAC*) for the BS to CN_i

$MAC_{BS - CN_i}$: the *MAC* for the CN_i to authenticate BS

msg_{CN_i} : the receiving message of the *i*-th cluster node from the decrypted messages of the *p* sensor nodes

msg_{start}, msg_{finish} : the start message and finish message, respectively

$msg_{location}$: the location message broadcasted by the cluster node

3.2. Initialization Phase

In this phase, the BS pre-distributes the polynomial scheme parameters to both the sensor nodes and cluster nodes.

Step 1: The BS selects a random number a_i and computes the hash chain:

$$\begin{aligned} a_i^0 &= a_i \\ a_i^1 &= h(a_i^0) \\ a_i^2 &= h(a_i^1, a_i^0) \\ &\vdots \\ a_i^h &= h(a_i^{h-1}, a_i^{h-2}), \quad (1 \le i \le m) \end{aligned} \tag{5}$$

It stores $((a_1^1, a_1^0), ..., (a_m^1, a_m^0))$ to the nodes' dynamic key list $SNKEY_{list}$

$$SNKEY_{list} = ((a_1^1, a_1^0), (a_2^1, a_2^0), ..., (a_m^1, a_m^0)), \quad (1 \le i \le m) \tag{6}$$

It then stores (ID_{SN_i} , (a_i^1, a_i^0) , $Seed$, K_{msg}) to the *i*-th sensor node.

Step 2: The BS selects a random number b_i and builds the hash chain as:

$$\begin{aligned} b_i^0 &= b_i \\ b_i^1 &= h(b_i^0) \\ b_i^2 &= h(b_i^1, b_i^0) \\ &\vdots \\ b_i^h &= h(b_i^{h-1}, b_i^{h-2}), \quad (1 \le i \le n) \end{aligned} \tag{7}$$

Then the BS stores $((b_1^1, b_1^0), (b_2^1, b_2^0), ..., (b_n^1, b_n^0))$ in the dynamic key list $CNKEY_{list}$ of the *n* cluster nodes,

$$CNKEY_{list} = ((b_1^1, b_1^0), (b_2^1, b_2^0), ..., (b_n^1, b_n^0)), \quad (1 \le i \le n) \tag{8}$$

The BS randomly selects two polynomials from the *k* ones for *n* cluster nodes, and then stores the bivariate polynomial on these nodes:

$$K_{CN_i - CN_j} = f_{CN_i - CN_j}(ID_{CN_i}, y) \tag{9}$$

$$K_{BS-CN_i} = f_{BS-CN_i}(ID_{BS}, ID_{CN_i}) \quad (10)$$

The BS selects a nonce N_{CN_i}, and then stores $(ID_{CN_i}, (b_i^1, b_i^0), N_{CN_i}, K_{CN_i-CN_j}, K_{BS-CN_i})$ to the *i*-th cluster nodes.

3.3. Location-based Routing Plan Determination

In this phase, the cluster nodes can establish the best route on the basis of the received broadcast location message in a monitoring area.

Step 1: After the Initialization phase, the sensors and cluster nodes have stored the operating parameters and then distributed the associated messages within their operating environment.

Step 2: The BS broadcasts each sensor network start message msg_{start} to the cluster nodes.

Step 3: Upon receiving the start message, the cluster node (equipped with a GPS receiver) broadcasts the message $msg_{location}$ concerning its location to the neighbor cluster nodes.

Step 4: After receiving the message $msg_{location}$, the cluster nodes know the location of the source neighbor cluster so that it can transmit the monitor data to the cluster node that is nearest to the BS.

For example, in Figure 1, the cluster node R_5 can receive the nearest distance message to the BS from the neighbor cluster nodes R_1, R_2, R_3, R_4, R_6, R_7, R_8 and R_9; It can compare the received location message to select the nearest node from the BS and establish the multi-hop routing path to the cluster node R_1. The cluster node R_1 will be used to relay communications to the BS, so the best path of the cluster node R_5 will be established as follows: $R_5 \rightarrow R_1 \rightarrow BS$.

On the basis of the shortest distance between the cluster node and the BS, each cluster node will establish the best routing path. In Figure 2, the cluster node R_9 can determine that the neighbor cluster node on the best path is R_5, and the cluster node R_5 and R_1 can determine the R_1 and BS, respectively. The best path for the cluster node R_9 can be established as follows: $R_9 \rightarrow R_5 \rightarrow R_1 \rightarrow BS$. In the same way, the cluster node R_3 can determine the best path: $R_3 \rightarrow R_2 \rightarrow R_1 \rightarrow BS$.

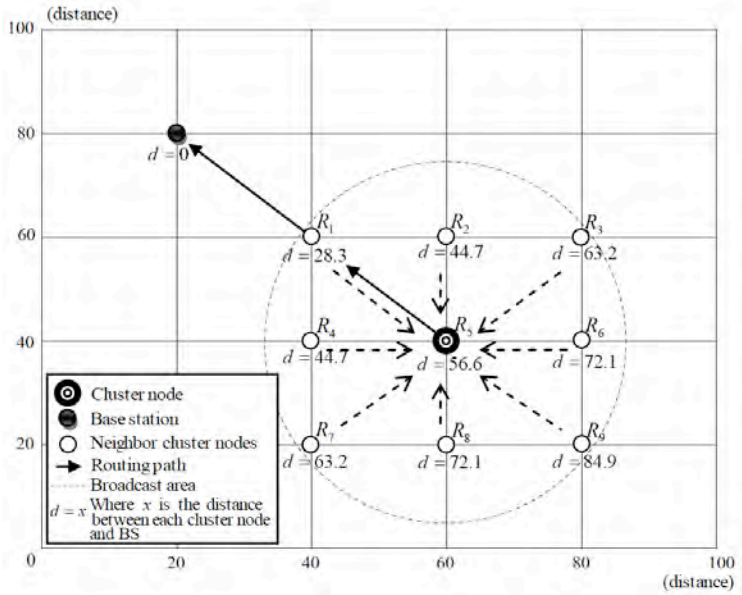


Fig. 1. Each cluster node broadcasts its location to its neighbor cluster nodes

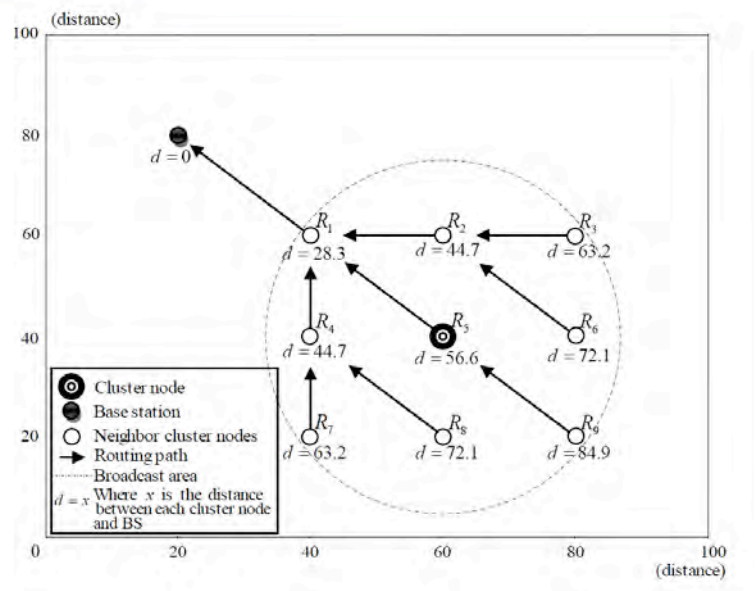


Fig. 2. Establishing the best routing overview

Every pair of nodes along the resulting multi-hop path can establish a pairwise key for encrypted communication in such a way that each intermediate node can relay data towards the BS in a totally secure way. Location awareness

also increases the probability that the geographically closest node pairs establish a pairwise session key along the best path to the BS, with the effect of saving energy on all the nodes involved in multi-hop routing.

3.4. Establishing the Polynomial Session Key Phase

The cluster nodes can use the bivariate polynomial to establish the pairwise session keys along the previously determined multi-hop paths. Each cluster node CN_i broadcasts its unique ID_{CN_i} to the cluster node CN_j and the cluster node CN_j replies with its unique ID_{CN_j} to cluster node CN_i. The cluster nodes receive the related unique ID from the neighbor cluster nodes and compute the session key as follows:

$$K_{CN_i-CN_j} = f_{CN}(ID_{CN_i}, ID_{CN_j}) \tag{11}$$

$$K_{CN_j-CN_i} = f_{CN}(ID_{CN_j}, ID_{CN_i}) \tag{12}$$

3.5. Cluster Node Requests Session Key Phase

Each cluster node, when collecting the monitoring data, receives the associated messages from its *p* sensor nodes, and then decrypts them properly. In order to do this, the cluster node needs to request the session key to the BS. The session key request scenario is shown in Figure 3.

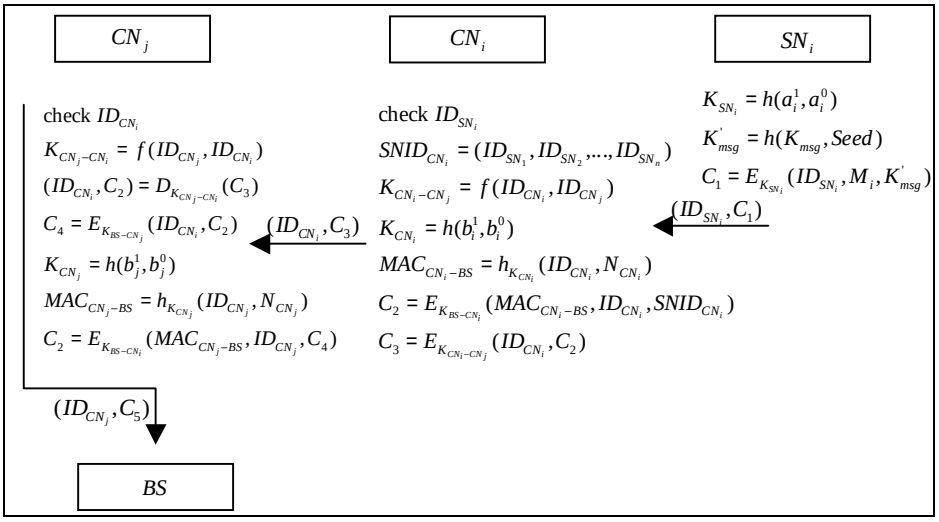


Fig. 3. Overview of the cluster node session key requests

Step 1: First, the sensor node SN_i uses the parameters (a_i^1, a_i^0) to compute the dynamic session key K_{SN_i},

$$K_{SN_i} = h(a_i^1, a_i^0) \tag{13}$$

The sensor node computes the finish message key $K_{msg}^{'}$,

$$K_{msg}^{'} = h(K_{msg}, Seed) \tag{14}$$

Then the sensor node uses the dynamic session key K_{SN_i} to encrypt $(ID_{SN_i}, M_i, K_{msg}^{'})$,

$$C_1 = E_{K_{SN_i}}(ID_{SN_i}, M_i, K_{msg}^{'}) \tag{15}$$

The sensor node then sends (ID_{SN_i}, C_1) to the cluster node.

Step 2: The cluster node CN_i checks ID_{SN_i} and collects the monitoring data of *p* sensor nodes to make an identity list $SNID_{CN_i}$,

$$SNID_{CN_i} = (ID_{SN_1}, ID_{SN_2}, ..., ID_{SN_p}) \tag{16}$$

Then the cluster node CN_i computes the session key $K_{CN_i-CN_j}$,

$$K_{CN_i-CN_j} = f(ID_{CN_i}, ID_{CN_j}) \tag{17}$$

The cluster node CN_i then uses the parameters b_i^1 and b_i^0 to compute the dynamic session key K_{CN_i},

$$K_{CN_i} = h(b_i^1, b_i^0) \tag{18}$$

The cluster node CN_i uses the ID_{CN_i} and nonce N_{CN_i} to compute the message authentication code MAC_{CN_i-BS},

$$MAC_{CN_i-BS} = h_{K_{CN_i}}(ID_{CN_i}, N_{CN_i}) \tag{19}$$

The cluster node CN_i uses the session key K_{BS-CN_i} to encrypt $(MAC_{CN_i-BS}, ID_{CN_i}, SNID_{CN_i})$,

$$C_2 = E_{K_{BS-CN_i}}(MAC_{CN_i-BS}, ID_{CN_i}, SNID_{CN_i}) \quad (20)$$

Then the cluster node CN_i uses the session key $K_{CN_i-CN_j}$ to encrypt (ID_{CN_i}, C_2),

$$C_3 = E_{K_{CN_i-CN_j}}(ID_{CN_i}, C_2) \quad (21)$$

And the cluster node CN_i sends the message (ID_{CN_i}, C_3) to CN_j.

Step 3: Upon receiving the message (ID_{CN_i}, C_3), the cluster node CN_j checks ID_{CN_i} and computes the session key $K_{CN_j-CN_i}$ to decrypt the message C_3,

$$K_{CN_j-CN_i} = f(ID_{CN_j}, ID_{CN_i}) \quad (22)$$

$$(ID_{CN_i}, C_2) = D_{K_{CN_j-CN_i}}(C_3) \quad (23)$$

Then the cluster node CN_j uses the session key K_{BS-CN_j} to encrypt the forwarding message (ID_{CN_i}, C_2) of the cluster node CN_i,

$$C_4 = E_{K_{BS-CN_j}}(ID_{CN_j}, C_2) \quad (24)$$

The cluster node CN_j computes the message authentication code MAC_{CN_j-BS},

$$MAC_{CN_j-BS} = h_{K_{CN_j}}(ID_{CN_j}, N_{CN_j}) \quad (25)$$

And then, the cluster node CN_j encrypts the message $(MAC_{CN_j-BS}, ID_{CN_j}, C_4)$,

$$C_5 = E_{K_{BS-CN_j}}(MAC_{CN_j-BS}, ID_{CN_j}, C_4) \quad (26)$$

and sends the message (ID_{CN_j}, C_5) to the BS.

3.6. Authentication Phase

In this phase, the BS authenticates the cluster nodes. Moreover, the cluster node can also authenticate the BS accordingly. The overview of the authentication phase is shown in Figure 4.

Step 1: Once receiving the message (ID_{CN_j}, C_5), the BS checks the ID_{CN_j} and decrypts C_5,

$$(MAC_{CN_j-BS}, ID_{CN_j}, C_4) = D_{K_{BS-CN_j}}(C_5) \tag{27}$$

Then it computes the message authentication code MAC'_{CN_j-BS} and checks whether or not it is if equal to MAC_{CN_j-BS} ,

$$MAC'_{CN_j-BS} = h_{K_{CN_j}}(ID_{CN_j}, N_{CN_j}) \tag{28}$$

$$MAC'_{CN_j-BS} \stackrel{?}{=} MAC_{CN_j-BS} \tag{29}$$

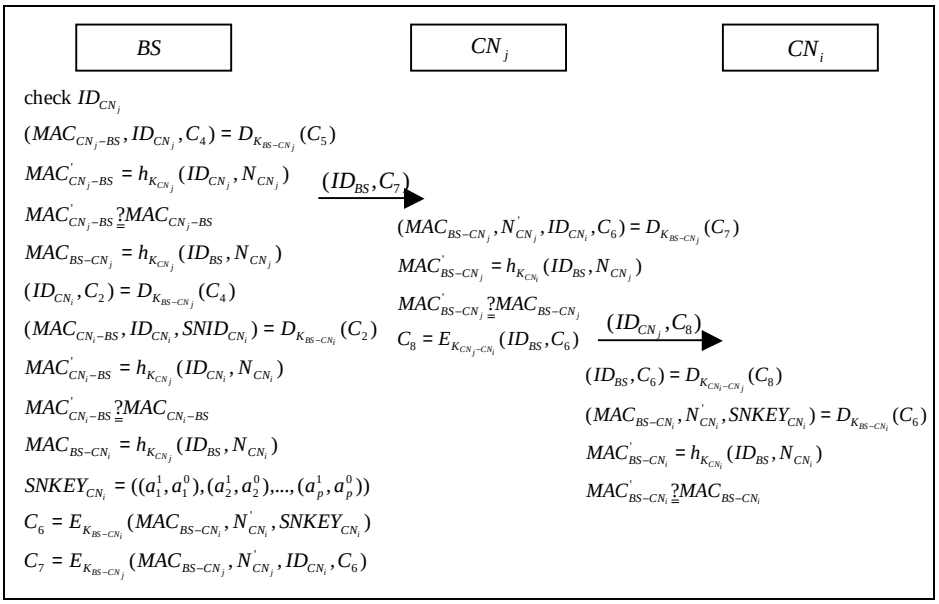


Fig. 4. Overview of the authentication phase

The BS uses the ID_{BS} and nonce N_{CN_j} to compute the message authentication code MAC_{BS-CN_j},

$$MAC_{BS-CN_j} = h_{K_{CN_j}}(ID_{BS}, N_{CN_j}) \quad (30)$$

After that, the BS decrypts the message C_4,

$$(ID_{CN_i}, C_2) = D_{K_{BS-CN_j}}(C_4) \quad (31)$$

Then it decrypts C_2, computes the message authentication code MAC'_{CN_i-BS} and checks whether or not it is equal to MAC_{CN_i-BS},

$$(MAC_{CN_i-BS}, ID_{CN_i}, SNID_{CN_i}) = D_{K_{BS-CN_i}}(C_2) \quad (32)$$

$$MAC'_{CN_i-BS} = h_{K_{CN_i}}(ID_{CN_i}, N_{CN_i}) \quad (33)$$

$$MAC'_{CN_i-BS} \stackrel{?}{=} MAC_{CN_i-BS} \quad (34)$$

Then the BS uses the ID_{BS} and the nonce N_{CN_i} to compute the message authentication code MAC_{BS-CN_i},

$$MAC_{BS-CN_i} = h_{K_{CN_i}}(ID_{BS}, N_{CN_i}) \quad (35)$$

After authentication, the BS uses the ID list $SNID_{CN_i}$ to find the dynamic key of the sensor node in the $SNKEY_{list}$; it then stores the dynamic session key to the key list $SNKEY_{CN_i}$ of the *p* members of the cluster node CN_i,

$$SNKEY_{CN_i} = ((a_1^1, a_1^0), (a_2^1, a_2^0), ..., (a_p^1, a_p^0)), \quad 1 \le i \le p \quad (36)$$

Then the station uses K_{BS-CN_i} to encrypt $(MAC_{BS-CN_i}, N'_{CN_i}, SNKEY_{CN_i})$,

$$C_6 = E_{K_{BS-CN_i}}(MAC_{BS-CN_i}, N'_{CN_i}, SNKEY_{CN_i}) \quad (37)$$

It uses the session key K_{BS-CN_j} to encrypt $(MAC_{BS-CN_j}, N'_{CN_j}, ID_{CN_i}, C_6)$,

$$C_7 = E_{K_{BS-CN_j}}(MAC_{BS-CN_j}, N'_{CN_j}, ID_{CN_i}, C_6) \quad (38)$$

The BS sends (ID_{BS}, C_7) to the cluster node CN_j.

Step 2: Upon receiving (ID_{BS}, C_7), the cluster node CN_j uses the session

key K_{BS-CN_j} to decrypt C_7,

$$(MAC_{BS-CN_j}, N'_{CN_j}, ID_{CN_i}, C_6) = D_{K_{BS-CN_j}}(C_7) \quad (39)$$

Then it computes the message authentication code MAC'_{BS-CN_j} and checks whether or not it is equal to MAC_{BS-CN_j},

$$MAC'_{BS-CN_j} = h_{K_{CN_j}}(ID_{BS}, N_{CN_j}) \quad (40)$$

$$MAC'_{BS-CN_j} \stackrel{?}{=} MAC_{BS-CN_j} \quad (41)$$

The cluster node CN_j uses the session key $K_{CN_j-CN_i}$ to encrypt the message (ID_{BS}, C_6),

$$C_8 = E_{K_{CN_j-CN_i}}(ID_{BS}, C_6) \quad (42)$$

Since the cluster node CN_j has the message ID_{CN_i}, the cluster node CN_j sends the message (ID_{CN_j}, C_8) to the cluster node CN_i.

Step 3: After receiving the message, the cluster node CN_i uses the session key $K_{CN_i-CN_j}$ to decrypt the message C_8,

$$(ID_{BS}, C_6) = D_{K_{CN_i-CN_j}}(C_8) \quad (43)$$

Then the cluster node CN_i decrypts the message C_6 by using the session key K_{BS-CN_i},

$$(MAC_{BS-CN_i}, N'_{CN_i}, SNKEY_{CN_i}) = D_{K_{BS-CN_i}}(C_6) \quad (44)$$

Then it computes the message authentication code MAC'_{BS-CN_i} and checks whether or not it is equal to MAC_{BS-CN_i},

$$MAC'_{BS-CN_i} = h_{K_{CN_i}}(ID_{BS}, N_{CN_i}) \quad (45)$$

$$MAC'_{BS-CN_i} \stackrel{?}{=} MAC_{BS-CN_i} \quad (46)$$

3.7. Dynamic Key Management Phase

The cluster node decrypts the message from the members, collects the monitor data, and sends it to the BS. When the finish message is sent out, the BS and all the sensors update the dynamic key. The overview of the dynamic key management phase is shown in Figure 5.

Step 1: After the authentication, the cluster node CN_i uses the key list $SNKEY_{CN_i}$ to find the dynamic key parameter of *p* members

$$SNKEY_{CN_i} = ((a_1^1, a_1^0),(a_2^1, a_2^0),...,(a_p^1, a_p^0)) \quad (47)$$

The cluster node CN_i gets the dynamic key K_{SN_i} of the sensor nodes; it can decrypt the message C_1 of the monitoring data,

$$K_{SN_i} = h(a_i^1, a_i^0) \quad (48)$$

$$(ID_{SN_i}, M_i, K_{msg}^{'}) = D_{K_{SN_i}}(C_1) \quad (49)$$

After that, the cluster node CN_i gets the message $(M_0, M_1,..., M_P)$ from the decrypted messages of the sensor nodes; it then stores into msg_{CN_i},

$$msg_{CN_i} = (M_0, M_1,..., M_P) \quad (50)$$

Then the cluster node CN_i encrypts $(ID_{CN_i}, msg_{CN_i}, msg_{finish})$,

$$C_9 = E_{K_{BS-CN_i}}(ID_{CN_i}, msg_{CN_i}, msg_{finish}) \quad (51)$$

The cluster node CN_i uses the finish message key $K_{msg}^{'}$ to encrypt the finish message msg_{finish},

$$C_5 = E_{K_{msg}^{'}}(msg_{finish}) \quad (52)$$

The cluster node then sends (ID_{CN_i}, C_9) and (ID_{CN_i}, C_{10}) to the BS and sensor nodes, respectively. The cluster node updates the N_{CN_i} and the dynamic session key K_{CN_i},

$$N_{CN_i} = N_{CN_i}^{'} \quad (53)$$

$$K_{CN_i}^{'} = h(K_{CN_i}, b_i^1), \quad (\text{where } K_{CN_i} = b_i^2) \quad (54)$$

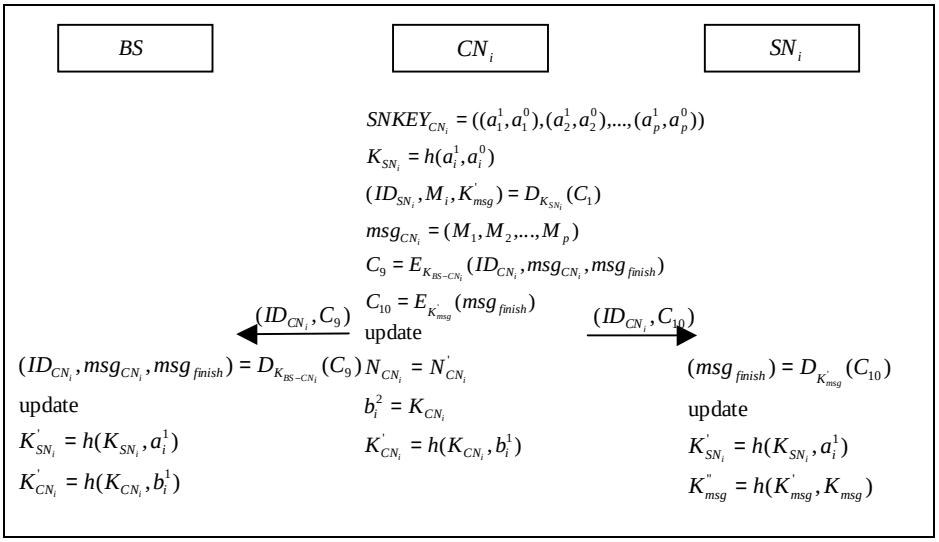


Fig. 5. Overview of the dynamic key management phase

Step 2: Once receiving the message, the BS checks the ID_{CN_i} and decrypts the message C_9 ,

$$(ID_{CN_i}, msg_{CN_i}, msg_{finish}) = D_{K_{BS-CN_i}}(C_9) \tag{55}$$

The BS gets the messages msg_{SN_i} of the sensor nodes, and the finish message msg_{finish}; then the BS updates the dynamic keys K'_{SN_i} and K'_{CN_i} of the sensor node and the cluster node, respectively.

$$K'_{SN_i} = h(K_{SN_i}, a_i^1), \quad (\text{where } K_{SN_i} = a_i^2) \tag{56}$$

$$K'_{CN_i} = h(K_{CN_i}, b_i^1), \quad (\text{where } K_{CN_i} = b_i^2) \tag{57}$$

Step 3: The sensor node SN_i decrypts C_{10} and gets the message msg_{finish} ,

$$msg_{finish} = D_{K'_{msg}}(C_{10}) \tag{58}$$

it updates the dynamic session key K'_{SN_i} and the finish message key K''_{msg}

$$K'_{SN_i} = h(K_{SN_i}, a_i^1), \quad (K_{SN_i} = a_i^2) \tag{59}$$

$$K''_{msg} = h(K'_{msg}, K_{msg}) \tag{60}$$

4. Analysis and Discussion

4.1. Sensor Network Resilience

Since the cluster nodes have the GPS location capability, when the BS sends the start message to the cluster node, the cluster node can find its location, broadcasts the location message to the neighbor cluster nodes and finds the nearest cluster node of the BS. Therefore, if the cluster nodes are compromised or the sensors are at low-energy, the cluster node broadcasts to the new neighbor cluster nodes that are nearest to the BS. In this way, our scheme achieves the network resilience.

4.2. Resistance to Sensor Node Capture Attack

Node capture attack is a serious threat in WSNs deployed in hostile environments. Due to their hardware limitations the nodes usually are not tamper-resistant and hence any adversary that captures a sensor can easily extract its stored secret data to break the underlying security scheme. In our scheme, because the coefficients a_{ij} ($0 \le i, j \le k$) are randomly chosen from a finite *GF(Q)*, where *Q* is a prime number that is large enough to accommodate a cryptographic key, each cluster node pair has a unique pairwise session key $K_{CN_i-CN_j}$ and $K_{CN_j-CN_i}$, so the security cannot be compromised between cluster nodes. if one of them is compromised. Only a legal ID pair ID_{CN_i} and ID_{CN_j} can compute the right session keys:

$$K_{CN_i-CN_j} = f_{CN}(ID_{CN_i}, ID_{CN_j}) \tag{61}$$

$$K_{CN_j-CN_i} = f_{CN}(ID_{CN_j}, ID_{CN_i}) \tag{62}$$

If we use a *k*-degree bivariate polynomial our scheme is guaranteed to be (k + 1)-secure. That is, no less than (k + 1) nodes holding polynomial shares have to be captured in order to reconstruct it.

4.3. Mutual Authentication

We can consider two fundamental cases:

(1) The BS authenticates the *i*-th cluster node

The cluster node uses the dynamic session key to compute the message authentication code MAC_{CN_i-BS} of the N_{CN_i}; it then sends it to the BS.

$$MAC_{CN_i-BS} = h_{K_{CN_i}}(ID_{CN_i}, N_{CN_i}) \quad (63)$$

Since the BS received the message authentication code MAC_{CN_i-BS} from the cluster node CN_i, it can compute the message authentication code MAC'_{CN_i-BS} and checks whether or not it is equal to MAC_{CN_i-BS},

$$MAC'_{CN_i-BS} = h_{K_{CN_i}}(ID_{CN_i}, N_{CN_i}) \quad (64)$$

$$MAC'_{CN_i-BS} \stackrel{?}{=} MAC_{CN_i-BS} \quad (65)$$

(2) The *i*-th cluster node authenticates the BS

The BS uses the dynamic session key to compute the message authentication code MAC_{BS-CN_i} of N_{CN_i}, and sends it to the cluster node,

$$MAC_{BS-CN_i} = h_{K_{CN_i}}(ID_{BS}, N_{CN_i}) \quad (66)$$

Upon receiving the message authentication message MAC_{BS-CN_i}, the cluster node computes the message authentication code MAC'_{BS-CN_i} and checks whether or not it is equal to MAC_{BS-CN_i},

$$MAC'_{BS-CN_i} = h_{K_{CN_i}}(ID_{BS}, N_{CN_i}) \quad (67)$$

$$MAC'_{BS-CN_i} \stackrel{?}{=} MAC_{BS-CN_i} \quad (68)$$

After authentication, the BS selects a new nonce N'_{CN_i} and sends it to the cluster node N_{CN_i}; the cluster node N_{CN_i} updates the nonce after sending the monitoring data back to the BS. Because the nonce N_{CN_i} and the session key K_{CN_i} are updated in each session, our scheme achieves the mutual authentication between cluster node and BS.

4.4. Dynamic Key Management

In the dynamic key management phase, the session keys of the sensor and cluster nodes are generated as follows:

$$K'_{SN_i} = h(K_{SN_i}, a_i^1), \quad (\text{where } K_{SN_i} = a_i^2) \quad (69)$$

$$K'_{CN_i} = h(K_{CN_i}, b_i^1), \quad (\text{where } K_{CN_i} = b_i^2) \tag{70}$$

As we mentioned, the next parameters of the hash seed (K_{SN_i}, a_i^1) and (K_{CN_i}, b_i^1) are updated in each session.

Our scheme can solve the replay problem in random pairwise keys pre-distribution. It only needs two pre-stored parameters to produce the session key by using the hash function. When the information transmission is finished, the cluster and the sensor nodes should update the session key and prevent the replay attack after each session.

4.5. Discussion

The proposed scheme supports mutual authentication and fully dynamic key agreement. It is worth mentioning that it embeds the polynomial key function and the GPS location capability in cluster nodes. Each cluster node can leverage on GPS information to find out its best path to the BS, and the polynomial key function can be easily used to create the cluster node session keys $K_{CN_i-CN_j}$ necessary for encrypted communications between adjacent cluster node pairs (*i, j*) along this path. If any of these cluster node gets corrupted, the other ones can use the broadcast message msg_{start} to find out their new best path to the BS. As Table 1 shows that our scheme is superior to other related works.

Table 1. Comparison of the proposed scheme with the most significant related works

Protocol	Our scheme	KMTD[8]	IKDM[7]
Captured attack analysis	Yes	Yes	Yes
Detail security analysis	Yes	Yes	Partial (captured attack)
Stored cost (Cluster node)	One session key, two polynomial function	Two session keys, one base station ID	One session key, two polynomial function
Stored cost (Sensor node)	Two session keys, one cluster node ID	Two session keys, one cluster node ID	Two session keys, one base station ID
Sensor network model	Hierarchical	Hierarchical	Hierarchical
Sensor's homogeneity	Hierarchical	Hierarchical	Homogeneous
Mutual authentication	Yes	Yes	N/A
Dynamic key agreement	Yes	Yes	N/A
GPS capability	Yes (cluster node)	N/A	N/A
Routing protocol	Yes	N/A	N/A

5. Conclusions

In this paper, we proposed a dynamic key management scheme. Our scheme can achieve the following goals:

(1) We provide a dynamic key management to prevent the replay attack.
(2) We use the GPS technology to find the nearest node of the BS to the neighbor cluster nodes.
(3) We proposed a nonce-based mechanism to complete the mutual authentication between the BS and cluster nodes. It can enhance the information security.
(4) We coped with the storage and energy consumption limitations and reduced the computation cost of the sensors.

In the future, we envision that our scheme could be extended to apply polynomial key techniques in different WSNs for more efficient transmission combined with energy control mechanism or for implementing alternate key management strategies.

6. References

1. Diffie, W., Hellman, M.E.: New Directions in Cryptography. IEEE Transactions on Information Theory, Vol. 22, No. 6, 644-654. (1976)
2. Rivest, R.L., Shamir, A. Adleman, L.: A Method for Obtaining Digital Signatures and Public-key Cryptosystems. Communications of the ACM, Vol.21, No. 2, 120-126. (1978)
3. Xia, F., Yang, X., Liu, H., Zhang, D., Zhao, W.: Energy-efficient Opportunistic Localization with Indoor Wireless Sensor Networks. Computer Science and Information Systems, Vol. 8, No. 4, 973-990. (2011)
4. Martin, K.M., Paterson, M.: An Application-Oriented Framework for Wireless Sensor Network Key Establishment. Electronic Notes in Theoretical Computer Science, Vol. 192, No. 2, 31–41. (2008)
5. Chan, H., Perrig, A., Song, D.: Random Key Predistribution Schemes for Sensor Networks. 03 Proceedings of the 2003 IEEE Symposium on Security and Privacy, IEEE Computer Society Washington, DC, 11-14 May,2003, USA. 197-213. (2003)
6. Sheu, J.P., Cheng, J.C.: Pair-wise Path Key Establishment in Wireless Sensor Networks. Computer Communications, Vol. 30, No. 11-12, 2365-2374. (2007)
7. Cheng, Y., Agrawal, D.P.: An Improved Key Distribution Mechanism for Large-Scale Hierarchical Wireless Sensor Networks. Ad Hoc Networks, Vol. 5, No. 1, 35–48. (2007)
8. Chen, C.L., Tsai, Y.T., Shih, T.F.: A Novel Key Management of Two-tier Dissemination for Wireless Sensor Network. 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous

Computing (IMIS 2012), Palermo, Italy. 576-579. (2012)
9. Qian, Q., Shen, X., Chen, H.: An Improved Node Localization Algorithm based on DV-Hop for Wireless Sensor Networks. Computer Science and Information Systems, Vol. 8, No. 4, 953-972. (2011)
10. Wang, X., Ma, J., Wang, S., Bi, D.: Distributed Energy Optimization for Target Tracking in Wireless Sensor Networks. IEEE Transactions on Mobile Computing, Vol. 9, No. 1, 73-86. (2010)
11. Blundo, C., De Santis, A., Herzberg, A., Kutten, S., Vaccaro, U., Yung, M.: Perfectly-Secure Key Distribution for Dynamic Conferences. Journal Information and Computation, Vol. 146, No. 1, 1-23. (1998)
12. Palmieri, F.; Ricciardi, S.; Fiore, U.: Evaluating Network-Based DoS Attacks under the Energy Consumption Perspective: New Security Issues in the Coming Green ICT Area," International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), 2011 pp.374-379. (2011)

Chin Ling Chen, PhD, is a member of the Chinese Association for Information Security. From 1979 to 2005, he was a senior engineer at the Chunghwa Telecom Co., Ltd. He is currently a professor of the Department of Computer Science and Information Engineering at Chaoyang University of Technology, Taiwan. His research interests include cryptography, network security and electronic commerce. Dr. Chen had published over 50 SCI/SSCI articles on the above research fields in international journals.

Yu Ting Tsai, born in 1988. Currently he is pursuing his master's degree at the Department of Computer Science and Information, Chaoyang University of Technology. His research interests include WSN and cryptography.

Aniello Castiglione, PhD, joined the Computer Science department ``R. M. Capocelli'' of University of Salerno, Italy, in 2006. He is an active member of IEEE, ACM and IISFA (International Information System Forensics Association). His research interests include Communication Networks, Digital Forensics, Security and Privacy, Security Standards and Cryptography.

Francesco Palmieri, PhD, is an assistant professor at the Engineering Faculty of the Second University of Napoli, Italy. His research interests concern high performance networking protocols and architectures, routing algorithms and network security. He has been closely involved with the development of the Internet in Italy as a senior member of the Technical-Scientific Advisory Committee and of the CSIRT of the Italian NREN GARR.

Received: September 07, 2012; Accepted: March 05, 2013

DOI: 10.2298/CSIS121108023Z

Evaluation on the Influence of Internet Prefix Hijacking Events

Jinjing Zhao[1,2], Yan Wen[1,2]

[1]Beijing Institute of System Engineering,Beijing, China
misszhaojinjing@hotmail.com

[2]National Key Laboratory of Science and Technology
on Information System Security,Beijing, China
celestialwy@gmail.com

Abstract. The inter-domain routing system based on the BGP protocol is a kernel establishment in the Internet. There have been many incidents of IP prefix hijacking by BGP protocol in the Internet. Attacks may hijack victim's address space to disrupt network services or perpetrate malicious activities such as spamming and DoS attacks without disclosing identity. The relation between prefix hijacking and the Internet hierarchy is presented in this paper. The Internet is classified into three tiers based on the power-law and commercial relations of autonomous systems. The relation between network topology and prefix hijacking influence is presented for all sorts of hijacking events in different layers. The results assert that the hierarchical nature of network influences the prefix hijacking greatly.

Keywords: IP prefix hijacking; Power law; BGP; Inter-domain routing system; Internet Service Providers

1. Introduction

The inter-domain routing system based on the BGP protocol is a kernel establishment in the Internet. It is not only the basic mechanism of exchanging the reachable information, but also the key way to inter-connect the autonomous systems and establish the policy control in ISPs. Unfortunately, the limited guarantees provided by BGP sometimes contribute to serious instabilities and outages. Prefix hijacking are probably the most straightforward type of BGP attack.

Prefix hijacking is also known as BGP hijacking, because to receive traffic destined to hijacked IP addresses, the attacker has to make those IP addresses known to other parts of the Internet by announcing them through BGP. Because there is no authentication mechanism used in BGP, a mis-behaving router can announce routes to any destination prefix on the Internet and even manipulate route attributes in the routing updates it sends to

neighboring routers. Taking advantage of this weakness has become the fundamental mechanism for constructing prefix hijack attacks. They occur when an AS announces a route that it does not have, or when an AS originates a prefix that it does not own. In the recent past, there have been many instances of prefix hijacking in the Internet wherein an Autonomous System hijacks routes simply by advertising the corresponding prefixes [1]. On January 22, 2006, a network (AS-27506) wrongly announced the IP prefix 65.173.134.0/24 representing an address block of 224 IP addresses, into the global routing system. This prefix belonged to another network (AS-19758) and because routers do not have a means to accurately verify the legitimate origin of each prefix, they accepted announcements from both the true origin (AS-19758) and the false one (AS-27506), and selected one of them based on the local routing policies and other criteria. As a result, some networks sent for data traffic destined to 65.173.134.0/24, to AS-27506 instead of the true owner.

Irrespective of whether it is caused by a misconfiguration or a malicious entity, the AS that hijacks a prefix can blackhole and intercept all the hijacked traffic and thus, cause a denial-of-service attack or a man-in-the-middle attack against the prefix owner [2, 3]. Because the current BGP protocol implements little authentication and often assumes a high level of trust to its neighbor routers, IP hijacking can easily succeed.

Previous efforts on prefix hijacking are presented from two aspects: hijack prevention and hijack detection. Generally speaking, prefix hijack prevention solutions are based on cryptographic authentications [4-8] where BGP routers sign and verify the origin AS and AS path of each prefix. While hijack detection mechanisms [9-15] are provided when a prefix hijack is going to happen which correction steps must follow.

Because there is a lack of a general understanding on the impact of a successful prefix hijack, it is difficult to assess the overall damage once an attack occurs, and to provide guidance to network operators on how to prevent the damage. Ballani et. al. [16] presents a study of Internet prefix hijacking and interception, which analyzes the probability of an AS hijacking the traffic to a prefix from another AS and the proposal of the prefix interception. Lad et. al. [17] estimate the resilience of Prefix hijacks through simulation across the Internet's AS-level topology.

In this paper, we conduct a systematic study on the impaction of prefix hijacks launched at different position in the Internet hierarchy. The Internet is classified into three hierarchies—core layer, forwarding layer and marginal layer based on the power-law and commercial relations between autonomous systems (ASes). Two impaction parameters—affected ASes set Nc and affected paths factor μ, are analyzed for typical nine types of prefix hijacking events in different layers.

The remainder of this paper is organized as follows: The section 2 descripts the hierarchy model of the Internet based on the power-law and relationships between ASes. Based on section 2, section 3 builds the attack model of IP prefix hijacks on a comprehensive attack taxonomy relying on the Internet hierarchy model and BGP protocol policies. The impaction

analysis of the prefix hijacks attack is also presented. The related works are discussed in section 4 and section 5 concludes the paper.

2. Internet Hierarchical Model

2.1. Internet Topology

The internet structure has been the subject of many recent works. Researchers have looked at various features of the Internet graph, and proposed theoretical models to describe its evolvement. Faloutsos et al. [18] experimentally discovered that the degree distribution of the Internet AS and router level graphs obey a power law. Opposite to negative exponential distribution, the curve of ASes' degree is declining very smoothly and heavy-tailed. So large numbers of ASes have little connections, but the ASes with rich connections are quite a few. The Internet AS structure is shown to have a core in the middle and many tendrils connected to it. A more detailed description is that around the core there are several rings of nodes all have tendrils of varying length attached to them. The average node degree decreases as one moving away from the core. We call these core nodes "hub" nodes, whose degree is very high. As a "storage-forwarding" network, the node degree is an important merit for evaluating a node's forwarding ability.

In order to consider the influence on the inter-domain system of the power-law nature, we classify the nodes by its forwarding ability. We build our model based on the traditional Transit-Stub model and also consider the power-law nature of the Internet. The Transit-Stub model [19] classifies ASes into two types, transit nodes and stub nodes. The transit nodes have the routing ability, but the stub nodes haven't. The transit nodes are interconnected into a core of the network, and the stub nodes connect to the core around. In order to emphasize the importance of the power-law, we classify the transit nodes into two kinds, hub nodes and middle nodes. So the whole inter-domain system can be classified into three layers, the core layer (hub nodes), the forwarding layer (middle nodes) and the marginal layer (stub nodes). We call this model power law-hierarchy model.

The power-law and hierarchy of the Internet fits each other very well. Generally, the nodes in the core layer have rich connections and the lots of the nodes in the marginal layer have few connections which need not forwarding for other nodes. The environment in the forwarding layer is more complex, but the node degrees are also decreasing with the hierarchy increased. A few nodes with high degree makeup the core layer, and a large numbers of node with few connections on the margin form the marginal layer, and between them is the most complex layer- forwarding layer.

2.2. Model Establishment

The main consideration of the power law-hierarchy model is to distinguish the performance of different layer nodes in the BGP convergence process. We build our model according to the hierarchy of the inter-domain system, and then, we testify it with the power-law rules. If the result is right, then the hierarchical model is also a power law-hierarchy model.

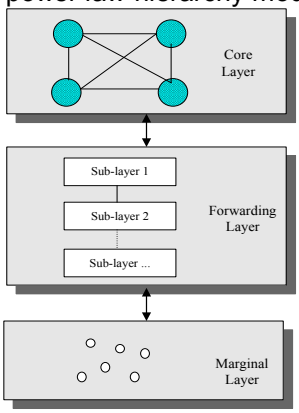


Fig. 1. Three Tiers Model Literature of Internet. a)The set of nodes who have no providers forms a clique is the core layer.b) If the nodes don't forward data for others, then it belongs to the marginal layer.c) The node that belongs to neither the core layer nor the marginal layer belongs to the forwarding layer.

The work in [20] presents a hierarchical formalization method for Internet. In [21], a five-hierarchy model of the Internet is presented based on the commercial relation between ASes. These models are too complex to analyze for BGP convergence. In [22], we build a three-hierarchy model of the Internet and give an efficient arithmetic for it. The model is organized as follows:

a) The set of nodes who have no providers forms a clique (interconnection structure), which is the core layer.

b) If the nodes don't forward data for others, then it belongs to the marginal layer.

c) The node that belongs to neither the core layer nor the marginal layer belongs to the forwarding layer. And the forwarding layer has several sub-layers.

By analysis on the number of nodes and connections of different layers which drawn from the route table data of RouteViews Project from 2005 to 2010, we can see that:

a) The average node degree of the core layer is about 880, however the one of the forwarding layer is about 7.4, and for the marginal layer, it's only 1.12, so the power-law is obeyed.

b) The proportion of the nodes number between forwarding layer and marginal layer is steady, which is about 1/6.

c) The number of the interconnections in the margin layer is zero, which means that its peer connections couldn't be observed by upper layers.

Table 1. Statistics of Connections of Layers.

	Core Layer	Forwarding Layer	Marginal Layer
Core Layer	78	2992	8374
Forwarding Layer	2992	7310	17699
Marginal Layer	8374	17699	0

In this way, we build the power law-hierarchy model of the inter-domain system based on the commercial relations between ASes, which also obey the heavy-tailed rule of power-law.

3. Prefix Hijacking Attack Model

3.1. Model Description

Prefix-hijacking occurs when a malicious or misconfigured AS announces to its peers that a block of IP-address space belongs to themselves, when, in fact, it does not. After a short delay, routes based on this bad announcement propagate through the internet at large and the malicious AS may be able to send and receive traffic using addresses it does not own. This hijacked space can be - and has been - used to send unsolicited mass e-mails, download copyrighted works, launch break-in attempts, or anything else generally considered to be illegitimate network use.

Prefix hijacking can happen in one of three ways - a block containing unallocated space can be announced, a sub-block of an existing allocation can be announced, or a competing announcement for exactly the same space as an existing allocation can be announced. Upon receiving these fabricated advertisements, other BGP routers may be fooled into thinking that a better or more specific route has become available towards the target prefix and start forwarding future traffic along the false path. As a result of the prefix hijacking, part (if not all) of the traffic addressed to the target prefix will be forwarded to the attacker instead of the target prefix.

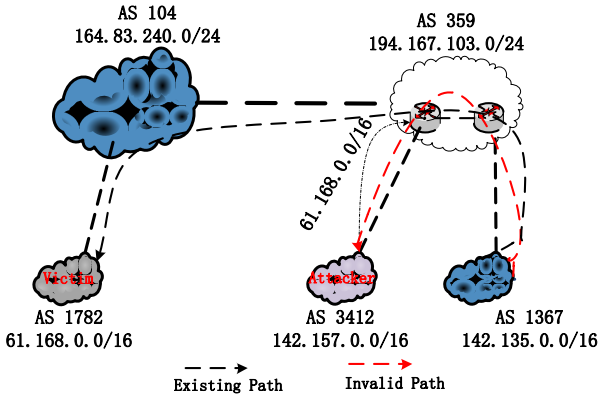


Fig 2. AS 3412 hijacks the prefix of AS 1782. AS 3412 has the IP prefix of 142.157.0.0/16 and AS 1782 has the IP prefix of 61.168.0.0/16. But AS 3412 in order to hijack the flow of AS 1782, it announces a BGP Update that it has the IP prefix 61.168.0.0/16. AS 1367 changes its path to the destination prefix 61.168.0.0/16, and its data to AS 1782 would be transmitted to AS 3412.

In Fig 2, Obviously, AS 1367's choice depends on both its existing route and the newly-received invalid route to 61.168.0.0/16. On the other hand, the success of the hijacking is also relies on its BGP routing policies. An AS will pick the shortest path to the destination in most cases, but the selected path must be valley-free and have no loops.

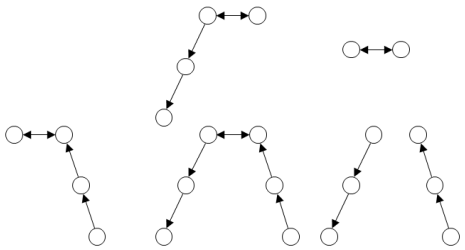


Fig 3. The Valley-free Path in BGP Policy. A valley-free path is a path has zero or several customer-provider sequences followed by one or zero peer-peer edge and then followed by zero or several provider-customer sequences.

Ballani et. al. [16] illustrated the influence of AS commercial relations between the prefix hijack path selections. Nine cases are analyzed according to different types of the existing paths and the hijacking invalid paths. Measurement studies in the past have shown that a large majority of ASes on the Internet tend to assign higher local-preference values to customer-routes than to peer-routes than to provider-routes. Since local-preference values are the first step of the BGP decision process, ASes prefer customer routes to peer routes to provider routes. In this paper, for the simpleness of our analysis, the ASes would not change their routes if the existing paths are the

same type and the same length as the hijacking invalid paths. The hijacking cases of different types if the existing paths and the invalid paths are summarized in Table 2

Table 2. Statistics of Connections of Layers.

Invalid Path / Existing path	Length	Customer	Peer	Provider
Customer	≤n	●	●	●
	>n	●	●	●
Peer	≤n	○	●	●
	>n	○	○	●
Provider	≤n	○	○	●
	>n	○	○	○

We model the Internet as a directed graph G= (V, E), nodes V represent the set of ASs in the Internet, links E are connections between them; a link e= (u, v) exists if node u will send update packets to v (but not vice versa). r = { v1,…,vk }is a simple path in G, for $\forall$ 1≦i, j≦k and if i≠j, then vi ≠vj and e=(vi, vj) ∈E, the length of r is |r|=k. G is classified into three hierarchies according to the power law-hierarchy model in section 2, the core layer C, forwarding lay and the marginal layer S.

Definition 1 If vj is a provider of vi, then vj∈provider (vi), by the same token, customer(vi) and peer(vi) can be defined.

Definition2 Function h (vi) presents the layer level of vi, 1≤h(vi)≤3(1≤ i ≤n), n is the number of nodes.

vi belongs to the core layer,

iff $\{h(v_i)=1 \mid \forall v_j \notin \text{provider}(v_i),\ 1\le i,j\le n\}$

vi belongs to the marginal layer,

iff $\{h(v_i)=3 \mid \forall v_j \notin \text{customer}(v_i),\ 1\le i,j\le n\}$

vi belongs to the forwarding layer,

iff $\{h(v_i)=2 \mid v_i \notin C, \text{且} v_i \notin S, 1\le i\le n\}$

Definition3 Function l(ej)presents the layer level of ej=(uk,um), 1≤l(ej)≤6 (1≤ j ≤m), m is the number of edges.

l(ej)=1, iff uk∈C and um∈C.

l(ej)=2, iff uk∈C and um∈T.

l(ej)=3, iff uk∈C and um∈S.

l(ej)=4, iff uk∈T and um∈T.

l(ej)=5, iff uk∈T and um∈S.

l(ej)=6, iff uk∈S and um∈S.

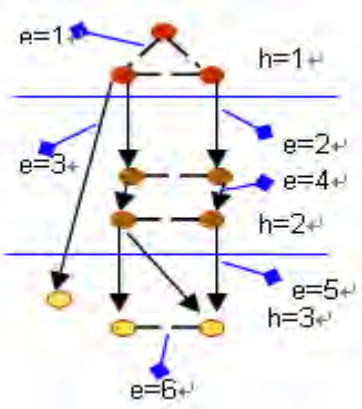


Fig 4. Hierarchy of the Nodes and Edges

To evaluate the influence if prefix hijacking events, two impaction parameters are introduced as follows:

Definition 4 Set of the affected nodes Nc: The set of nodes whose routing states might be changing because of the happening prefix hijacking event.

Definition 5 Affected path factor μ: The percentage of the paths might be changed because of the happening prefix hijacking event.

The affected path factor μ can be presented by an important parameter in graph theory, the betweenness centrality (BC) of a node.

Definition 6: Node BC $g(v)$: the BC of node v in the network is defined as: $g(v) = \sum_{s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$.

Where σ_{st} is the number of shortest paths going from s to t and σ_{st} (v) is the number of shortest paths going from s to t and passing through e. BC gives in transport networks an estimate of the traffic handled by the vertices.

BGP does not always use the shortest path between two ASes however. And the affected paths factor μ is depend on importance of the path in the network. Because of this we use a definition of path betweenness:

Definition 7: Path BC $p(v)$: the path BC of node v in the network is defined as: $p(v) = \sum_{s \neq t \in N} Path_{st}(v)$.

Where pathst(p) is the number of BGP paths between IP blocks in s and t that use path p, N is the set of the nodes in the network.

3.2. Classification of Hijack

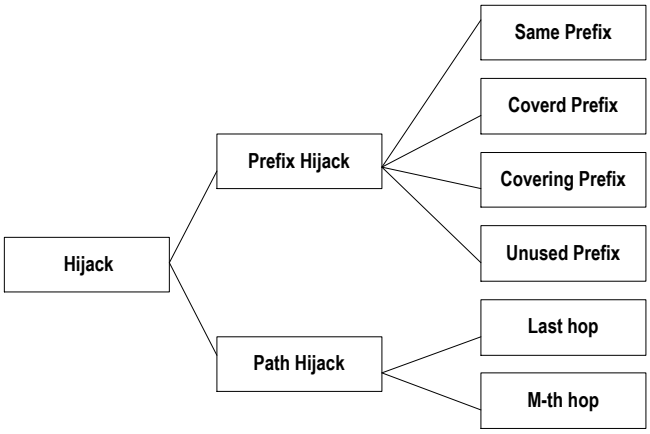


Fig 5. Types of prefix hijacks

The Hijack can be classified into two types by the way the attacker adopted, one is the node hijack and the other is the path hijack. The prefix hijack is done when an attacker announces an IP prefix which isn't belonging to him. And the path hijack is the way that the attacker announces a valid prefix, but reports an invalid path to a victim origin.

There are four types in the prefix hijack. An AS may pretend to be the owner of the same prefix of other's and originate the prefix resulting in a false origin hijack. If the attacker announces a sub-prefix of some valid prefix, termed as a covered prefix hijack, then routers in the Internet may contain routes to both the victim AS's prefix as well as the attacker's prefix. However, if the destination IP of a packet being routed, falls under the attacker's prefix space, then due to longest prefix match, the data would be forwarded to the attacker. An attacker AS may also announce a less specific prefix than a valid prefix, termed as a covering prefix hijack but will receive traffic, only when the route to the valid prefix is withdrawn. Finally, an AS may announce an invalid prefix that does not conflict with any used prefix space. For example, spammers are known to use unused prefixes for spam purpose.

To the path hijack, we separate the case of false last hop from false information on any other m-th hop in the path. The last hop hijack means that the hijacking AS announces a direct connection to the victim AS which is not existed. And m-th hop hijack is happened when the hijacking AS announces an m hops path to the victim AS, and the existing path is perhaps much longer than m or it even does not exist indeed.

The prefix hijacking events are illustrated in this paper, while the last two prefix hijacking types are not discussed. For the influence of hijacking a covering prefix hijack is the same as the hijacking the AS's prefix when the route to it is withdrawn. So the analysis to this type can be referenced to the same prefix hijack type. And the impact of the unused prefix hijack is not

determined by the hijacking event, but the activities after the hijacking, which is not focused in this paper.

3.3. Model Construction

The systematic study on the impaction of prefix hijacks launched at different position in the Internet hierarchy is descripted in this part, after the Internet hierarchy model and the prefix hijacking type are cleared.

For the simpleness of the description, the ASes whose prefixes being hijacked are expresses with V, and the hijack attack ASes are denoted by A. Furthermore, we suppose each AS only has one provider. The multi-home mechanism is not considered in this paper.

Firstly, the same prefix hijacking events are analyzed.

1. V ∈ *Core Layer*

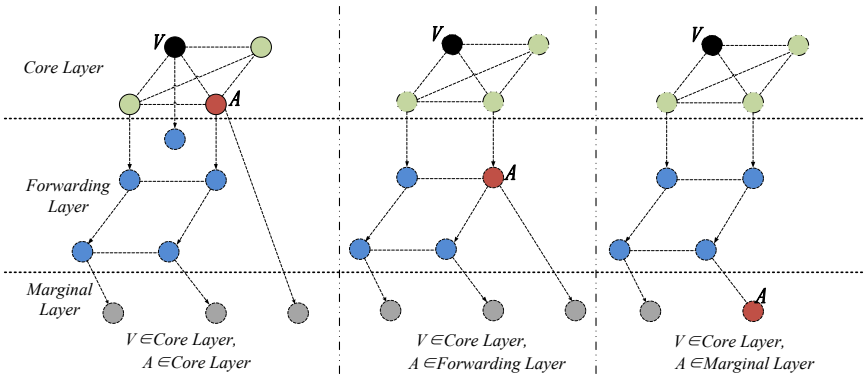


Fig 6. Hijacked AS in the Core Layer

1) A ∈ Core Layer

The hijacking and hijacked AS are both in the core layer.

Analysis:

All the ASes is the core layer are direct neighbors and they are peer nodes to each other. So, when the hijacking node A is trying to announce the same prefix of V to its neighbors, V will drop the update packet directly, and other ASes may find that they has a path to the prefix as {V} in their routing tables and they choose to ignore the announcement from A.

On the other hand, A will announce the same prefix to its customers. Its customer will accept the information and change their routing path to the prefix from {A,V} to {A}, and update the information to their customers. The update events will go on until the bottom nodes of the network who have the routes to the hijacking prefix receiving the update packets.

Parameters:

The set of the affected nodes Nc is:

$$Nc = \{A\} \cup \{\dot{Customer}(A))\}$$

2) A∈Forwarding Layer

The hijacked AS is in the core layer and the hijacking AS is in the forwarding layer.

Analysis:

If A hijacks the same prefix of V, it will announce the prefix to its providers, peers and customers. Because V is at the core layer, the routing information to it of the nodes in the forwarding layer must be received from its providers or peers before the hijacking event happening.

If A is the direct customer of an AS in core layer, its provider will ignore this announcement, because of the neighborhood between it and V. Otherwise, the provider of A will change their routing path to V, because the customer update has highest priority.

When A's peer nodes receive the announcement of hijacking prefix, they will judge where the existing route to V comes from. If the existing route is received from its providers, they will change the path to A. If the route is published by its peers, according to the rule of valley-free path, its peers can only announce the path from their customers. Because V is in the core layer, peers of A's peer could not get the path to V from their customers. So A's peer should accept he invalid path to the hijacking prefix.

When A's customers receive the announcement of hijacking prefix and the existing paths to V are coming from its providers, they will accept the update to V. If the existing paths to V are coming from its peers, they prefer to keep them, because the peer update has higher priority than the provider update.

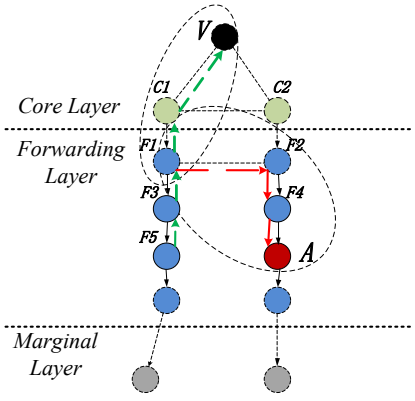


Fig 7. Existing paths compare with the invalidate paths

According to the valley-free rule of BGP path, after A's providers calculating there route to the hijacking prefix, they will announce the corresponding updates to their providers, peers and customers, while A's

peers or customers will announce updates to their customers. All the customers of A's providers or customers and their customers will changes their routing table. Whether the peers of A's n-th provider changing their routing tables or not depend on the length of their existing paths to V. In figure 7, AS F5 receives the update packets from its provider F3 to prefix which belongs to V before. The invalid path announced to F5 is {F3, F1, F2, F4, A}. And its existing path to the prefix is {F3, F1, C1, V}, which is shorter. So F5 would not accept the update event. The length of the invalid path is much correlated with A's hierarchy in the network. The higher it is, the larger hijacking impaction would be.

Parameters:

The set of the affected nodes *Nc* is in the range as:

$$\{\dot{provider}(A)\} \cup \{peer(\dot{provider}(A) \cup \{A\})\} \cup \{\dot{customer}(\{A\} \cup (\dot{provider}(A)))\} \leq Nc$$

$$\leq \{\dot{provider}(A)\} \cup \{peer(\dot{provider}(A) \cup \{A\})\} \cup \{\dot{customer}(peer(\dot{provider}(A) \cup \{A\} \cup (\dot{provider}(A))))\}$$

3) A ∈ Marginal Layer

The hijacked AS is in the core layer and the hijacking AS is in the marginal layer.

Analysis:

ASes in the marginal layer usually only have the provider relations. If A hijacks the same prefix of V, it will announce the prefix to its providers. If A is the direct customer of the core layer, its provider will ignore this announcement. Otherwise, the provider of A will change their routing path to V, and announce the update to its providers, peers and customers.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\{\dot{provider}(A)\} \cup \{peer(\dot{provider}(A) \cup \{A\})\} \cup \{\dot{customer}(\dot{provider}(A)\} \leq Nc \leq \{\dot{provider}(A)\} \cup \{peer(\dot{provider}(A))\} \cup \{\dot{customer}(peer(\dot{provider}(A)) \cup (\dot{provider}(A)))\}$$

2. V ∈ Forwarding Layer

1) A ∈ Core Layer

The hijacked AS is in the forwarding layer and the hijacking AS is in the core layer.

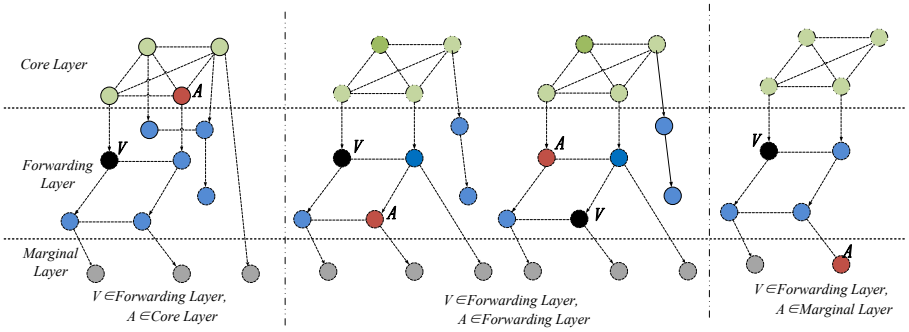


Fig 8. Hijacked AS in the Forwarding Layer

Analysis:

A announces the hijacked prefix of V to its peers. All the peers except the n-th provider of V in the core layer will update their path to V, because their existing path to V is announced by their peers and the invalid path is much shorter. The n-th provider of V would like to keep their existing path to V, because it came from their customers.

The peers, who accept the invalid path will update this information to their customers, withdraw the former paths and announce the new one. And their customers will do the update events to their customers. The procedure will going on. The ASes who is the peer of V's n-th providers or customers will reject, because their routes coms from the peer type update which has the high propriety than the provider type update.

Parameters:

The set of the affected nodes Nc is:

$$Nc = \{A\} \cup \{peer(A) - \{V\}\} \cup \{\dot{customer}(\{A\} \cup \{peer(A) - \{V\}\})\}$$
$$-\{peer(\dot{provider}(V) \cup \{V\} \cup \dot{customer}(V))\}$$

2) A ∈ Forwarding Layer

The hijacking and hijacked AS are both in the forwarding layer.

Analysis:

In the former case, if A hijacks the same prefix of V, it will announce the prefix to its providers, peers and customers.

There are three cases to analyze the peers of A: 1) the existing path is from their providers, the update from A has higher propriety; 2) the existing path is from their peers, the path length to V must be longer than one hop, so they will pick the path {A} to the hijacking prefix; 3) the existing path is from their customers, which is shown in Figure 9, they will keep their routing

information. If they change their routing table, their n-th customers will also do.

The customers of A will receive the withdraw update of their existing path to V through A and the new invalid path {A} to the hijacking prefix. And they do the same update events to their n-th customers.

The provider of A will change their routing path to V, because the customer update has highest priority. And the n-th provider of A will get the same decisions. But when they announce it to their peers, there are two cases, shown in Figure 9. When the peer of A's n-th providers is higher than V, they should not accept the update activity, because their existing path is announced by their customers. But when he peer of A's n-th providers is lower than V, they will update the paths.

To summarize the analysis above, the n-th providers of V would not affected by the hijacking events. To the n-th customers of V, if they has peer with other ASes they will accept the hijacking path. And if they has customer relations with other ASes and the path to A is shorter than the path to V, they will accept the hijacking path.

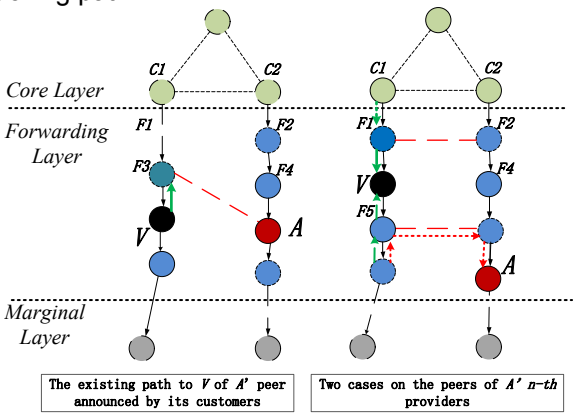


Fig 9. The hijacking and hijacked AS are both in the forwarding layer.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\{\dot{provider}(A)\} \cup \{\dot{customer}(A)\} \le Nc \le \{\dot{provider}(A)\} \cup \{\dot{peer}(\{A\} \cup \dot{provider}(A))\} \cup \{\dot{customer}(\dot{peer}(\dot{provider}(A)) \cup (\dot{provider}(A)))\}$$

3) A∈ Marginal Layer

The hijacked AS is in the forwarding layer and the hijacking AS is in the marginal layer.

Analysis:

A in the marginal layer, it can only announce the hijacking update to its provider. The provider of A will change their routing path to V, because the customer update has highest priority. And the n-th provider of A will get the same decisions. But when they announce it to their peers, there are also two cases. When the peer of A's n-th providers is higher than V, they should not accept the update. But when he peer of A's n-th providers is lower than V, they will update the paths.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\{\dot{provider}(A)\} \le Nc \le \{\dot{provider}(A)\} \cup \{peer(\dot{provider}(A))\}$$
$$\cup \{\dot{customer}(peer(\dot{provider}(A)) \cup (\dot{provider}(A)))\}$$

3. V ∈ *Marginal Layer*

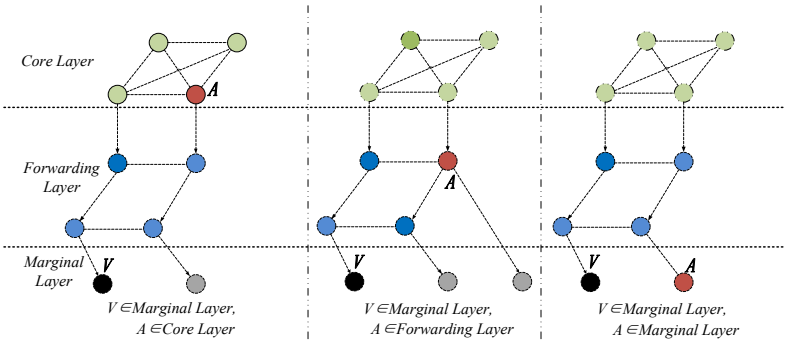


Fig 10. Hijacked AS in the Marginal Layer

1) *A ∈ Core Layer*

The hijacked AS is in the marginal layer and the hijacking AS is in the core layer.

Analysis:

A will announce the hijacking prefix to its customers and peers. Its n-th customer will accept the change unless its existing path to V is from its peers, which may be the providers of V. The peers of A except the n-th provider of V in the core layer will update their path to {A}. When they update the paths to its customers, if its customers are peers of V's n-th providers, they will reject the update. Otherwise, they will change the paths to the hijacking prefix.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\varnothing \le Nc \le \{peer(A)\} \cup \{\dot{customer}(peer(A)) \cup \{A\})\}$$

2) A∈ Forwarding Layer

The hijacked AS is in the marginal layer and the hijacking AS is in the forwarding layer.

Analysis:

When A announces the hijacking update to its providers, peers and customers, they will accept the path change to V, and send the update packets to its providers, peers and customers, unless their peers or customers are the V's n-th providers or their peers.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\{\overset{\bullet}{provider}(A)\} \le Nc \le \{\overset{\bullet}{provider}(A)\} \cup \{peer(\overset{\bullet}{provider}(A))\}$$
$$\cup \{\overset{\bullet}{customer}(peer(\overset{\bullet}{provider}(A)) \cup (\overset{\bullet}{provider}(A)))\}$$

3) A∈ Marginal Layer

The hijacking and hijacked AS are both in the marginal layer.

Analysis:

When A announces the hijacking update to its n-th providers, they will announce the updates to its peers and other customers. When they are the n-th provider of V, the hijacking attacks are successful to them.

Parameters:

The set of the affected nodes Nc is in the range as:

$$\{\overset{\bullet}{provider}(A)\} \le Nc \le \{\overset{\bullet}{provider}(A)\} \cup \{peer(\overset{\bullet}{provider}(A))\}$$
$$\cup \{\overset{\bullet}{customer}(peer(\overset{\bullet}{provider}(A)) \cup (\overset{\bullet}{provider}(A)))\}$$

To all the same prefix hijack types, the influenced ASes will change the paths to the hijacking prefix which is in the set of Nc. So, the affected path factor μ depends on proportion of Nc nodes in the whole network and the path BC of node V:

$$\mu = \frac{|Nc|}{N} \sum_{s \ne t \in N} Path_{st}(V)$$

N is the AS number of the network.

The covered prefix hijacking is much easier. Most all the ASes except for V will add a new path to the sub-prefix of V. So the set of the affected nodes Nc is all the ASes except for V.

$$Nc = All \setminus \{V\}$$

And the affected path factor μ is depends on the percentage of the sub-prefix hijacked in the prefix V assigned and the path BC of node V.

$$\mu = \eta \sum_{s \ne t \in N} Path_{st}(V)$$

η is the proportion of the sub-prefix in the prefix range of V.

From the analysis above, these results can be drawn:

1) The hijacked AS in the core layer is not the most awful thing. On the contrary, if the AS in the marginal layer being hijacked, the number of the affected nodes is the largest among the three levels;

2) The hijacked AS in the forwarding layer can affect more paths than the core layer or the marginal layer;

3) If the hijacked ASes are in the same level, the hijacking AS in the forwarding layer can affect more nodes than the core layer or the marginal layer, and the higher attacker is in, the larger its influence will be;

4) The sub-prefix hijack can affect more ASes than the same prefix hijack, and the lager sub-prefix range is, the bigger affected path factor μ will be.

4. Related Work

Various prefix hijack events have been reported to NANOG [23] mailing list from time to time. IETF's rpsec (Routing Protocol Security Requirements) Working Group provides general threat information for routing protocols and in particular BGP security requirements [24]. Recent works [3,25] give a comprehensive overview on BGP security. The prefix hijacking is one of the key problems being noticed to BGP in these papers.

Previous works on prefix hijacking can be sorted into two categories: hijack prevention and hijack detection. The former one is trying to prevent the hijacking in the protocol mechanism level, and the latter one is trying to find and alert the hijacking event after it happening. The methods can be categorized into two types: cryptography based and non-crypto based.

The cryptography methods, like [4-6, 27-31], imply that BGP routers sign and verify the origin AS and AS path of each prefix. Origin authentication [31] uses trusted database to guarantee that an AS cannot falsely claim to be the rightful owner for an IP prefix. However, the manipulator can still get away with announcing any path that ends at the AS that rightfully owns the victim IP prefix. Secure Origin BGP (soBGP) [30] provides origin authentication as well as a trusted database that guarantees that any announced path physically exists in the AS-level topology of the internetwork. However, a manipulator can still get away with announcing a path that exists but is not actually available. In addition to origin authentication, S-BGP [6] also uses cryptographically-signed routing announcements to provide a property called path verification. It effectively limits a single manipulator to announcing available paths. However, S-BGP does not prevent the manipulator from announcing the shorter, more expensive, provider path, while actually forwarding traffic on the cheaper, longer customer path. In SPV [32], the originator of a prefix establishes a single root value used to seed the generation of one-time signature structures for each hop in the PATH. However, the security of SPV is in some cases based on probabilistic arguments, which may be acceptable for some constrained environments, and it is unclear whether such arguments will be acceptable in the larger Internet. And it does not provide the requisite security to protect against path

modification. In addition to added router workload, these solutions require changes to all router implementations, and some of them also require a public key infrastructure. Due to these obstacles, none of the proposed prevention schemes is expected to see deployment in near future.

The non-crypto methods include [4, 9, 10, 12, 14]. PHAS [10] is predicated on the notion that a prefix owner is the only entity that can differentiate between real routing changes and those that take place as a result of a prefix hijacking attack. And if there are changes to the originator of a route, the owner of that prefix is notified through email. The system is incrementally deployable in that to join the system. A prefix owner need only register with the PHAS server; however, this server is also a single point of failure in the system, and if it is compromised, it could send out numerous false alarms to prefix owners. Additionally, the system relies on the validity of entities registering their prefixes; there is no protection against an adversary making a false registration. Hu and Mao examined prefix hijacking in greater detail and provided a mechanism for detecting prefix hijacking attacks in real time [14]. Their solution is based on fingerprinting techniques for networks and hosts. If there are conflicting origin ASes advertised, which is potential evidence of a prefix hijacking attack, the collected fingerprints are compared against probes sent to all origins. This approach relies on a real-time BGP UPDATE monitor, which sends differentiating probes if prefixes are advertised from multiple locations. The availability of the monitor is critical as, if updates are delayed, the ability to collect measures, such as probing and subsequent decision making, will be compromised. The Whisper protocol [4] is designed to validate the initial source of path information. The protocol seeks to alert network administrators of potential routing inconsistencies. A random value is initially assigned to each prefix by the originator. The value is repeatedly hashed at each hop as it is propagated from AS to AS. if the hash values are the same, then they must have come from the same source. Only the route originator can verify the route because of the non-invertibility of secure hash functions. Thus, the recipient would have to query the originator as to the veracity of the route, which is often outside of the purview of the originator's knowledge. Another recently-proposed alerting system is pretty good BGP (PGBGP) [12]. The key insight in this work is that misconfigurations and prefix hijacking attacks could be mitigated if routers exercise a certain amount of judgement with the routes that they adopt into their routing tables. MyASN[9] is an offline prefix hijack alert service provided by RIPE. A prefix owner registers the valid origin set for a prefix, and MyASN sends an alarm via regular email when any invalid origin AS is observed in BGP routing update.

5. Conclusion

This paper conducts a systematic study on the impaction prefix hijacks launched at different position in the Internet hierarchy based on the work in

paper [34]. The Internet is classified into three tiers—core layer, forwarding layer and marginal layer based on the power-law and commercial relations between ASes. Two impaction parameters—affected ASes set Nc and affected paths factor μ, are analyzed for the same prefix hijacking events and the covered prefix hijacking events in different layers. We studied nine type hijacking events based on the position of the hijacking ASes and the hijacked ASes.

The study shows that if the AS in the marginal layer being hijacked, the number of the affected nodes is the largest among the three levels. The hijacked AS in the forwarding layer can affect more paths than the core layer or the marginal layer. If the hijacked ASes are in the same level, the hijacking AS in the forwarding layer can affect more nodes than the core layer or the marginal layer, and the higher attacker is in, the larger its influence will be. The sub-prefix hijack can affect more ASes than the same prefix hijack, and the lager sub-prefix range is, the bigger affected path factor μ will be.

Acknowledgment. This research is supported by National Natural Science Foundation of China (Grant No. 61100223).

References

1 Mohit Lad, Ricardo Oliveira, Beichuan Zhang and Lixia Zhang ,Understanding Resiliency of Internet Topology Against Prefix Hijack Attacks. pp.368-377, 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07), 2007.
2 O. Nordstrom and C. Dovrolis, Beware of BGP attacks, SIGCOMM Comput. Commun. Rev., vol. 34, no. 2, 2004.
3 Kevin Butler, Patrick McDaniel and Jennifer Rexford. A Survey of BGP Security Issues and Solutions. Proceedings of the IEEE. Vol. 98, No. 1, January 2010
4 L. Subramanian, V. Roth, I. Stoica, S. Shenker, and R. H. Katz. Listen and whisper: Security mechanisms for BGP. In Proceedings of ACM NDSI 2004, March 2004.
5 J. Ng. Extensions to BGP to Support Secure Origin BGP. ftp://ftp-eng.cisco.com/sobgp/drafts/draft-ng-sobgpbgp-extensions-02.txt, April 2004.
6 S. Kent, C. Lynn, and K. Seo. Secure border gateway protocol (S-BGP). IEEE JSAC Special Issue on Network Security, 2000
7 S. S. M. Zhao and D. Nicol. Aggregated path authentication for efficient bgp security. In 12th ACM Conference on Computer and Communications Security (CCS), November 2005.
8 B. R. Smith, S. Murphy, and J. J. Garcia-Luna-Aceves. Securing the border gateway routing protocol. In Global Internet' 96, November 1996.
9 RIPE. Routing information service: myASn System. http://www.ris.ripe.net/myasn.html.
10 M. Lad, D. Massey, D. Pei, Y. Wu, B. Zhang, and L. Zhang. PHAS: A prefix hijack alert system. In 15th USENIX Security Symposium, 2006.

11 S. Qiu, F. Monrose, A. Terzis, and P. McDaniel. Efficient techniques for detecting false origin advertisements in interdomain routing. In Second workshop on Secure Network Protocols (NPSec), 2006.
12 J. Karlin, S. Forrest, and J. Rexford. Pretty good bgp: Protecting bgp by cautiously selecting routes. Technical Report TR-CS-2005-37, University of New Mexico, Octber 2005.
13 W. Xu and J. Rexford. MIRO: multi-path interdomain routing. In SIGCOMM 2006, pages 171–182, 2006.
14 X. Hu and Z. M. Mao, Accurate Real-time Identification of IP Prefix Hijacking, in Proc. of IEEE Security and Privacy (Oakland), 2007.
15 C. Zheng, L. Ji, D. Pei, J. Wang, and P. Francis, A Light-Weight Distributed Scheme for Detecting IP Prefix Hijacks in Realtime, in Proc. of ACM SIGCOMM, August 2007.
16 H. Ballani, P. Francis, and X. Zhang, A Study of Prefix Hijacking and Interception in the Internet. SIGCOMM Comput. Commun. Rev., vol. 37, no. 4, pp. 265–276, 2007.
17 M. Lad, R. Oliveira, B. Zhang, and L. Zhang, Understanding Resiliency of Internet Topology Against Prefix Hijack Attacks," in Proc. of IEEE/IFIP DSN, 2007.
18 Michalis Faloutsos, Petros Faloutsos, Christos Faloutsos. On Power-Law Relationships of the Internet Topology.1999.
19 Zegura, Calvert and Donahoo, “A quantitative comparison of graph-based models for Internet topology”, IEEE/ACM Transactions on Networking, December 1997.
20 R. Govindan and A. Reddy. An Analysis of Internet Inter-Domain Topology and Route Stability. In Proc. IEEE INFOCOM ‘97, March 1997.
21 GE Z, FIGUEIREDO D, JAIWAL S, and et al. On the hierarchical structure of the logical Internet graph [A]. Proceedings of SPIE ITCOM[C]. USA, August 2001.
22 Peidong Zhu, Xin Liu. An efficient Algorithm on Internet Hierarchy Induction. High Technology Communication.14: 358-361, 2004.
23 The NANOG Mailing List. http://www.merit.edu/mail.archives/nanog/.
24 B. Christian and T. Tauber. BGP Security Requirements. IETF Draft: draft-ietf-rpsec-bgpsecrec-04, March 2006.
25 Sharon Goldberg, Michael Schapira, Peter Hummon, Jennifer Rexford. How Secure are Secure Interdomain Routing Protocols? in Proc. of ACM SIGCOMM, August 30–September 3, 2010, New Delhi, India.
26 Y. Rekhter, T. Li, and S. Hares. Border Gateway Protocol 4. RFC 4271, Internet Engineering Task Force, January 2006.
27 RFC 4271, Internet Engineering Task Force, January 2006. S. S. M. Zhao and D. Nicol. Aggregated path authentication for efficient bgp security. In 12th ACM Conference on Computer and Communications Security (CCS), November 2005.
28 B. R. Smith, S. Murphy, and J. J. Garcia-Luna-Aceves. Securing the border gateway routing protocol. In Global Internet’ 96, November 1996.
29 T. Wan, E. Kranakis, and P. van Oorschot, Pretty Secure BGP, psBGP, in Proc. of NDSS, 2005.
30 R. White, Architecture and Deployment Considerations for Secure Origin BGP (soBGP), draft-white-sobgp-architecture-01, Nov 2005.
31 W. Aiello, J. Ioannidis, and P. McDaniel, Origin authentication in interdomain routing, in Proc. of conference on Computer and communications security (CCS), 2003.

32 Y.-C. Hu, A. Perrig, and M. Sirbu, B. SPV: Secure path vector routing for securing BGP, in Proc. ACM SIGCOMM, Portland, OR, Aug. 2004.
33 J. Karlin, S. Forrest, and J. Rexford, B. Autonomous security for autonomous systems, Comput. Networks, Oct. 2008.
34 Zhao JJ, Wen Yan, Li Xiang, etc. The Relation on Prefix Hijacking and the Internet Hierarchy, The 6th International Conference on Innovative Mobile and Internet Services (IMIS'12), Italy, July, 2012.
35 Judanov, M., Jacko, O., Jevtia, M.: Influence of Information and Communication Technologies on Decentralization of Organizational Structure. Computer Science and Information Systems, Vol. 6, No. 1, 93-109. (2009)

Jinjing Zhao received her B.S., M.S. and Ph.D. degrees in School of Computer from National University of Defense Technology, Changsha, China. She is currently an associate professor at Beijing Institute of System Engineering. Her major research interests include computer networks, and information security.

Yan Wen received his B.S., M.S. and Ph.D. degrees in School of Computer from National University of Defense Technology, Changsha, China. He is currently an assistant professor at Beijing Institute of System Engineering. His major research interests include virtualization technology, operating system, and information security.

Received: November 08, 2012; Accepted: March 25, 2013

DOI: 10.2298/CSIS120908026C

Two-Step Hierarchical Scheme for Detecting Detoured Attacks to the Web Server

Byungha Choi[1] and Kyungsan Cho[2]

[1] Graduate School, Dankook University
Yongin, Gyeonggi, Korea
notanything@hanmail.net
[2] Corresponding Author
Dept. of Software Science, Dankook University
Yongin, Gyeonggi, Korea
kscho@dankook.ac.kr

Abstract. In this paper, we propose an improved detection scheme to protect a Web server from detoured attacks, which disclose confidential/private information or disseminate malware codes through outbound traffic. Our scheme has a two-step hierarchy, whose detection methods are complementary to each other. The first step is a signature-based detector that uses Snort and detects the marks of disseminating malware, XSS, URL Spoofing and information leakage from the Web server. The second step is an anomaly-based detector which detects attacks by using the probability evaluation in HMM, driven by both payload and traffic characteristics of outbound packets. Through the verification analysis under the attacked Web server environment, we show that our proposed scheme improves the False Positive rate and detection efficiency for detecting detoured attacks to a Web server.

Keywords: detection scheme, two-step detection, detoured attack, signature-based, anomaly-based, outbound traffic

1. Introduction

Attacks to information systems have evolved steadily over a long time, and more Web-based attacks have replaced traditional attacks. Nowadays, more systems are reliant upon the Web server to get and exchange information through the Internet and the security shifts from lower layers of network protocol to the application layer. Thus, Web-based attacks focused on applications have become one of the most serious topics in the security field. That is, Web-based attacks focus on an application itself and functions on layer 7 of the OSI[13].

Application vulnerabilities could provide the means for malicious end users to break a system's protection mechanisms in order to take advantage of or gain access to private information or system resources. The most common Web-based attack types are SQL Injection, XSS (Cross-Site Scripting), buffer overflow, password cracking, spoofing, repudiation, information disclosure, denial of service, and evaluation of privileges[12,13].

Web-based attacks expose the weakness and vulnerability of the victim system, and disseminate malware to other hosts communicating with the victim system. Layered security systems with firewall, IDS (Intrusion Detection System) and WAF (Web Application Firewall) are provided to cope with the above attacks, and they protect the victim system very well. Traditionally, they detect external outsider threats by inspecting the traffic towards the system. Even though outsider attacks are constantly evolving and increasing, they are well detected and protected with the corresponding technical improvement. However, because an insider can directly access the Web server, these attacks should be coped with in other ways. It is found that insiders show unusual activities or abnormal behaviors when they access system resources for attacking purposes[21]. Thus, most works are based on identifying abnormal insider's behaviors and finding any significant change in insider's activities.

In addition, there are detoured attacks which bypass the traditional Web-based intrusion path. For example, there are e-mails with attached malware to insiders, or methods that use malicious USB memory or PDA with the aid of insiders and backdoor attacks. This unusual type of detoured attacks (including insider attacks) has become a more serious and common threat, and has already overtaken Web-based viruses and worm attacks[4]. Many detoured attacks to Web servers disclose confidential/private information, or disseminate malware codes to outside of the victim system, and these harms could be detected by inspecting the outbound traffic.

Thus, instead of inspecting inbound traffic, or analyzing a user's profiling and activity, we propose a scheme to inspect and detect abnormal outbound traffic caused by detoured attacks. Our proposed scheme is a two-step detection system composed of a signature-based IDS that uses Snort and an anomaly-based IDS that uses probability evaluation in HMM (Hidden Markov Model).

This paper is a revised and extended version of our previous work which was submitted to the MIST-2012 workshop[6]. The followings are major improvements to the earlier version:

1 To extend the scope of our detection, we improve the anomaly-based detection method, by adding a scheme detecting abnormal traffic characteristics from non-HTTP packets. In addition, we extend rules in the signature-based detection. With this extension, we verify the detection efficiency to backdoor attacks which produce abnormal non-HTTP traffic.
2 We evaluate our proposed two-step detection scheme with new datasets in terms of the FP (False Positive) rate, detection efficiency and detection rate. Through verification, we show that two detection methods in our proposal are complementary to each other, and our proposal is very efficient in detecting abnormal HTTP packets.
3 Most figures and tables are revised, and the evaluation results are extended, according to the above revision.

The rest of the paper is organized as follows. In Section 2, we review related work on security vulnerabilities of the Web Server and solutions to them. In Section 3, we propose our two-step hierarchical detection scheme. In Section 4, we

verify our scheme with real datasets collected under the attacked environment. In Section 5, a summary is provided.

2. Related Works

In 2010, OWASP announced the updated Top 10 most critical Web application security risks to educate about the consequences of the most important Web application security weaknesses and to provide basic techniques to protect against these high risk problem areas[17]. The WASC threat classification v2.0 shows proper classification of threats into attacks and weaknesses for a static/core view[1]. Both show the seriousness of Web-based attacks, with focus on the application layer of the protocol suite. Application vulnerabilities could provide the means for malicious end users to breach a system's protection mechanisms, in order to gain access to confidential and private information or system resources[13].

To detect Web-based outsider attacks, a layered Web security system is commonly used with firewall, IDS and Web application firewall. The security system protects the Web server from external attacks by inspecting the inbound traffic as shown in Figure 1.

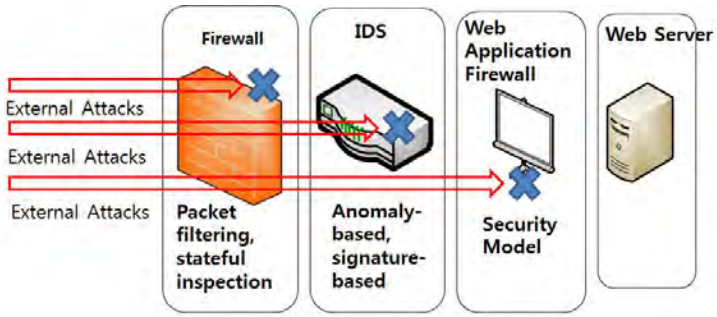


Fig. 1. Traditional Layered Security System for Web server

In the first stage of a layered security system, the firewall, which filters the specific network traffic between the network and the Web server, uses packet filtering and stateful inspection to detect simple intrusions. However, firewalls cannot prevent previously unknown attack types or insider attacks[14]. The second stage, IDS, uses signature-based or anomaly-based techniques, to protect against attacks that pass the firewall in the first stage. Signature-based IDS finds known patterns of misbehaviors in the message context, and anomaly-based IDS finds any deviation from the normal context patterns. IDS also can be classified according to the location and purpose as NIDS (Network-based

IDS) or HIDS (Host-based IDS). Mostly, NIDS is used in the second stage of a layered security system[23]. A Web application firewall filters packets by applying a set of rules. It uses a Positive Security Model or Negative Security Model or both. It filters packets which have already passed both the firewall and IDS[9].

Most current IDSs use only one of the two detection methods, signature-based detection or anomaly-based detection. However, each has its own limitations. Signature-based IDSs cannot detect any unknown attacks whereas anomaly-based IDSs cannot detect any untrained attacks. Thus, the detection accuracy of signature-based detection is extremely high, but the FP rate of anomaly-based IDSs is not negligible. Snort is a widely used IDS which allows pattern search for signature-based detection, and some works on using Snort rules in IDS have been proposed [11]. Security tools incorporating anomaly-based detection are proposed, and HMM, a statistical model of a system as a Markovian process with hidden states, has been shown to provide high level performance for detecting intrusions[5,15]. Lately, hybrid IDSs have been proposed that combine the two approaches into one system. For example, a hybrid IDS is obtained by combining packet header anomaly detection (PHAD) and network traffic anomaly detection (NETAD), or a hybrid intrusion detection system (HIDS) is configured with three sub-modules of misused detection module, anomaly detection module, and signature generation module[2,10]. However, by inspecting inbound traffic, they detect only external attacks.

Even though external attacks are constantly evolving and increasing, they are well detected and protected with the corresponding technical improvement. In fact, a layered security system pays little attention to what is happening inside the system. But, insiders show unusual activities or behaviors when they access system resources for attack purpose. Thus, information about a user's pattern of behavior and activity could be inspected for detection purposes. Most works on inside attacks are based on identifying abnormal insider's behavior, and finding any significant change in an insider's activity. Maloof and Stephens developed ELICIT, which detects insiders by inspecting insider's violating "need to know"[16]. However, this may not be enough to make a conclusion of a malicious act merely from knowing only a user's activity, and need further verification[21].

In addition to the insider attacks, detoured attacks which bypass the traditional security path to the Web server, are possible. For example, e-mail attacks with attached malware, attacks using USB memory/PDA attack with the aid of the insider, and backdoor attacks are detoured attacks. From hence, we address detour attacks as including insider attacks. If an e-mail containing a malware is sent to the insider and the insider accepts it, it causes the victim system to be remotely controlled by the outsider. If an outsider connects USB memory or PDA infected malware to a Web server with the aid of any insider, it causes detoured attack. A backdoor, which often refers to a backdoor program, is a hidden method for obtaining remote access to a computer that bypasses the traditional security mechanism. Thus, backdoor attack is a kind of detoured attack to be detected in our work. Backdoors can easily be installed on a victim

system that they aim to its exploit, thus detecting them requires considerable policies. A basic principle for backdoor detection is to find distinctive features of the activity of interest[22]. Table 1 shows possible detour attacks, and Figure 2 shows the possible detoured path of attacks described in Table 1.

Table 1. Types of Detoured Attacks

Type	Description
Through E-mail	Outsider sends e-mails containing malware procedures to the insider
Intentional Error	Insider intentionally makes programming errors
Through USB/ PDA	Insider connects contaminated USB/PDA with malware
Getting Information	Outsider gets account information from the insider
Backdoor Attacks	A hidden method of obtaining remote access to a computer system that bypasses the traditional security mechanism

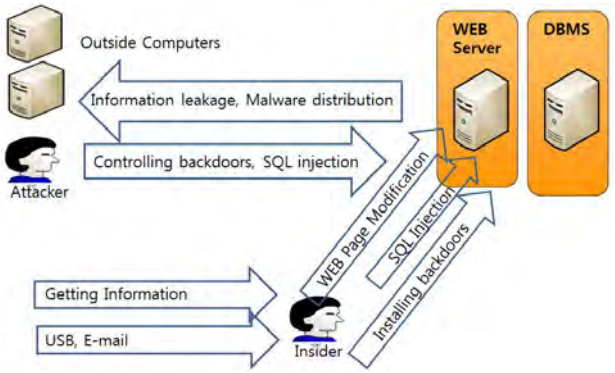


Fig. 2. Intrusion Path of Detoured Attacks

Detoured attacks to Web servers could use weaknesses of OWASP Top 10 or malware codes, thus causing altered HTML documents with Tags and JavaScript codes, SQL injection in DB, as well as altered traffic patterns. When a falsified Web page is activated by malware code, it could disclose confidential/ private information, and distribute malware codes[27]. However, conventional security solutions monitor network communication without paying much attention to outgoing traffic, due to the high processing cost of packet level network traffic analysis[24]. Lately, several works on inspecting outbound traffic for secu-

rity reasons have been proposed. For example, information leaks through HTTP were measured and detection of outbound malware traffic was proposed[3].

It is already known that most detoured attacks to a Web server show similar patterns in the HTTP outbound traffic[7], and a potential HTTP-based application-level attack exploits the features of Hypertext Markup Language (HTML)[26]. Thus, currently unknown detoured attacks could also be detected by inspecting outbound traffic. If any deviation from the normal context pattern for the specific traffic is found in the outbound traffic, it could be detected as ”attacked”. However, the outbound packets generated by detoured attacks, such as backdoor attacks, could be both HTTP packets and non-HTTP packets. It is shown that many samples use HTTP and continue with non HTTP-based damage functionality[20]. Snort also has TCP rules that have a port different from the HTTP port for non-HTTP traffic[18].

3. Proposed Detection Scheme with Two-Step Hierarchy

Our proposed scheme has a two-step hierarchy. The first step is signature-based detection using Snort and the second step is anomaly-based detection using HMM. We cannot find any other hybrid system that combines these two detection approaches to detect detoured attacks.

3.1. Overview of the proposal

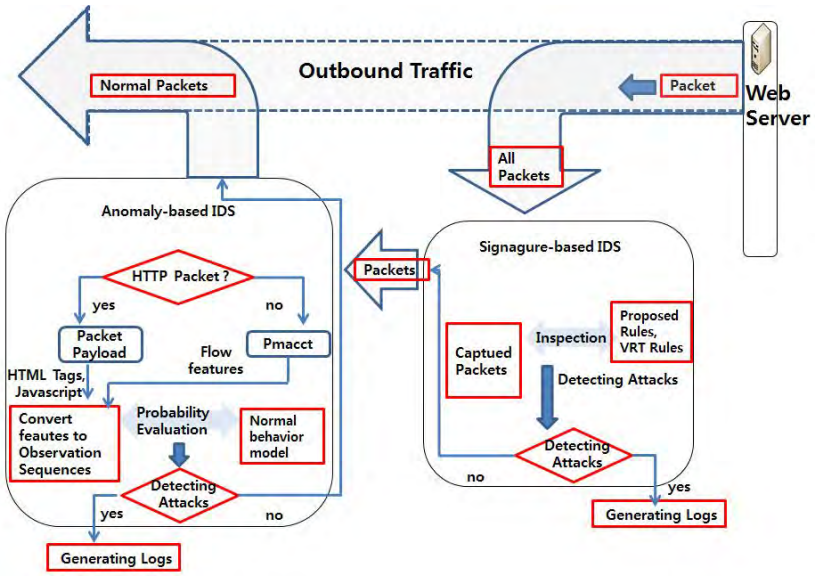


Fig. 3. Proposed Two-Step Detection Scheme

Traditional layered security systems detect intrusion, and do not pay attention to what happens after the intrusion. Unlike outsider attacks, detoured attacks make full use of this point, and bypass the traditional Web-based intrusion path. To protect from detoured attacks which the layered security system cannot detect, we proposed a two-step detection scheme. Detoured attacks show abnormal symptoms when packets are sent through the network, as discussed in Section 2. That is, abnormal contents of HTTP packets are transferred outwards through the network or outbound non-HTTP packets show abnormal traffic features. Therefore, instead of analyzing user's activities as done in the insider detection system, and instead of inspecting inbound packets as done in a traditional security system, we propose to inspect and detect abnormal outbound traffic caused by the detoured attack as shown in Figure 3.

3.2. Signature-based Detection

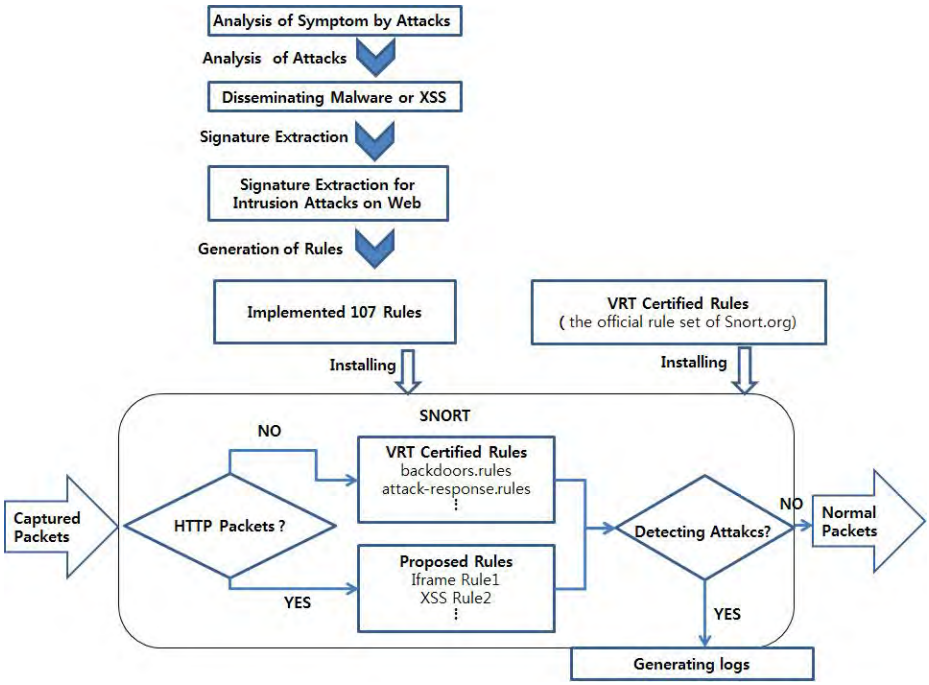


Fig. 4. Process of signature-based detection

The first step of our detection scheme is signature-based detection, which we implement using Snort which is an open source network intrusion prevention and detection system that combines the benefits of signature, protocol,

and anomaly-based inspection. A signature is a distinctive mark or characteristics contained in the context of a packet. Our signature-based IDS detects the symptoms of disseminating malware, XSS, URL Spoofing, information leakage and other abnormality from the Web server.

All these symptoms are represented as special forms of Tags and JavaScript codes as well as particular context in outbound packets. Thus, the above attacks could be detected by finding predefined signatures in the HTML documents transferred from the Web server.

Snort has a function to detect abnormal context in the outbound packets, if proper signatures and rules are provided. Thus, we define signatures found in abnormal HTTP packets with port number 80. We create new rules in Snort to inspect HTTP packets and detect predefined signatures. Our rules detect the actual vulnerability with signatures extracted in abnormal HTML documents. In addition, we use preexisting rules in Snort, called VRT certified rules, to detect abnormal non-HTTP packets generated by backdoor attacks. Figure 4 shows the detailed process of generating and applying rules in Snort from the signatures independent of other attributes.

3.3. Anomaly-based Detection

The second step of our detection scheme is anomaly-based detection which detects attacks by using HMM, and finding the probability of an observed sequence in a normal model. HMM is a statistical model of a system as a Markovian process with hidden states. HMM is characterized by the number of states N, the number of distinct observation symbols per state M, the state transition probability distribution A, the observation symbol probability distribution in a state B, and the initial state distribution π. Given appropriate values of N, M, A, B, and π, HMM can generate an observation sequence O. Thus, HMM requires specifications of two model parameters (N, M), observation symbols, and three probability measurements (A, B, π) and the compact notation λ=(π, A, B) is used to indicate the HMM model[19] . As an application of HMM to detecting attacks, we can find how to compute P(O| λ) under the given model λ=(π, A, B) with observation sequence O, and how to adjust the model parameters λ=(π, A, B) to maximize P(O| λ). With the Baum-Welch algorithm for this problem, we can train with the normal dataset in the same way of finding optimal values for π, A and B to maximize the probability of an observation sequence O given λ. Then, the probability evaluation, which finds the probability of the observation sequences(generated from tags/Javascript codes or flow features in outbound traffic) in the normal model, is used to detect attacks. We have already proposed an IDS with HMM[8].

Our proposed anomaly-based detection checks two events: 1) whether Tags or JavaScript codes in the HTTP outbound packet are normal, and 2) whether non-HTTP packets have abnormal flow features. This process requires the six phases shown in Figure 5. Normal behavior models are created according to

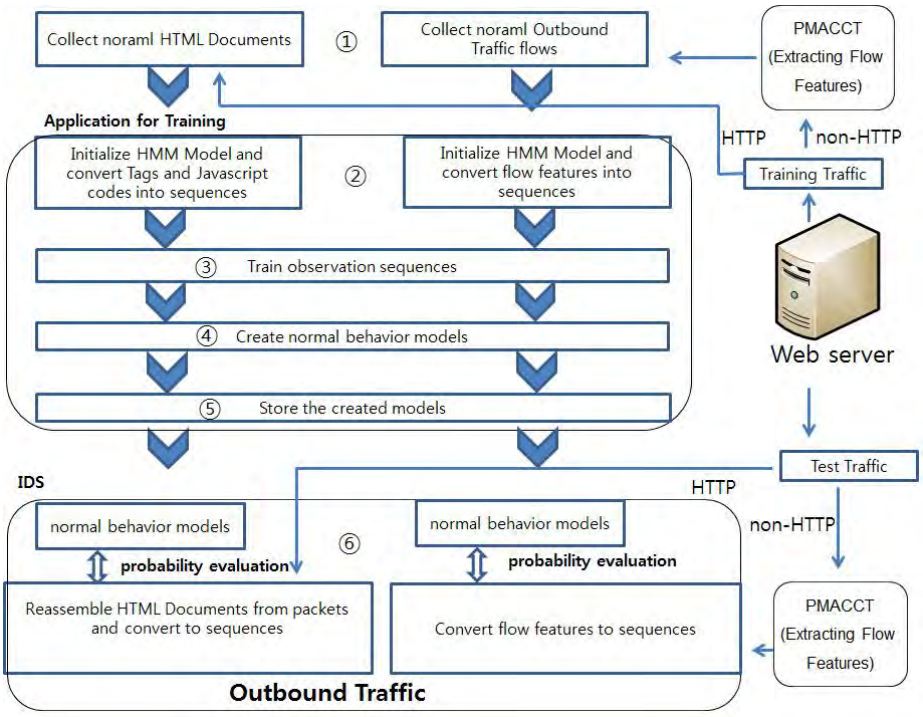


Fig. 5. Process of anomaly-based detection

the training results from both Tags or JavaScript codes in HTTP packets and statistical analysis of non-HTTP packets.

For HTTP packets, tags and JavaScript codes in each packet are applied to probability evaluation. For non-HTTP packets, each packet is monitored by the pmacct which is a small set of passive network monitoring tools to measure, account, classify, aggregate and export IPv4 and IPv6 traffic[25]. Flow features provided by the pmacct are applied to probability evaluation. The selected flow features in the detection are destination IP address, destination port number, source port number, TCP flags, average number of packets per flow, average number of bytes per flow, and time duration between two flows.

Both Tags or JavaScript codes in each HTTP outbound packet and selected features of each flow, are applied to probability evaluation, and attacks are detected according to the normal models that are configured as shown in Figure 5.

4. Verification Analysis

4.1. Test Environment

Table 2 shows a detailed description of our test environment.

4.2. Datasets

We use the following three datasets for the verification of detecting abnormal HTTP packets:

– Dataset1 with 670 Normal HTML documents generated by the common Web server with DB installed in the test environment. This dataset is used to evaluate the False Positive rate.

– Dataset2 with 100 altered HTML documents provided by one of the Korean Security Agencies. They are generated by real inside attacks and detoured attacks. They include obfuscated JavaScript codes or HTML tags, which work in their attempt to download and install malware or adware. This dataset is used to evaluate the detection rate.

– Dataset3 generated in real time by HDSI which is an SQL Injection Hacking Tool. This dataset is used to show detection efficiency.

The following dataset is used as non-HTTP outbound traffic:

– Dataset4 with packets generated by a Web server, in which 10 backdoor programs ($TrojanDropperBackdoorSpyware, GOV-bundestrojaner, APT-RTLO, Crime_Kelihos.B, Crime_Sinowal-Mebroot-Torpigandavariant malware, A16977E9CCBF86168CE20DFC33E0A93C, BBBreport, prorat, trojanSiscosBackdoor-as-FlashInstaller$) are installed. This dataset is

Table 2. Test Environment

device	item	description
Web server for dataset3	DBMS	MS - SQL 2000
	Server	IIS 5.0
	Web Programming Language	ASP
	Virtual machine	MS Virtual PC
Web server for other datasets	DBMS	MS - SQL 2000
	Server	IIS 7.5
	Web Programming Language	ASP
	Virtual machine	MS Virtual PC
Signature-based IDS	IDS	Snort 2.8.6.1
	Packet capture Library	winpcap 4.0
	Virtual machine	MS Virtual PC
Anomaly-based IDS	Packet capture Library	Jpcap 0.7
	HTML Parser	Jericho HTML Parser
	JavaScript Parser	Rhino 1.7 R3
	HMM Library	JaHMM
	JDK (language)	Oracle JDK 1.6 (Java)
	Virtual Machine	MS Virtual PC
	Flow Monitoring Tool	pmacct

used to show the efficiency in detecting abnormal non-HTTP packets generated by backdoor attacks.

4.3. Verification

The performance of the proposed scheme is analyzed in terms of the FP rate, detection rate, and detection efficiency.

Table 3. False Positive rate of each detection scheme

	normal documents	FPs	FP rate (%)
signature-based IDS	670	0	0.000
anomaly-based IDS		3	0.0044

To verify the FP rate, 670 normal HTML documents are randomly requested to the Web server. Table 3 shows the result. The FP rate of signature-based detection is negligible; it shows no error for 670 documents. The signature-based detector checks each packet in the outbound traffic, and classifies the packet as ”attacked” only if any defined signature equals any part of the payload of the packet. Thus, the FP rate, which means the evaluated result is ”attacked” for the unattacked packet, must be negligible. However, the FP of anomaly-based detection happens because of untrained abnormality caused by programming errors and exceptional payment documents.

Table 4. Detection rate of each detection scheme

	normal documents	detection rate
signature-based IDS	100 HTML documents	73%
anomaly-based IDS		89%

Table 5. Detection rate of two-step detection scheme

altered documents	passed documents after 1st step	passed documents after 2nd step	detection rate (%)
100	27 (73 are detected)	2 (98 are detected)	98

To verify the detection rate, we use dataset2 of 100 altered HTML documents. First, the two detection schemes are tested individually. As shown in Table 4, the detection rate of any single scheme is below 90%. Then, we test the same dataset in two steps; first signature-based detection and then anomaly-based detection. The evaluated results show that almost all altered documents generated by the attacks are detected through the two steps and the detection rate is 98%, as shown in Table 5.

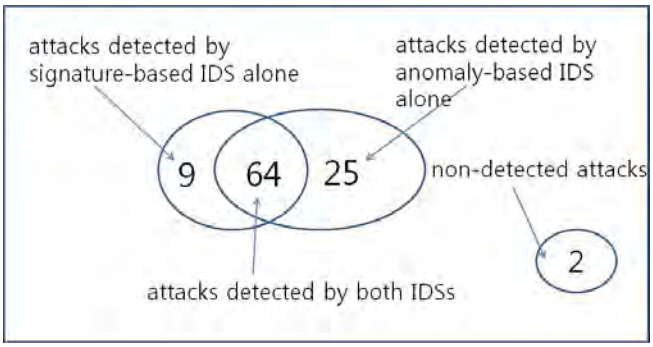


Fig. 6. Venn Diagram of detected attacks

Figure 6 shows how two detection methods in our scheme are complimentary to each other. The signature-based detector cannot detect any unknown or new attacks, but the anomaly-based detector can detect them if they have the same abnormalities as the known attacks have. In Figure 6, 25 detections by anomaly-based IDS alone represent these attacks. On the other hand, the anomaly-based IDS cannot detect any attacks producing untrained abnormality, or indistinctive tags or normal traffic features but the signature-based detector can detect them if they have special context in the outbound packets. 9 detections by signature-based IDS alone represent these attacks. Thus, our proposed two-step detection scheme increases the detection rate by detecting both unknown and untrained attacks, and the evaluated result agrees with the proposal that the two detection methods are complementary to each other.

Table 6. Detection of attacks generated by HDSI

Stage	Signature-based IDS	Anomaly-based IDS
1st stage	none	19
2nd stage	none	14
3rd stage	none	86
4th stage	none	311
5th stage	30	90

As a test of the detection efficiency, we use dataset3 generated by inputting various SQL queries as Web parameters in HDSI. This test is performed under the most vulnerable Windows 2000. HDSI tries to attack a DB in the Web server through 5 stages. Each stage generates attacks from the Web server, in order to get detailed data. An anomaly-based detector detects abnormal documents in each stage: 19 anomalies in the 1st stage, 14 anomalies in the 2nd stage, 86 anomalies in the 3rd stage, 311 anomalies in the 4th stage and finally 90 anomalies (3 anomalies per each e-mail, 30 e-mails) in the 5th stage, as shown in Table 6. From the analysis result, we can find the efficiency of anomaly-based detection for different anomalies caused by various attacks.

On the other hand, the signature-based detector detects the 30 e-mail attacks in the 5th stage only. This is because error messages only are generated in the 1st - 4th stages and no signatures are found in them. Only e-mails disclosed in the 5th stage have predefined signatures.

For the evaluation of the detection of backdoor attacks, we install 10 backdoor programs to remotely access the Web server. Then, abnormal traffic patterns in non-HTTP packets generated by the backdoor attacks are inspected.

Table 7. FP rate of abnormal non-HTTP packets

	Normal Flows	FPs	FP rate (%)
Signature-based IDS	926 Flows	0	0
Anomaly-based IDS		17	1.83

Table 8. Detection rate of abnormal non-HTTP packets

Malicious Outbound flows	After 1st step (Detection rate)	After 2nd step (Detection rate)
443	22 (4.9%)	443 (100%)

As shown in Table 7, the FP rate of signature-based detection is 0%, because of the same reason as in Table 3. However, the untrained traffic features in normal flows cause the FP rate of anomaly-based detection as high as 1.83%. Thus, we need more efficient traffic features to reduce the FP rate. From the detection rate shown in Table 8, our two-step scheme detects all abnormal non-HTTP flows. However, the detection rate of signature-based detection in the first step is relatively low at 4.9%. Thus, we need more effective signatures to increase this rate.

5. Summary

Even though external outsider attacks are constantly evolving and increasing, they are well detected and protected with the corresponding technical improvement. Detoured attacks (including insider attacks,) unusual type of attacks, become more serious, and pose a common threat that bypasses the traditional Web-based attack path for malicious purpose. It is found that many detoured attacks to the Web server disclose confidential/private information or disseminate malware codes through the Web, and this harm could be protected from, by inspecting their outbound traffic.

In this paper, we propose an improved detection scheme for detoured attacks by inspecting outbound traffic based on the analysis addressed in the previous sections. Our proposed scheme has a two-step hierarchy; the first step is signature-based detection using Snort and the second step is anomaly-based detection using HMM. To detect both abnormal Tags/JavaScript codes and abnormality caused by non-HTTP traffic, the second step inspects both payload and traffic features of the packets for probabilistic evaluation in HMM. We cannot find any other hybrid system that combines two approaches to detect detoured attacks.

Through the verification analysis under real attacked Web server environments, it has been shown that the proposed scheme causes a satisfactory false positive rate and detection efficiency for attacks to the Web server. In particular, the evaluated result agrees with the analysis that the two detection methods in our scheme are complementary to each other for detecting altered HTTP packets as shown in Table 9. In addition, our anomaly-based detector shows good efficiency in detecting abnormal non-HTTP packets caused by backdoor attacks. Our work on detecting abnormal non-HTTP packets is in progress; we need to supplement more signatures and traffic features for complete two-step detection as shown in the verification analysis.

Table 9. Comparison of two detection methods for abnormal HTTP packets

Features	Signature-based	Anomaly-based
Weakness	unknown attack	untrained abnormality
Method	signature matching	HMM
Complexity	simple	complex (time consuming)
FP rate	almost none	very low
Detection rate	relatively low	very high

As an extension, our proposed scheme may be applicable to client-side hosts and mobile devices. In particular, mobile malware has emerged as a serious threat to resource-constrained handheld devices over the past few years. For example, outbound traffic from a smartphone becomes more dangerous, due to confidential information leakage, the scanning of private documents and DDOS attack.

Acknowledgments. The research was conducted by the research fund of Dankook University in 2013.

References

1. The WASC Threat Classification v2.0. Tech. rep., The Web Application Security Consortium (2010), http://projects.webappsec.org/w/page/13246978/Threat%20Classification
2. Aydn, M.A., Zaim, A.H., Ceylan, K.G.: A hybrid intrusion detection system design for computer network security. Computers & Electrical Engineering 35(3), 517 – 526 (2009), http://www.sciencedirect.com/science/article/pii/S0045790609000020
3. Borders, K., Prakash, A.: Quantifying information leaks in outbound web traffic. In: Proceedings of the 2009 30th IEEE Symposium on Security and Privacy. pp. 129–140. SP '09, IEEE Computer Society, Washington, DC, USA (2009), http://dx.doi.org/10.1109/SP.2009.9
4. Bowen, B.M., Hershkop, S., Keromytis, A.D., Stolfo, S.J.: Baiting inside attackers using decoy documents
5. Cho, S.B., Han, S.J.: Two sophisticated techniques to improve hmm-based intrusion detection systems. In: RAID. pp. 207–219 (2003)
6. Choi, B., Cho, K.: Detection of insider attacks to the web server(mist 2012 volume 1). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA) 3(4), 35–45 (12 2012)
7. Choi, B., Choi, S., Cho, K.: An Efficient Detection Scheme of Web-based Attacks through monitoring HTTP Outbound Traffics. Journal of The Korea Society of Computer and Information 16(1), 125–132 (2011)
8. Choi, B., Choi, S., Cho, K.: Anomaly Detection Scheme of Web-based Attacks by applying HMM to HTTP Outbound Traffic. Journal of The Korea Society of Computer and Information 17(5), 33–40 (2012)
9. Desmet, L., Piessens, F., Joosen, W., Verbaeten, P.: Bridging the Gap Between Web Application Firewalls and Web Applications. In: Proceedings of the 2006 ACM Workshop on Formal Methods in Security Engineering. pp. 67–77 (2006)
10. Ding, Y.X., Xiao, M., Liu, A.W.: Research and implementation on snort-based hybrid intrusion detection system. In: Machine Learning and Cybernetics, 2009 International Conference on. vol. 3, pp. 1414 –1418 (july 2009)
11. Gómez, J., Gil, C., Padilla, N., Baños, R., Jiménez, C.: Design of a snort-based hybrid intrusion detection system. In: Proceedings of the 10th International Work-Conference on Artificial Neural Networks: Part II: Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. pp. 515–522. IWANN '09, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-02481-8_75
12. Jovicic, B., Simic, D.: Common web application attack types and security using asp.net. Comput. Sci. Inf. Syst. 3(2), 83–96 (2006), http://dblp.uni-trier.de/db/journals/comsis/comsis3.html#JovicicS06
13. Justin Crist: Web Based Attacks. Tech. rep., SANS Institute (2010), http://www.sans.org/reading_room/whitepapers/application/web-based-attacks_2053
14. K., S., P., H.: Guidelines on Firewalls and Firewall Policy. Tech. rep., Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg (2009)

15. Khreich, W., Granger, E., Sabourin, R., Miri, A.: Combining hidden markov models for improved anomaly detection. In: Proceedings of the 2009 IEEE international conference on Communications. pp. 965–970. ICC'09, IEEE Press, Piscataway, NJ, USA (2009), http://dl.acm.org/citation.cfm?id=1817271.1817451
16. Maloof, M.A., Stephens, G.D.: Elicit: a system for detecting insiders who violate need-to-know. In: Proceedings of the 10th international conference on Recent advances in intrusion detection. pp. 146–166. RAID'07, Berlin, Heidelberg (2007), http://dl.acm.org/citation.cfm?id=1776434.1776446
17. Mike Boberski, Juan Carlos Calderon et al.: OWASP Top 10 - The Ten Most Critical Web Application Security Risks. Tech. rep., OWASP (2010), https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project
18. Pontarelli, S., Greco, C., Nobile, E., Teofili, S., Bianchi, G.: Exploiting dynamic reconfiguration for fpga based network intrusion detection systems. In: Field Programmable Logic and Applications (FPL), 2010 International Conference on. pp. 10 –14 (31 2010-sept 2 2010)
19. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Proceedings of the IEEE. pp. 257–286 (1989)
20. Rossow, C., Dietrich, C.J., Bos, H., Cavallaro, L., van Steen, M., Freiling, F.C., Pohlmann, N.: Sandnet: network traffic analysis of malicious software. In: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security. pp. 78–88. BADGERS '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1978672.1978682
21. Salem, M., Hershkop, S., Stolfo, S.: A Survey of Insider Attack Detection Research. In: Stolfo, S., Bellovin, S., Keromytis, A., Hershkop, S., Smith, S., Sinclair, S. (eds.) Insider Attack and Cyber Security, Advances in Information Security, vol. 39, pp. 69–90. Springer US (2008), http://dx.doi.org/10.1007/978-0-387-77322-3_5
22. Salimi, E., Arastouie, N.: Backdoor detection system using artificial neural network and genetic algorithm. In: Proceedings of the 2011 International Conference on Computational and Information Sciences. pp. 817–820. ICCIS '11, IEEE Computer Society, Washington, DC, USA (2011), http://dx.doi.org/10.1109/ICCIS.2011.103
23. Shaimaa Ezzat Salama, Mohamed I. Marie, L.M.E.F., Helmy, Y.K.: Web Anomaly Misuse Intrusion Detection Framework for SQL Injection Detection. International Journal of Advanced Computer Science and Applications 3(3), 123–129 (2012)
24. Skrzewski, M.: Analyzing outbound network traffic. In: Kwiecie, A., Gaj, P., Stera, P. (eds.) Computer Networks, Communications in Computer and Information Science, vol. 160, pp. 204–213. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-21771-5_22
25. pmacct project team: pmacct project, http://www.pmacct.net/
26. Wang, X., Luo, J., Yang, M., Ling, Z.: A potential http-based application-level attack against tor. Future Generation Computer Systems 27(1), 67 – 77 (2011), http://www.sciencedirect.com/science/article/pii/S0167739X10000713
27. Yim, K., Hori, Y.: Information leakage prevention in emerging technologies (mist 2012 volume 2). Journal of Internet Services and Information Security (JISIS) 2(3/4), 1–2 (2012)

Byungha Choi received the MS degree from the Dept. of Information and Communication Technology, Dankook University. He is currently a Ph.D. student of Dept. of Computer Science and Engineering at Dankook University. His research interest is Network Security.

Kyungsan Cho(Corresponding Author) received his B.Sc. in Electronics Engineering(Seoul National University, 1979), master degree in Electrical and Electronic Engineering(KAIST, 1981), and his Ph.D. degree in Electrical and Computer Engineering(the University of Texas at Austin, 1988). During 1988-1990, he served as a senior R&D engineer at Samsung Electronics Company. He joined Dankook University in March 1990, where he is currently a professor in the department of software science. He authored several books in Computer Architecture and Computer Networks and published over 40 academic papers. His research interests include mobile networks, network security and traffic analysis.

Received: September 8, 2012; Accepted: March 11, 2013.

DOI: 10.2298/CSIS120702027H

An Efficient GTS Allocation Scheme for IEEE 802.15.4 MAC Layer

Der-Chen Huang, Yi-Wei Lee and Hsiang-Wei Wu

National Chung Hsing University, Taiwan, R.O.C.

huangdc@nchu.edu.tw

Abstract. Based on IEEE 802.15.4, the contention-free period (CFP) adopts a guarantee time slot (GTS) mechanism to ensure each device can access the radio channel. However, it is hard to get the authority to access the radio channel due to more competitor access the radio channel simultaneously. To cope with this issue, we proposed a guarantee time slot mechanism to enhance the performance and utilization by using CFP. Our proposed method ensures each device has the authority to access the radio channel without any additional step. By comparing with the method of IEEE 802.15.4, the experimental results show that data average transmission delay and energy consumption can be reduced dramatically. In addition, the bandwidth and performance of network is improved since the pre-allocation mechanism can reduce the number of control packets. Several experiments have been conducted to demonstrate the performance of our work

Keywords: ZigBee, GTS, Cluster Tree, Beacon, MAC Layer.

1. Introduction

In recent decade, Wireles Sensor Networks (WSNs) have been improved and applied popularly in many fields because of low cost, low power consumption, small size and short transmission distance [1-4]. Hence, the ZigBee standard has been proposed to satisfy these requirements to achieve the goal of longer lifetime with higher reliability [5]. In general, the ZigBee tree routing is a popular mechanism and has been applied in many applications. Because the routing table is not necessary when transmitting data, the tree routing mechanism is suitable for small, cost oriented and resource limited network applications. However, the number of intermediate nodes is increased dramatically while applying to a large range network. It requires more energy consumption and delay time such that the performance and life cycle of total system are decrease. To cope with this issue, [6] provides a shortcut tree routing method. This paper provides a shot cut path for transmission to save more energy and reduce delay time.

The risk of losing channel access authority is possibly happened in the star or tree topology network. This is because that IEEE 802.15.4 adopts the

policy of CSMA/CA to access the radio channel. [7] presents a back off assignment mechanism to avoid collision while accessing radio channel. The IEEE 802.15.4 proposes a guarantee time slot (GTS) mechanism to ensure the requirement of bandwidth and delay can be satisfied. However, the channel access successful possibility is highly related with the number of access nodes while using CSMA/CA mechanism. Moreover, the CAP is used to obtain the authority of using the radio channel in GTS mechanism. This increases the difficulty of accessing radio channel. In our method, we propose a pre-allocation time slot mechanism to ensure the radio channel to be accessed without considering the contention activities. By means of cancelling the contention access period (CAP) in CSMA/CA, this paper proposes a new concept of guarantee time slot (GTS) to ensure the total superframe to be occupied by contention free period (CFP) to increase the bandwidth of GTS. This method has the advantage of: (1) to reduce the energy consumption caused by the CSMA/CA mechanism, (2) to decrease the number of control packets and (3) to improve the delay effect.

2. Related Works

Regarding the GTS mechanism, NGA [8] and i-GAME [9] provide bandwidth allocation method to decrease the waste of bandwidth. NGA divides the CFP into 16 time slots with equal size to allocate the time slot to decrease the waste of CFP. Its drawback is to compete with other devices to acquire the access authority by means of CAP. Hence, the bandwidth of CFP becomes small and the probability of allocation is smaller due to more competitors. In contrast, the i-GAME provides a mechanism to allow more devices in one GTS. It induces more delay if there are more devices in one time slot. Most of the previous mentioned papers are focused on using the CFP area to allocate the GTS. Following the CSMA/CA protocol, each device must get the authority of allocating one time slot during the period of CAP to access the channel. However, the access probability is decrease if more devices want to access the channel such that more energy consumption is required during this compete period. Thus, a new GTS allocation scheme had been proposed to reduce the transmission power consumption and delay time [10]. [11] presented a slotted beacon scheduling to reduce the power consumption while considering the hierarchical tree topology and beacon-enabled network. The energy efficiency for each different method had been compared and analyzed in [12]. It proposed several new schemes to reduce power consumption for personal area network (PAN).

This paper proposes a new mechanism to combine the CAP and CFP together to allocate the time slot again for each device. We adopt the ZigBee tree routing structure to demonstrate the efficiency of our work. This paper applies a pre-allocation schema to solve the energy consumption induced from CSMA/CA mechanism such that it can be used to decrease the GTS requests, to remove the control packet, and to reduce the delay and energy

consumption. Here, we have illustrated the cluster tree based network as an example to demonstrate the performance of our proposed method for simplicity. In addition, our method can be easily further applied to different network topology. However, the node number calculation algorithm is required to develop again for each different network.

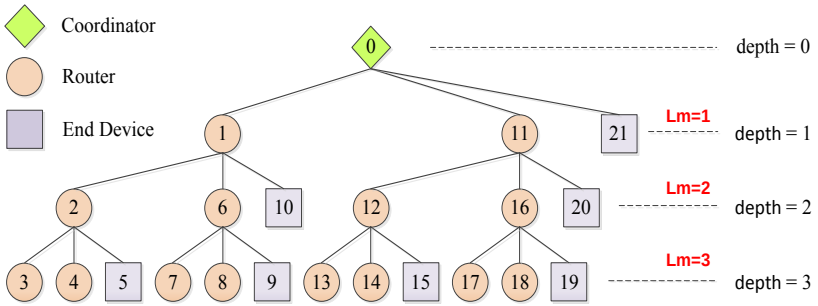


Fig. 1 Zigbee tree topology

3. New GTS Allocation Scheme

Based on the network address allocation method of IEEE 802.15.4 specification, we assign each device in the Zigbee network with a dedicate network address. After that, the topologic relation between each device can be constructed. The router or coordinate device (i.e. parent device) can assign time slot to each child device as a GTS for communication. In Zigbee, the device network address assigned mechanism is fulfilled in a distributed manner. To obtain network address, there are three topological parameters to be defined: the number of child devices (Cm), the network maximum depth value (Lm) and the maximum number of router under parent node (Rm). In Equation 3.1, the PAN coordinator (i.e. parent) can compute the Cskip(d) as an offset to derive network starting address for each child device if three parameters have been provided for any given parent device with depth value d.

$$\begin{cases} 1 + Cm \cdot (Lm - d - 1) & Rm = 1 \\ \frac{1+Cm-Rm-Cm \cdot Rm^{Lm-d-1}}{1-Rm} & Rm \neq 1 \end{cases} \quad (3.1)$$

Our metho oe n't requ re the AP for o ng the SMA/ A act v t e while considering to accessing the communication channel. Thus, the CAP can be used for transmitting data between parent node and child node to avoid collision and save energy. Once the ZigBee network has been set up, each parent device (coordinator or router) decide the GTS size for each child device according to the parameter Cm. In this paper, the maximum Cm should not be over 15 because each superframe includes 16 time slots and

one is already dedicated to beacon. In other words, each allocated GTS size is identical to one time slot size in superframe as the Cm is equal to 15.

According to the specification of IEEE 802.15.4, beacon interval (BI), superframe duration (SD), beacon order (BO) and superframe order (SO) can be used to described the relation between beacon frame and superframe. The superframe duration and beacon interval can be computed as Equations 3.2 and 3.3, where aBaseSuperframeDuartion is a time unit. The length of each time slot can be obtained by using Equation 3.4. Superframe Duration can be denoted as $F_S(SO)$,which can be divided into 16 portions. The new superframe contention free period can be found as Equation 3.5 in which one time slot is already used by beacon signal. Thus, the allocated GTS size for each device can be computed as Equation 3.6 for the case with Cm child device.

$$F_S(SO)=\text{Superframe Duration(SD)} = \text{aBaseSuperframeDuartion}*2^{SO} \tag{3.2}$$

$$F_B(BO)=\text{Beacon Interval(BI)} = \text{aBaseSuperframeDuartion}*2^{BO} \tag{3.3}$$

$$\text{Length of Time Slot : } L_{TS}(SO) = \frac{F_S(SO)}{16} \tag{3.4}$$

$$newCFP(SO) = F_S(SO) - L_{TS}(SO) \tag{3.5}$$

$$\text{Length of GTS : } L_{GTS}(Cm,SO) = \left\lfloor newCFP(SO) \times \frac{1}{Cm} \right\rfloor \tag{3.6}$$

We propose a method to compute the GTS for each device. The device can be divided as coordinator, router and end device, where only the coordinator and router can issue the beacon signal and the GTS to their corresponding child device. We illustrate the tree topologic structure as our basis to demonstrate our proposed method.

Algorithm 1 GTS communication scheme in the coordinator

1. MAC layer use GTS allocation scheme
2. define the Superframe Duration(SD) = $F_S(SO)$
3. define the Beacon Interval(BI) = $F_B(BO)$
4. define the length of time slot $L_{TS}(SO) = \frac{F_S(SO)}{16}$
5. newCFP (SO) = $F_S(SO) - L_{TS}(SO)$
6. Each length of GTS : $L_{GTS}(Cm,SO) = \left\lfloor newCFP(SO) \times \frac{1}{Cm} \right\rfloor$
7. for i_{th}=1 to $i_{th} \leq Cm$

8. Starting time of $GTS(i_{th})$ = $\frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (i_{th} - 1)$
9. end for
10. Send a beacon signal for synchronization

To ensure the integration operations smoothly, we have presented three algorithms in corresponding to coordinator, router and end device, respectively. First, for the coordinator algorithm as shown in algorithm 1, the row 2 to 4 gives the definition of parameter used in this algorithm. The row 5 finds the size of new CFP. The size of allocated GTS for each device is obtained by the row 6, where Cm represents the number of child devices under the coordinator. The starting time of GTS for each device can be obtained in the row 7 and 8. The Equation 3.7 indicates that the starting time of GTS for each device is related to the assigned network address of child device (i.e. *i*). The assigned sequence of GTS for each child device follows the network address of joining parent net. The parent device can communicate with the child device during this GTS period after the synchronization is initialized by the beacon signal.

$$GTS(i_{th}) = \frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (i_{th} - 1), \text{where } i = i \text{ th Cm} \quad (3.7)$$

Algorithm 2 GTS communication scheme in the router

---

1. MAC layer use GTS communication scheme
2. define the Superframe Duration(SD) = F_S(SO)
3. define the Beacon Interval(BI) = F_B(BO)
4. define the length of time slot L_{TS}(SO) = $\frac{F_S(SO)}{16}$
5. newCFP (SO) =$F_S(SO) - L_{TS}$(SO)
6. Each length of GTS : L_{GTS}(Cm,SO) =$\left\lfloor newCFP(SO) \times \frac{1}{Cm} \right\rfloor$
7. if receive a Beacon signal with synchronization
8. then k is computed by: $k = \frac{A_k - A_{parent} - 1 + Cskip_{parent}(d)}{Cskip_{parent}(d)}$
9. Starting time of $GTS_{parent} = \frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (k - 1)$
10. end if
11. After SD_{parent} of period then Allocate GTS for children
12. for i_{th}=1 to $i_{th} \leq Cm$
13. Starting time of $GTS(i_{th})$ = $\frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (i_{th} - 1)$
14. end for
15. Send a beacon signal for synchronization

Second, for algorithm 2, the row 2 to 5 is the same as algorithm 1. The row 6 is to compute the GTS size L_{GTS} of each router device. The row 7 to 10 is to

find the starting time of GTS of router device, which is assigned by the parent device.

$$k = \frac{A_k - A_{parent} - 1 + Cskip_{parent}(d)}{Cskip_{parent}(d)}\ , \qquad \text{where k = k th Router} \tag{3.8}$$

$$GTS_{parent} = \frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (k-1) \tag{3.9}$$

In Equation 3.8, A_k is the network address of router, A_{parent} is the network address of parent and $Cskip_{parent}(d)$ is the offset address for each different router. The k denotes the k_{th} router of parent device. Thus, we can compute the starting time of accessing channel for each different router in terms of k as Equation 3.9 shows. The router communicates with its parent during time slot GTS_{parent}. The row 11 indicates that the processing sequence is from parent to child. After the Superframe Duration of parent, the router assigns GTS to its child device by following the sequence of network address. The rows 12, 13 and 14 compute the starting time of GTS for each child device. We can assign each different starting time for each child device based on a different i_{th}, where the i_{th} denotes the number of child device for a router. The router communicates with its child device during time slot $GTS(i_{th})$.

Algorithm 3 GTS communication scheme in the end device

1. MAC layer use GTS communication scheme with parent
2. define the Superframe Duration(SD) = $F_S(SO)$
3. define the Beacon Interval(BI) = $F_B(BO)$
4. define the length of time slot L_{TS}(SO) = $\frac{F_S(SO)}{16}$
5. CFP_{parent} = newCFP(SO) =$F_S(SO) - L_{TS}$(SO)
6. Each length of GTS_{parent} =$\left\lfloor newCFP(SO) \times \frac{1}{Cm} \right\rfloor$
7. if receive a beacon signal with synchronization
8. then
   $n = A_n - A_{parent} - Cskip_{parent}(d) \times Rm$
9. Starting time of $GTS(i_{th}) = \frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (n + Rm - 1)$
10. end if

Similarly, for the algorithm 3, the row 2 to 6 gives definitions regarding superframe duration, beacon interval and length of time slot to find the size of CFP_{parent} and GTS_{parent}. The row 7 indicates that the beacon signal is received for synchronization. The row 8 identifies the end device is the n_{th} device in parent net based on Equation 3.10.

$$n = A_n - A_{parent} - Cskip_{parent}(\mathrm{d}) \times Rm$$
$$\mathrm{n} = \mathrm{n}_{th} \text{ end device} \qquad (3.10)$$

In Equation 3.10, A_n is the network address of end device assigned by its parent device, A_{parent} is the network address of parent device, $Cskip_{parent}(\mathrm{d})$ is the offset and *Rm* is the number of router in child devices. The *n+Rm* can be used to represent the i_{th} *Cm* child device in parent net. We can compute the staring time of GTS for each child device based on Equation 3.11. The end device communicates with its parent in time slot $GTS(i_{th})$.

$$GTS(i_{th}) = \frac{F_S(SO)}{16} + L_{GTS}(Cm, SO) \cdot (n + Rm - 1)$$
$$\text{,where } n+Rm=i_{th}\ Cm \qquad (3.11)$$

To evaluate these algorithms, we illustrate an example to demonstrate our ro o e metho . Let' llu trate an exam le u ng a Z gBee tree ba e topologic as shown in Fig. 1. In this example, the coordinator locates start at the network address 0 with depth equal to 0, where we assume Cm=3, Lm=3 and Rm=2. Then, the Cskip can be computed as Cskip(0)=10, Cskip(1)=4 and Cskip(2)=1. The network address for the device with depth =1 can be assigned as 1, 11 and 21 by the coordinator. The status inside superframe is shown in Fig. 2 after executing the synchronization of parent and child devices. The contention free period includes GTS for each different device because of no channel contention access required in our method. The communication with coordinator is necessary to obtain the GTS allocation command. In this case, the coordinator (0) generates the beacon signal to synchronize with devices 1, 11 and 21 in the first time slot and then the coordinator (0) communicates with device 1, 11 and 21 within the period GTS1, GTS2 and GTS3, respectively. After that, the device 1 starts to synchronize with devices 2, 6 and 10 during this inactive period of coordinator (0). Similarly, the status inside superframe for the case between router and its child devices can be shown in Fig. 3, where the communication with router is required. The device 6 starts to synchronize with devices 7, 8 and 9 within this inactive period of device 1. The communication between the parent and child devices can be conducted in the pre-allocate GTS. We can ensure that the fairness of accessing bandwidth based on the pre-allocate GTS mechanism. The total quantity of packet and delay time can thus be reduced and improved because the child device doe n't requ re to asking its corresponding parent device to send or cancel the control packet for GTS. In addition, the energy consumption can be reduced since the CSMA/CA mechanism is not required in our research. The superframe for the routers 1, 11, 2, 6, 12 and 16 with depth varied from 1 to 3 can be shown together as Fig. 4.

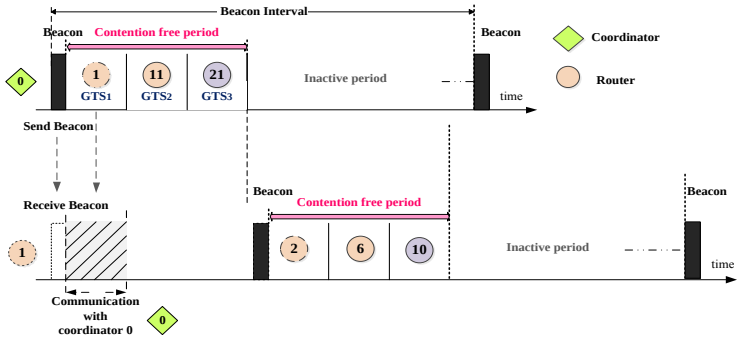


Fig. 2 The relation between coordinator and child device

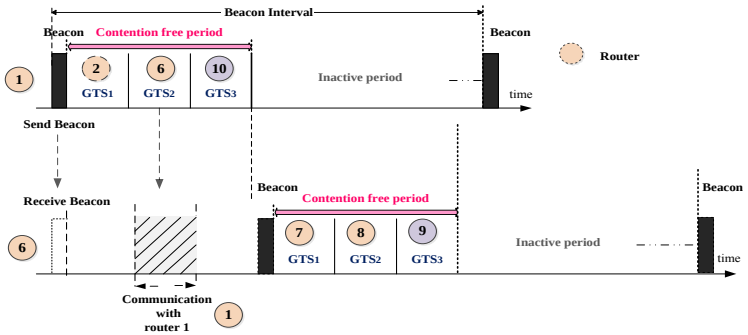


Fig. 3 The relation between router and child device

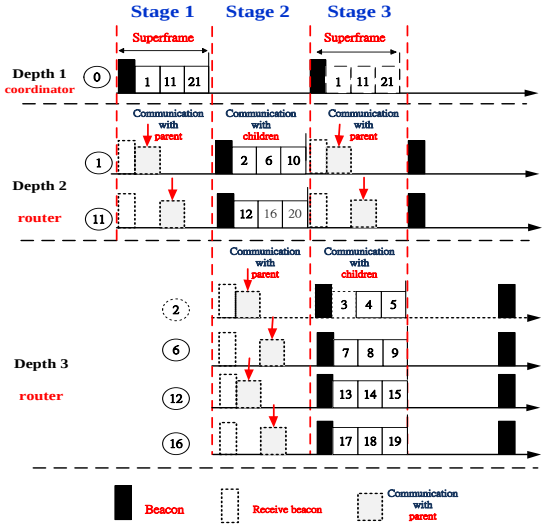


Fig. 4 The relation between depth 1 to 3 for tree-topology

4. Experimental Results

To evaluate our proposed method, our simulation environment is based on NS2 version 2.34 as a basis to demonstrate the performance and correctness of our work. In this experiment, we illustrate a tree type sensor network as an example shown in Fig. 1, where its corresponding parameters can be set as Cm=3, Rm=2 and Lm=3. For simplicity, we assume that the traditional GTS allocation mechanism (OSA GTS) with CSMA/CA can be 100% success without any loss. In other words, the compared method with us is an optimal case. In here, we assume SO from 2 to 4 and BO various from 3 to 5 to observe the variation after increasing the beacon interval.
To examine the delay effects in our method, we treat the coordinate as a sink node to accept the signal from other nodes. First, we assume depth=3 to simulate the packet is transmitted from the node with depth=3 to the sink node. Similarly, the rest nodes follow the same way by transmitting data repeatedly along with the directions as Fig. 5. The simulation results can be shown in Fig. 6 to 11 with assigning SO=2~4 and BO=3~5, where the input data arrival rate can vary from 0.1 to 1. The Fig. 6, 8 and 10 show the transmit delay with respect to BO=3, BO=4 and BO=5, respectively. Similarly, the Fig. 7, 9 and 11 show the energy consumption for the BO=3, BO=4 and BO=5. First, for the average delay results, the OSA GTS average delay is larger than the NSA GTS over 100 % becau e the NSA GTS oe n't nee 'acqu re' an 'cancel' act vities because of pre-allocate GTS scheme. Hence, the communication can be executed immediately between parent and child ev ce after rece v ng the arent ev ce' beacon gnal. However, the OSA GTS needs to wait for beacon signal in the next superframe and then receive the related GTS information to allocate the radio channel. After that, the communication between parent and child devices can be started. Therefore, the delay overhead of OSA GTS is more than the NSA GTS. For the OSA GTS, the delay is increased as if the arrival rate is becoming greater (i.e. 0.1 to 1) since the arrival packets during the CAP period will be totally handled in the next GTS. In contrast, the NSA GTS keeps the uniform delay regardless of the change of arrival rate. The energy consumption of OSA GTS is larger than NSA GTS scheme because OSA GTS requires one more superframe to transmit data packet if the radio channel is allocated successfully. To implement such a system, we need to modify the sensor node by adding our proposed algorithm.

To further compare the delay effects, we propose the formula:

$$\text{ADRI}=\frac{OSA\ GTS\ Average\ Delay-NSA\ GTS\ Average\ Delay}{OSA\ GTS\ Average\ Delay}\times 100\%$$

as Average Delay Reduction Index (ADRI) to indicate the average delay. The ADRI with SO=2, 3, 4 and BO=3, 4, 5 can be shown in Fig. 12, 13 and 14, respectively. Based on the ADRI with SO=2~4 and BO=3~5, the average improvement rate is over 51%, especially for the case with SO=4, BO=5 and

arrival rate=0.1, where its improve rate is 80%. It is obviously to indicate that low transmission delay can be obtained by the NSA GTS method.

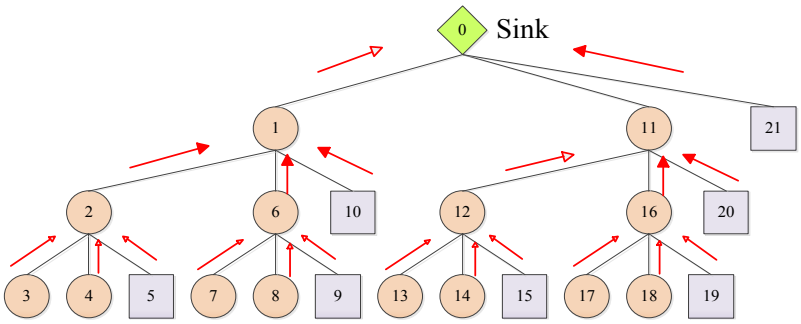


Fig. 5 Data packet transmission path

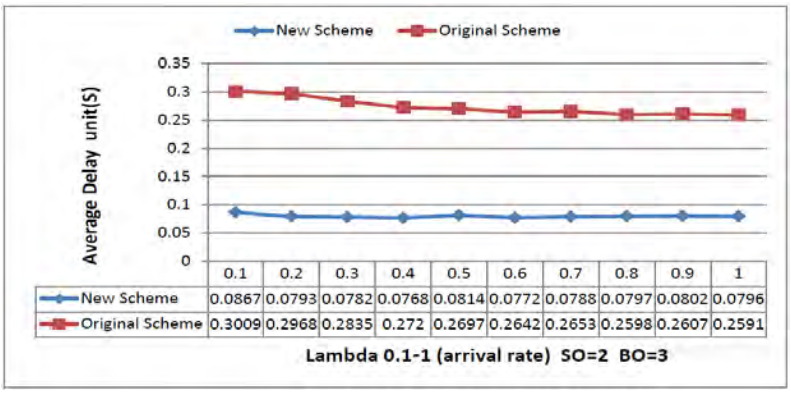


Fig. 6 Packet average transmission delay (SO=2 BO =3)

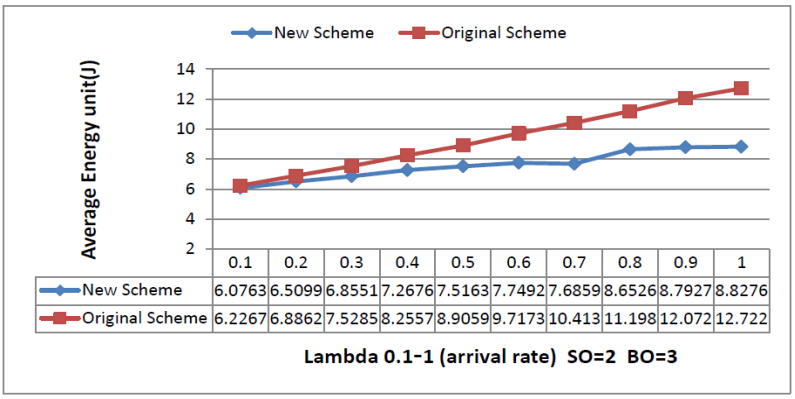
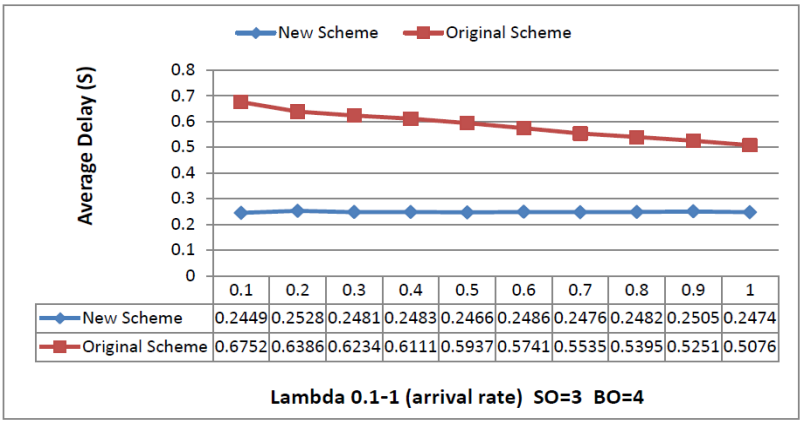
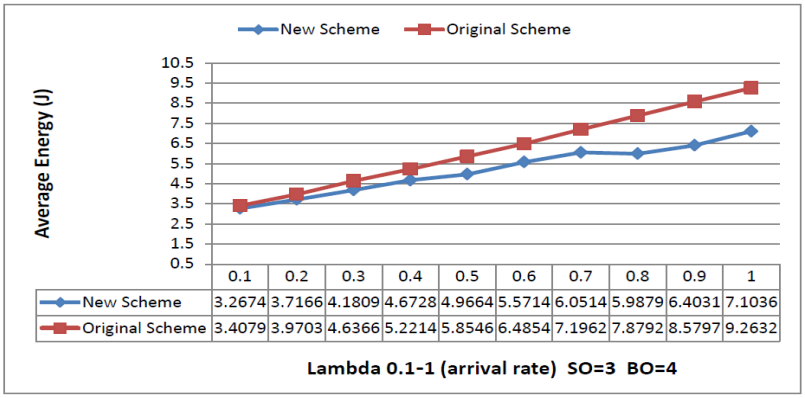


Fig. 7 Average energy consumption (SO=2 BO =3)



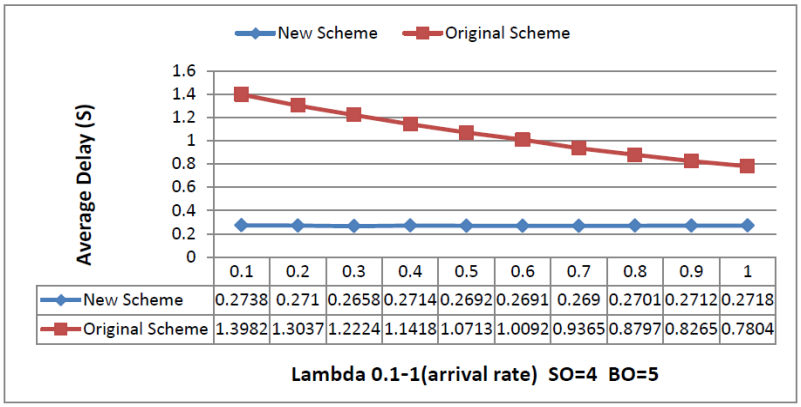
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
New Scheme	0.2449	0.2528	0.2481	0.2483	0.2466	0.2486	0.2476	0.2482	0.2505	0.2474
Original Scheme	0.6752	0.6386	0.6234	0.6111	0.5937	0.5741	0.5535	0.5395	0.5251	0.5076

Fig. 8 Packet average transmission delay (SO=3 BO =4)



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
New Scheme	3.2674	3.7166	4.1809	4.6728	4.9664	5.5714	6.0514	5.9879	6.4031	7.1036
Original Scheme	3.4079	3.9703	4.6366	5.2214	5.8546	6.4854	7.1962	7.8792	8.5797	9.2632

Fig. 9 Average energy consumption (SO=3 BO =4)



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
New Scheme	0.2738	0.271	0.2658	0.2714	0.2692	0.2691	0.269	0.2701	0.2712	0.2718
Original Scheme	1.3982	1.3037	1.2224	1.1418	1.0713	1.0092	0.9365	0.8797	0.8265	0.7804

Fig. 10. Packet average transmission delay (SO=4 BO =5)

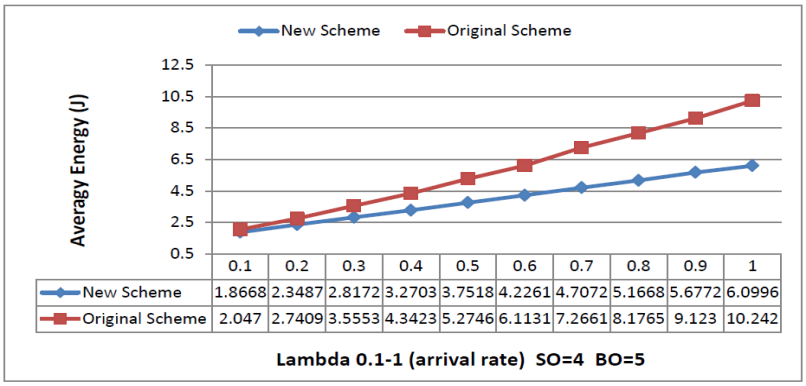


Fig. 11 Average energy consumption (SO=4 BO =5)

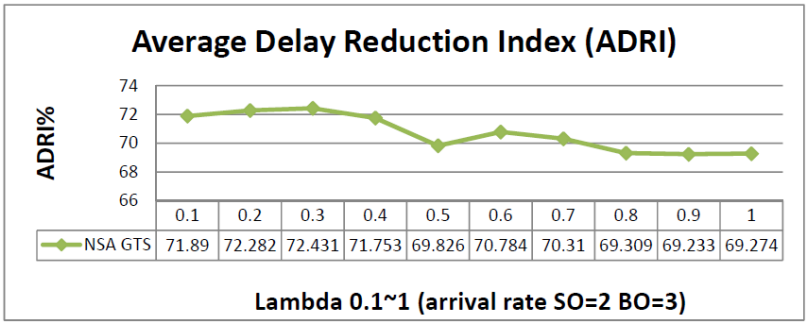


Fig. 12 SO=2 BO=3 ADRI

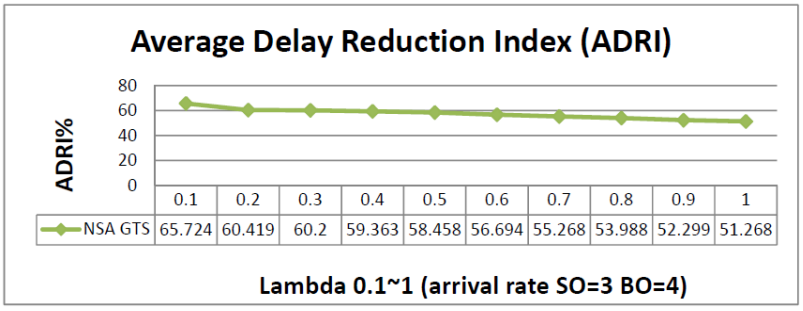


Fig. 13 SO=3 BO=4 ADRI

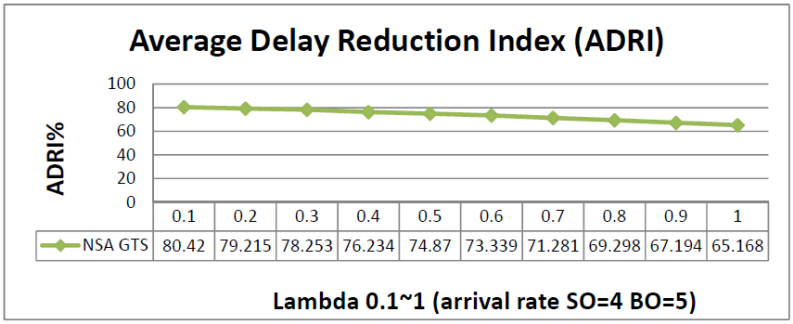


Fig. 14 SO=4 BO=5 ADRI

5. Conclusions

This paper proposes a new pre-allocation mechanism to resolve the drawbacks of traditional time slot access mechanism in the IEEE 802.15.4 specification. This method ensures each device to allocate a fix bandwidth based on the pre-allocat on cenar o. Our ro o e metho oe n't nee control mechanism to obtain guarantee time slot such that the control packet can be reduced to decrease delay and power consumption. We adopts pre-allocation schema to allocate the time slot in advance to increase the utilization of bandwidth and keep data transmission in real-time manner. Based on the experimental results, the proposed GTS mechanism is better than the IEEE 802.15.4 for the delay effect and energy consumption. To further enhance the performance of our method, we will propose an algorithm to improve the counting of sensor node number more precisely so that the power consumption and transmission delay can be optimized.

References

1. L. Q. Zhuang, K. M. Goh and J. B. Zhang：The wireless sensor networks for factory automation: Issue and challenges. In Proceeding of Emerging Technologies & Factory Automation, 141-148. (2007)
2. K. Sha, W. Shi and O. Watkins: Using Wireless Sensor Networks for Fire Rescue Applications: Requirements and Challenges. In Proceeding of Electro/information Technology, 239-244.(2006)
3. Hu. Jingtao: The Application of Wireless Sensor Networks to In-Service Motor Monitoring and Energy Management. In Proceeding of Intelligent Networks and Intelligent Systems, 165-169. (2008)
4. Jin Wang, Tinghuai Ma, Jinsung Cho and Sungoung Lee: An Energy Efficient and Load Balancing Routing Algorithm for Wireless Sensor Networks. In Computer Science and Information Systems Journal, Vol. 8, No. 4, 991-1007. (2011)

5. IEEE 802 Working Group: Standard for Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low Rate Wireless Personal Area Networks (LR-WPANs). ANSI/IEEE 802.15.4 (2003)
6. T. Kim, D. Kim, N. Park, S. E. Yoo and T. S. Lopez: Shortcut tree routing in ZigBee networks. In Proc. IEEE International Symposium on Wireless Pervasive Computing, San Juan, Puerto Rico (2007)
7. Jae Yeol Ha, Kim, T.H., Hong Seong Park, Sunghyun Choi and Wook Hyun Kwon: An Enhanced CSMA-CA Algorithm for IEEE 802.15.4 LR-WPANs. In Communications Letters, Vol. 11, No. 5, 461- 463. (2007)
8. Liang Cheng, A. G. Bourgeois and Xin Zhang: A new GTS allocation scheme for IEEE 802.15.4 networks with improved bandwidth utilization. In Proceeding of Communications and Information Technologies, 1143-1148. (2007)
9. A. Koubaa, M. Alves and E. Tovar: i-GAME: an implicit GTS allocation mechanism in IEEE 802.15.4 for time-sensitive wireless sensor networks. In Proceeding of the 18th Euromicro Conference on Real-Time Systems, 183-192. (2006)
10. Der-Chen Huang, Yi-Wei Lee and Hsiang-Wei Wu: A Cluster-Tree-Based GTS Allocation Scheme for IEEE 802.15.4 MAC Layer. In the third International Workshop on Mobility Modeling and Performance Evaluation (MoMoPE), Palermo, Italy. (2012)
11. Saeyoung Ahn, Jaejoon Cho and Sunshin An: Slotted Beacon Scheduling Using ZigBee Cskip Mechanism. In the second International Conference on Sensor Technologies and Applications, 103-108. (2008)
12. Liang Cheng and Bourgeois, Anu G.: Energy efficiency of different data transmission methods in IEEE 802.15.4: study and improvement. In the 2nd International Symposium on Wireless Pervasive Computing. (2007)

Der-Chen Huang received the B.S. degree in electronic engineering from Fung-Chia University, Taiwan, in 1983, the M.S. degree in computer engineering from Florida Institute of Technology, U.S.A, in 1991, and the Ph.D. degree in computer engineering from the Department of Computer Science and Information Engineering, Chung-Cheng University, Chiayi, Taiwan, R.O.C. in 2000. From 1983 to 1989, he worked as a design engineer with the Computer Communication Lab. (CCL)/Industrial Technology Research Institute (ITRI) and Chung-Shan Institute and Science of Technology (CSIST) when he was assigned to a partnership project at General Dynamics, Fort Worth, Texas. U.S.A. He was an associate professor with the Department of Electronic Engineering, National Chin-Yi University of Technology, Taichung, Taiwan, R.O.C. from 1991 to 2004. In 2004, he joined the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, R.O.C.. He was a director of Computer and Information Center of Chung Hsing University from 2007 to 2011. Currently, he is a professor of Chung Hsing University. Dr, Huang served as a reviewer for various technical journal and conferences and a member of editorial board of Journal of Internet Technology. He received the Best Paper Award from the 5th International Conference on Future Information Technology, Korea, in 2010. His research interests include VLSI design for testability and diagnosis, VLSI Digital Signal Process, Communication, Information Technology and Medical Image.

Yi-Wei Lee received the M.S. degree from the Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, R.O.C.. His research interests include computer network and communication.

Hsiang-Wei Wu received his B.S. degree from the Department of Electronic Engineering of National Chin-Yi University of Technology and M.S. degree from the Department of Electronic Engineering of Fung-Chia University, Taichung, Taiwan, R.O.C.. Currently, he is Ph.D. Candidate of the Department of Computer Science and Engineering, National Chung Hsing University. His research interests include computer network and communication.

Received: July 07, 2012; Accepted: January 15, 2013

DOI: 10.2298/CSIS121104028W

Efficient Verifiable Fuzzy Keyword Search over Encrypted Data in Cloud Computing

Jianfeng Wang[1], Hua Ma[1], Qiang Tang[2], Jin Li[3],
Hui Zhu[4,5], Siqi Ma[6], and Xiaofeng Chen[4⋆]

[1] Department of Mathematics, Xidian University, China
wjf01@163.com,ma_hua@126.com
[2] APSIA group, SnT, University of Luxembourg
6, rue Richard Coudenhove-Kalergi, L-1359 Luxembourg
qiang.tang@uni.lu
[3] School of Computer Science, Guangzhou University, China
jinli71@gmail.com
[4] State Key Laboratory of Integrated Service Networks,
Xidian University, China
xfchen@xidian.edu.cn
[5] Network and Data Security Key Laboratory of Sichuan
Provincezhuhui@xidian.edu.cn
[6] School of Computer Science and Technology,
Xidian University, China
xdmasiqi@hotmail.com

Abstract. As cloud computing becomes prevalent, more and more sensitive data is being centralized into the cloud by users. To maintain the confidentiality of sensitive user data against untrusted servers, the data should be encrypted before they are uploaded. However, this raises a new challenge for performing search over the encrypted data efficiently. Although the existing searchable encryption schemes allow a user to search the encrypted data with confidentiality, these solutions cannot support the verifiability of searching result. We argue that a cloud server may be selfish in order to save its computation ability or bandwidth. For example, it may execute only a fraction of the search and returns part of the searching result. In this paper, we propose a new verifiable fuzzy keyword search scheme based on the symbol-tree which not only supports the fuzzy keyword search, but also enjoys the verifiability of the searching result. Through rigorous security and efficiency analysis, we show that our proposed scheme is secure under the proposed model, while correctly and efficiently realizing the verifiable fuzzy keyword search. The extensive experimental results demonstrate the efficiency of the proposed scheme.

Keywords: searchable encryption, verifiable fuzzy search, cloud computing.

⋆ The corresponding author: Xiaofeng Chen, xfchen@xidian.edu.cn

1. Introduction

As cloud computing becomes prevalent, storage outsourcing is widely used to reduce operational costs or private backups. By outsourcing their data in the cloud, data owners can obtain high quality data storage services, while reducing the burden of data storage and maintenance. To securely store the outsourced data on an untrusted cloud server, sensitive data should be encrypted before outsourcing [16], [18]. However, it is intractable for data owners to search the encrypted data in the server efficiently. The trivial solution of downloading the whole database and decrypting locally is clearly impractical, due to the huge amount of communication and computation cost [20]. Moreover, data owners may share their outsourced data with a large number of users. The individual users might want to only retrieve certain specific data files they interested in during a given session [15]. It is desirable to support the searching functionality on the server side, without decrypting the data and loss of data confidentiality. A popular method is searchable encryption, which can offer the user to selectively retrieve files through keyword-based search. In addition, the keyword privacy should be protected effectively since keyword usually contains important information of the data files.

Although various searchable encryption schemes have been proposed to perform search securely and effectively without decrypting the data files, it is assumed that the server is "honest-but-curious". Specifically, the cloud server will follow our proposed protocol, but try to find out as much secret information as possible based on their possessions. However, we noticed that the cloud server may be selfish in order to save its computation ability or bandwidth, which is significantly beyond the conventional "honest-but-curious" server model. We consider a stronger adversary called "semi-honest-but-curious" server [9]. That is, the server may execute only a fraction of the search and returns part of the searching result honestly. Chai et al. firstly addressed this problem and proposed a verifiable keyword search scheme (VSSE) in [9]. In their solution, when the search behavior is completed, the server needs to prove to the user that the search result is correct and complete, which is named as `verifiable searchability`. However, the solution only supports the exact keyword search. In 2010, Li et al. [15] proposed a fuzzy keyword search scheme over encrypted data in cloud computing. However, they have not considered the issue of verifiable keyword search.

In this paper, we propose a new efficient verifiable fuzzy keyword search scheme, which not only supports verifiable fuzzy keyword search, but also reduces the verifying computation cost to $O(1)$. Specifically, our contribution can be summarized as follows:

- To the best of our knowledge, we propose the first verifiable fuzzy keyword search (VFKS) scheme, which not only enables fuzzy keyword search over encrypted data, but also maintains keyword privacy and the verifiability of the searching result.
- Through rigorous security analysis, our solution is secure and privacy preserving, while supporting the verifiability of the searching result.

– Our solution is highly efficient. For each query, the verifying computation cost is a constant complexity. Compared with the solution in [9], we reduce theverifying computation cost from $O(L)$ to $O(1)$, where L is the length of the searched keyword.

The organization of this paper is as follows. The related works are analyzed in Section 2 . Some preliminaries are given in Section 3.The proposed verifiable fuzzy keyword search scheme is given in Section 4. The extension scheme in hybrid cloud is given in Section 5. The security and performance analysis is given in Section 6 and 7. Finally, conclusion will be made in Section 8.

2. Related Work

Recently, plaintext fuzzy keyword search solutions have been proposed [4], [14], [13]. These solutions are based on approximate string matching techniques, which allow user to search without using try-and-see approach for finding relevant information. At a first glance, it seems possible for one to directly apply these string matching algorithms to the context of searchable encryption by computing the trapdoors on a character base within an alphabet. However, this trivial construction suffers from the dictionary and statistics attacks and fails to achieve the search privacy.

Searchable encryption is a broad concept that deals with searches in encrypted data. The goal is to outsource encrypted data and be able to conditionally retrieve or query data without having to decrypt all the data [2]. Traditional searchable encryption schemes (SSE) [3], [5], [6], [7], [10], [11], [12], [19] have been proposed in recent years. Among those works, most are focused on efficiency improvements and security definition formalizations. The first practical Searchable encryption scheme in the symmetric setting was proposed by Song et al. [19] in 2000. In their solution, each word of a document is encrypted independently with a special two-layered encryption construct. Unfortunately, the scheme is not secure against statistical analysis across multiple queries and can leak the positions of the queried keywords in a document. The searching overhead is linear to the whole file collection length. To achieve more efficient search, Goh [12] proposed to use Bloom filters to construct the index for each file. The index makes the search scheme independent of the file encryption. Moreover, the complexity of each search request is roughly proportional to the number of files in the collection. Chang et al. [10] developed a similar per-file index scheme. Curtmola et al. [11] presented the formal security notion of searchable encryption. Furthermore, they proposed similar “index” approaches, where a single encrypted hash table index is built for the entire file collection. In the index table, each entry consisting of the trapdoor of a keyword and an encrypted set of related file identifiers. Bao et al. [3] proposed a searchable encryption scheme in multi-user setting, where a group of users share data in a way that can contribute searchable contents and can search an encrypted file collection without sharing their secrets.

Searchable encryption has also been studied in the asymmetric setting. The first public-key based searchable encryption scheme is presented by Boneh et al. [6] in 2004, where anyone with the public key can encrypt data but only authorized users with the private key are able to search. Subsequently, Abdalla et al. [1] proposed a novel public-key encryption with temporary keyword search. Compared to symmetric searchable encryption, public key solutions are usually very computationally expensive.

All existing secure index based schemes support only exact keyword search. Hence, such schemes are not suitable for cloud computing. Li et al. [15] proposed the first fuzzy keyword search over encrypted data in cloud computing, which utilized the multi-way tree to enhance the search efficiency. However, note that the semi-honest-but-curious cloud server may be selfish in order to save its computation ability or bandwidth. It may execute only a fraction of the search and returns part of the searching result honestly. To solve this problem, Chai et al. [9] proposed a verifiable SSE (VSSE) scheme, which ensures that the user can verify the correctness and completeness of the search result.

3. Preliminaries

3.1. Notions

$C = (F_1, F_2, \ldots, F_n)$: a set of n encryption files;
$W = \{w_1, w_2, \ldots, w_p\}$: the set of distinct keywords of C;
$ID\{F_i\}$: the identifier of document F_i;
ID_{w_i}: the identifiers of documents containing the keyword w_i;
$F_{K,\cdot}$: a pseudo-random function; defined as $\{0,1\}^\star \times K \rightarrow \{0,1\}^l$;
$\{T_{w'}\}$: the trapdoor set of all fuzzy keywords of $w^{'} \in S_{w,d}$;
$\triangle = \{\alpha_i\}$: the predefined symbol set, where $|\triangle| = 2^n$, and $\alpha_i \in \triangle$ can be denoted by n bits;
G_W: a tree covering all fuzzy keywords of $w \in W$ is built up based on symbols in $|\triangle|$;
$T_w[i]$: the i-th symbol of the symbol sequence of trapdoor T_w;
$ord(T_w[i])$: the alphabetic order of the character $T_w[i]$ in $\triangle$;

3.2. Definitions

A verifiable fuzzy keyword search scheme (VFKS) consists of the polynomial-time algorithms ($\mathbf{Keygen}$, $\mathbf{Buildindex}$, $\mathbf{Trapdoor}$, $\mathbf{Search}$), which are similar to those of standard symmetric searchable encryption scheme (SSE), as well as a new algorithm $\mathbf{Verify}$. These algorithms are defined as follows:

– $\mathbf{Keygen}(\lambda)$: This algorithm is run by the data owner to setup the scheme. It takes a security parameter λ as input, and outputs the trapdoor generation key sk and secret key k.

- $\mathbf{Buildindex}(sk, W)$: This algorithm is run by the data owner to create the index. It takes a secret sk and the distinct keyword set W of the document collection C as inputs, and outputs a symbol-tree G_W.
- $\mathbf{Trapdoor}(sk, S_{w,d})$: This algorithm is run by the user to generate trapdoors for all fuzzy keywords of the user input keyword w. It takes a secret key sk and a fuzzy keyword set $S_{w,d}$ as inputs, and outputs a trapdoor set $\{T_{w'}\}_{w' \in S_{w,d}}$.
- $\mathbf{Search}(G_W, \{T_{w'}\})$: This algorithm is run by the server in order to search for the files in C that contain keyword w. It takes the symbol-tree G_W of the file collection C and a trapdoor set $\{T_{w'}\}$ of the fuzzy keyword set $S_{w,d}$ as inputs, and if search is successful outputs ID_w and the $proof$, otherwise outputs the $proof$.
- $\mathbf{Verify}(k, proof)$: This algorithm is run by the user to test whether the server is honest. It takes a secret k and $proof$ as inputs, and outputs $True$ if pass, otherwise outputs $False$.

Edit Distance Edit distance is a measure of similarity between two strings. The edit distance $ed(w_1, w_2)$ between two words w_1 and w_2 is the minimum number of operations required to transform one to the other. There are three primitive operations. (1) Substitution:changing one character to another in a word; (2) Deletion: deleting one character from a word; (3) Insertion: inserting a single character into a word. Given a keyword w, we let $S_{w,d}$ denote the set of keywords $w^{'}$ satisfying $ed(w, w^{'}) < d$ for a certain integer d.

Trapdoors of Keywords Trapdoors of the keywords are realized by applying a hash function f as follows: Given a keyword w, we compute the trapdoor of w as $T_w = f(sk, w)$, where the sk is the user's index generation key.

Verifiable of Keyword Search The server executes search for the user when receiving the search request, and returns the search result and the $proof$. If the server executes all operations honestly, the probability that the search result is incorrect should be negligible; but if the server just returns a fraction of the search result honestly, the user can detect the cheating behavior with overwhelming probability through the $verify$ algorithm.

3.3. System Model

In this paper, we consider a cloud data-outsourcing system, which consists of three different entities: the data owner, the user and the cloud server. The data owner has a collection of n encrypted data files $C = (F_1, F_2, \cdots, F_n)$ to be stored in the cloud server. A predefined set of distinct keywords in C is denoted as $W = (w_1, w_2, \cdots, w_p)$. The cloud server performs fuzzy search for the authorized users over the encrypted data. We assume the authorization between the

data owner and users is appropriately done. In the initialization phase, the data owner shares the trapdoor generation key sk with authorized users, and builds an index G_W for the encrypted file collection C together with the encrypted files outsourcing to the cloud server. To search the file collection for any keyword w, an authorized user generates the trapdoor of w, and sends it to the cloud server. Upon receiving the search request by the user, the server performs the search over the index G_W and returns all the encrypted files containing the specific keyword w. For the fuzzy keyword search, the server returns the closest possible results based on pre-specified similarity metrics. An architecture of verifiable fuzzy keyword search is shown in Fig. 1.

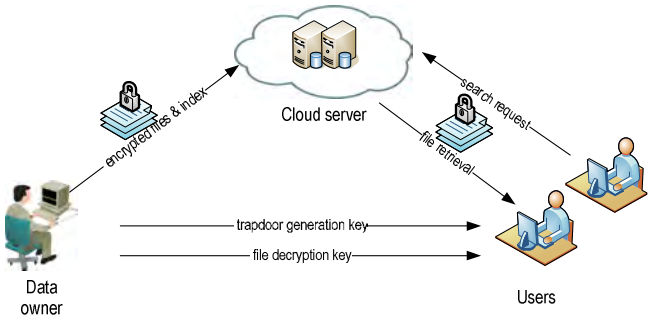


Fig. 1. Architecture of verifiable fuzzy keyword search

3.4. Security Model

In this work, we consider a "semi-honest-but-curious" cloud server, which is different with most previous searchable encryption schemes. We assume the cloud server acts in an "semi-honest" fashion, that is to say, it may not correctly follow our proposed protocol but forge part of search result or execute only a fraction of searching operations honestly. In addition, the cloud server tries to analyze the message flow received during the protocol in order to learn additional information. When designing verifiable fuzzy keyword search scheme, we follow the security definition deployed in the traditional searchable encryption [11]. Namely, it is required that noting should be leaked from the remotely stored files and index beyond the outcome and the pattern of search queries.

3.5. Design Goals

To support verifiable fuzzy keyword search over encrypted data using the above system and security models, our system design should achieve the following design goals: 1) to construct storage-efficient fuzzy keyword set and design efficient and effective fuzzy keyword search scheme; 2) to prevent the server

from learning either the data files or the searched keywords beyond the search pattern and the access pattern; 3) to design efficient verifiable fuzzy keyword search scheme and enable user to verify the correctness and completeness of search result.

4. A New Verifiable Fuzzy Keyword Search Scheme

4.1. Construction of the VFKS scheme

In this section, we present the proposed scheme in detail. We assume the data files are separately encrypted by a symmetric cipher in a conventional manner before the user build the index. Our scheme consists of five algorithms $(\mathbf{KeyGen}, \mathbf{Buildindex}, \mathbf{Trapdoor}, \mathbf{Search}, \mathbf{Verify})$.

- $\mathbf{Keygen}$
  In this process, the data owner generates the index generation key sk and a secret key k. The $\mathbf{Keygen}$ is a randomized key generation algorithm, which generate the key in this way : $sk, k \xleftarrow{R} \{0,1\}^k$.

Algorithm 1 Generate Fuzzy Set (w_i, d)

Input: Keyword w_i and Edit distance d
Output: Fuzzy keyword set $S_{w_i,d}$

```
if d >= 1 then
    Generate Fuzzy Set (w_i, d − 1);
end if
if d = 0 then
    Set S_{w_i,d} = {w_i};
else
    for k ← 1 to |S_{w_i,d}| do
        for j ← 1 to 2 × |S_{w_i,d}[k]| + 1 do
            if j is odd then
                Set Temp = |S_{w_i,d}[k]|;
                Insert ⋆ at position j + 1/2;
            else
                Set Temp = |S_{w_i,d}[k]|;
                Replace j/2-th character with ⋆;
            end if
            if Temp is not in S_{w_i,d−1} then
                Set S_{w_i,d} = S_{w_i,d} ∪ {Temp};
            end if
        end for
    end for
end if
return S_{w_i,d}
```

- **Buildindex**
  In this process, we utilize a symbol-based trie-traverse search scheme, where a multi-way tree is constructed for storing the fuzzy keyword set $\{S_{w_i,d}\}_{w_i \in W}$ over a finite symbol set. The key idea behind this construction is that all trapdoors sharing a common prefix may have common nodes. The root is associated with an empty set and the symbols in a trapdoor can be recovered in a search from the root to the leaf that ends the trapdoor. All fuzzy keywords in the trie can be found by a depth-first search. Assume $\Delta = \{\alpha_i\}$ is a predefined symbol set, where the number of different symbols is $|\Delta| = 2^n$ and each symbol $\alpha_i \in \Delta$ can be denoted by n bits.

  1. **Initialization**
     - The data owner scans the C and builds W, the set of distinct keywords of D.
     - The data owner outsources the encryption file collection D to the server and receives the identifiers of each file (denote as $ID\{F_i\}$). For all files of containing the keyword w_i, denote the identifier set as $ID_{w_i} = ID\{F_1\} \| ID\{F_2\} \ldots, \| ID\{F_i\}$.
  2. **Build Fuzzy Keyword Set**
     To build a storage-efficient fuzzy keyword set, we utilize the wildcard technique proposed in [15]. The idea is to consider the positions of the three primitive edit operations. Namely, we use a wildcard " $\star$ " to denote all edit operations at the same position. The wildcard-based fuzzy keyword set of w_i with edit distance d is denoted as $S_{w_i,d} = \{S^{'}_{w_i,0}, \cdots, S^{'}_{w_i,d}\}$, where $S^{'}_{w_i,d}$ denotes the set of keywords $w_i^{'}$ with d wildcards. For example, for the keyword cat with the pre-set edit distance 1, its wildcard-based fuzzy keyword set can be constructed as $S_{cat,1} = \{cat, \star cat, \star at, c\star at, c\star t, ca\star t, ca\star, cat\star\}$. The procedure for the wildcard fuzzy keyword set construction is shown in Algorithm 1.
  3. **Build Symbol-based Index Tree**
     - The data owner computes $T_{w_i^{'}} = f(sk, w_i^{'})$ for each $w_i^{'} \in S_{w_i,d} (1 \leq i \leq p)$ with the index generation key sk. Then he divides the hash value into a sequence of symbols as $\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_{l/n}}$, where l is the output length of one-way function $f(x)$.
     - The data owner builds up a trie G_W covering all the fuzzy keywords $w_i \in W$. each node in G_W has a two-tuples (r_0, r_1), r_0 stores the symbol; r_1 stores a globally unique value $path||mem||F_k(path||mem)$, where $F_k(\cdot)$ is a pseudo random function. The $path$ contains a sequence of symbols from root to the current node and the mem is a bitstream of length 2^n, which represents the set of children of the current node. For example, if the current node is a child of root and has only one child whose r_0 is the i-th symbol in Δ, the i-th bit of the bitstream of length 2^n is set to "1" while other bit positions are set to zero. That is, $path = \alpha_i$, $mem = 0 \cdots 1 \cdots 0$. The r_1 of leaf nodes is different to the other nodes, the r_1 can be denote as $r_1 = path||ID_{w_i}||F_K(path||ID_{w_i})$

- The data owner attaches $\{ID_{w_i}|g_k(ID_{w_i})\}_{1\leq i\leq p}$ to G_W and outsources G_W with encrypted files to the cloud server.

- **Trapdoor**
  - For search input w, the user generates the fuzzy keyword set $S_{w,d}$ using the Algorithm 1;
  - For each $w^{'} \in S_{w,d}$, the user computes $T_{w'} = f(sk, w^{'})$ and sends $\{T_{w'}\}_{w'\in S_{w,d}}$ to the cloud server. Meanwhile, the user needs to temporary storage the $\{T_{w'}\}_{w'\in S_{w,d}}$, which is used during the verify process.

Algorithm 2 Searching Tree $(G_W, \{T_{w'}\})$

Input: A trapdoor set $\{T_{w'}\}$ and The index tree G_W
Output: The set of proof $ProofSet$ and The set of files ID $IDSet$;

```
for i ← 1 to |{T_w'}| do
  Set currentnode as root of (G_W;
  for j ← 1 to l/n do
    Set α as α_ij in the i-th T_w';
    if no child of currentnode contains α then
      Append currentnode.proof to ProofSet;
      break;
    end if
    Set currentnode as child containing α;
  end for
  if currentnode is leafnode then
    Append currentnode.proof to ProofSet;
    Append currentnode.ID to IDSet;
    if i = 1 then
      return ProofSet and IDSet;
    end if
  end if
end for
return ProofSet and IDSet;
```

- **Search**
  Upon receiving the search request, the server divides each $T_{w'}$ into a sequence of symbols, then performs the search over G_W using Algorithm 2 and returns the file identifiers ID_{w_i} and $proof$ to the user. According to the ID_{w_i}, the user can retrieve the files of his interest. Note that the $proof$ is the r_1 of each node, which is a globally unique value.
- **Verify**
  In this process, we introduce the method of verifying the searching result. The idea is that each node in G_W has a globally unique value, called $proof$. Due to the construction of G_W, the path of each node is unique, without the secret key k, the attacker can not forge a valid $proof$. The data owner shares the k with all authorized users. The authorized user can verify the correctness of the search result by reconstructing the $proof$.

- When the search is successful, firstly, the user utilizes the $IDSet$ to test the completeness of search result. Specifically, he computes $g_k(\hat{ID}_w)$ and tests whether $g_k(\hat{ID}_w)$ is equal to the received $g_k(ID_w)$, where $\hat{ID}_w$ is the concatenation of identifiers received by the user. If pass, then he utilizes the $ProofSet$ to test the correctness of search result. Similarly, the user computes $F_k(path||\hat{|}mem)$ and test whether $F_k(path||\hat{|}mem)$ is equal to the received $F_k(path||mem)$, where $path||\hat{|}mem$ is the former part of r_1 of the current node returned by the server. If $F_k(path||\hat{|}mem)$ is not equal to the received $F_k(path||mem)$, the user can defect that the server is not honest.
- When the search is not successful, the user directly tests the correctness of searching result. The process contains two steps:
  * The user tests whether $F_k(path||\hat{|}mem) = F_k(path||mem)$, if not equal, the user can defect that the server is not honest.
  * If step 1 pass, he tests whether $mem[ord(T_w[i+1])] = 1$, where the $T_w[i+1]$ is the next character of the current node in the symbol sequence of the trapdoor. If not equal, the user can defect that the server is not honest.

4.2. Performance Comparison

We compare the proposed algorithm with Li's scheme [15] and Chai's scheme [9]. To make the comparison easier, we assume that N is the total number of keywords and M is the maximum size of the fuzzy keyword set $S_{w_i,d}$. Table 1 presents the comparison of the search efficiency and the verifiability among the above schemes.

Table 1. Comparison of the three scheme

	Li's scheme [15]	Chai's scheme [9]	Our scheme
Storage cost	$O(MN)$	$O(N)$	$O(MN)$
Search cost	$O(1)$	$O(L)$	$O(1)$
Verifiable searchability	No	Yes	Yes
Fuzzy searchability	Yes	No	Yes
Verify cost	-	$O(L)$	$O(1)$

Compared with Li's scheme, our scheme achieve verifiable of search result. The verify cost shows the computation of verifying per trapdoor in each query. In Chai's scheme, it requires L times decryption operations while in our scheme the computation is only one hash operation, where the L is the length of the searched keyword. Note that our scheme can reduce the computation from $O(L)$ to $O(1)$, due to the query is performed frequently, we can reduce a large amount of computation at the user.

Though our scheme need more space to store the fuzzy keyword($O(MN)$), it achieves the the fuzzy searchability. All in all, our scheme not only supports the fuzzy search but also achieves the verifiable searchability more efficiently.

5. Verifiable Fuzzy Keyword Search in Hybrid Cloud

5.1. System and Security Model

Recently, Bugiel et al. [8] provided an architecture consisting of two clouds for secure outsourcing of data and arbitrary computations to an untrusted commodity cloud. Their approach consists of a private cloud and a public cloud. The private cloud performs the security-critical operations, whereas the public cloud performs the performance-critical ones. This allows maximum utilization of the expensive resources of the private Cloud, while high loads of queries can be processed on-demand by the public Cloud. Based on their two clouds architecture, we consider to address the privacy-preserving fuzzy keyword search problem simultaneously supporting verifiability of search result in hybrid cloud model.

In this section, we consider a extension from single cloud model to hybrid cloud model. There are four entities defined in the hybrid cloud model, that is, the data owner, the user, the private cloud and the public cloud. The data owner outsources the encrypted files to the public cloud and shares them with the authorized users. The user performs fuzzy keyword search and decrypts the encrypted files retrieved from the public cloud. The private cloud is additionally introduced to facilitate user's secure usage of cloud service. Specifically, since the computing resources at user side are restricted and the public cloud is not fully trusted in practice, the private cloud is able to provide users with an execution environment and infrastructure working as an interface between user and the public cloud. The interface offered by the private cloud allows user to securely submit files and queries to be securely stored and computed respectively. An architecture of verifiable fuzzy keyword search is shown in Fig.2.

In this model, we assume that the public cloud is "semi-honest-but-curious", which may not correctly follow our proposed protocol but forge part of search result or execute only a fraction of searching operations honestly. As for the private cloud, we assume that it is "honest-but-curious", which will follow our proposed protocol, but try to find out as much secret information as possible based on their possessions. In addition, the user's input keywords are allowed to be known by the private cloud. Actually, approximately relaxing security demands by allowing keywords leakage to private cloud is innocuous because the private cloud in practice is located in the premises of the organization [17].

5.2. Scheme Description

In single cloud model, data owner/user has to compute trapdoors for all the relevant fuzzy keywords for both index generation and search requesting, which

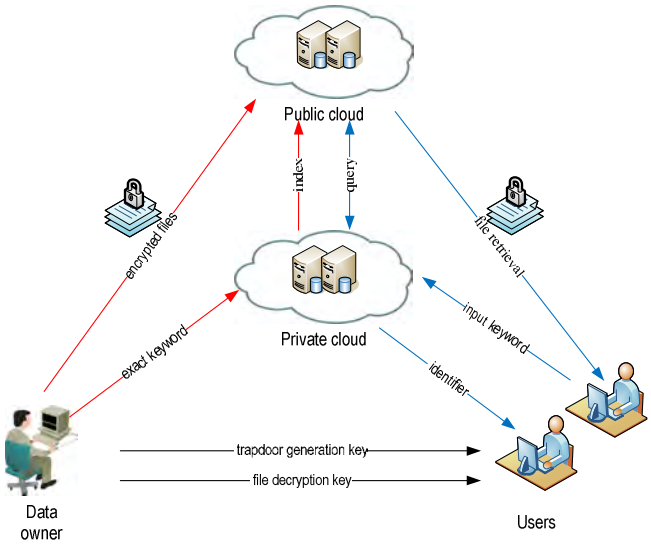


Fig. 2. Architecture of verifiable fuzzy keyword search

leads to a large amount of overhead at user side. In addition when receiving the search result retrieved from the public cloud, the user has to compute Pseudo-random function values for all the search results.

To fill the gap, we will show how to achieve efficient verifiable fuzzy keyword search under hybrid cloud model. The key idea is to outsource the expensive operation (i.e. trapdoor generation and search result verification) to the private cloud and only left the light-weight computation (file encryption and decryption) at user side.

Consider that the $Keygen$ operate identically to that in Section 4.1, we just provide the other four algorithms as follows.

- **Buildindex** When outsourcing a file F, the data owner performs an encryption on F by himself but outsources the task of generating the fuzzy keyword set $\{S_{w,d}\}_{w \in W}$ and building the index to the private cloud.
- **Trapdoor** and **Verify** When retrieving the interest files, the private cloud works as a proxy of the user. Firstly, it translates user's query into a set of trapdoors and sends to the public cloud. Later, upon receiving the search results returned by public cloud, the private cloud is to perform verification on them to test whether the public cloud is honest.
- **Search** Upon receiving the search request from the private cloud, the public cloud divides each trapdoor into a sequence of symbols, then performs the search and returns the search result to the private cloud.

Note that in hybrid cloud model, verifiable fuzzy keyword search can be presented soundly and efficiently. For soundness, since the private cloud possesses all the data owner/ user's resources except for the file encryption key sk,

it can perform the operations of index generation and search result verification. Moreover, due to the private cloud performs all the overhead operations for the data owner/user, the whole process can be processed more efficiently.

6. Security Analysis

In this section, we prove the correctness and security of the proposed verifiable fuzzy keyword search scheme.

Theorem 1. *The intersection of the fuzzy sets $S_{w,d}$ and $S_{w_i,d}$ for keyword w and w_i is not empty if and only if $ed(w, w_i) \leq d$.*

Proof. First, we prove that $S_{w,d} \bigcap S_{w_i,d}$ is not empty if $ed(w, w_i) \leq d$. To prove this, it is enough to find out an element in $S_{w,d} \bigcap S_{w_i,d}$. According to the definition of edit distance, we can transform w to w_i after $ed(w, w_i)$ edit operations. From w, we get an element $w^\star$ by marking the positions of those $ed(w, w_i)$ operations on w as $\star$. From $w^\star$, we can perform the $ed(w, w_i)$ edit operations on the same positions containing $\star$ at $w^\star$ and transform $w^\star$ to w_i. Since $ed(w, w_i) \leq d$, $w^\star$ is an element in both $S_{w,d}$ and $S_{w_i,d}$.

Next, we prove that $ed(w, w_i) \leq d$ if $S_{w,d} \bigcap S_{w_i,d}$ is not empty. We use $w^\star$ to denote the common element in $S_{w,d} \bigcap S_{w_i,d}$. Assume the number of $\star$ in $w^\star$ is k, there are two cases should be considered: If $k = 0$, it means that we do not need any edit operation to transform w to w_i. That is, $w = w_i = w^\star$. Obviously, $ed(w, w_i) = 0 \leq d$. If $k > 0$, for any $\star$ in $w^\star$, we can perform edit operation on the position of the $\star$ and transform it to the corresponding character in w and w_i. We use $w_1^\star$ and $w_{i_1}^\star$ to denote the result variants, respectively. Due to the two variants only have at most one position is different, we can transform $w_1^\star$ to $w_{i_1}^\star$ by at most one edit operation. That is, $ed(w_1^\star, w_{i_1}^\star) \leq 1$. After all the k $\star$ be performed in $w^\star$, we get w and w_i, respectively. Due to $w^\star \in S_{w,d} \bigcap S_{w_i,d}$, the number of $\star$ in $w^\star$ is not greater than d. we get that $ed(w, w_i) = k \leq d$.

Theorem 2. *The verifiable fuzzy keyword search scheme is secure regarding the search privacy.*

Proof. Similar to [15], suppose the proposed scheme is not achieve the index privacy against the indistinguishability under the chosen keyword attack (IND-CPA), which means there exists an algorithm $\mathcal{A}$ who can get the underlying information of keyword from the index. Then, we build an algorithm $\mathcal{A}^{'}$ that utilizes $\mathcal{A}$ to determine whether some function $f^{'}(\cdot)$ is a pseudo-random function such that $f^{'}(\cdot)$ is equal to $f(sk, \cdot)$ or a random function. $\mathcal{A}^{'}$ has an access to an oracle $O_{f'(\cdot)}$ that takes as input secret value x and return $f^{'}(x)$. Upon receiving any request of the index computation, $\mathcal{A}^{'}$ answers it with request to the oracle $O_{f'(\cdot)}$. After making these trapdoor queries, the adversary outputs two keywords w_0^* and w_1^* with the same length and edit distance, which can be relaxed by adding some redundant trapdoors. $\mathcal{A}^{'}$ picks one random $b \in \{0, 1\}$

and sends w_b^* to the challenger. Then, $\mathcal{A}^{'}$ is given a challenge value y, which is either computed from a pseudo-random function $f(sk, \cdot)$ or a random function. $\mathcal{A}^{'}$ sends y back to $\mathcal{A}$, who answers with $b^{'} \in \{0, 1\}$. Suppose $\mathcal{A}$ guesses b correctly with nonnegligible probability, which indicates that the value is not randomly computed. Then, $\mathcal{A}^{'}$ makes a decision that $f^{'}(\cdot)$ is a pseudo-random function. As a result, based on the assumption of the indistinguishability of the pseudo-random function from some real random function, $\mathcal{A}$ at best guesses b correctly with approximate probability 1/2. Thus, the search privacy is obtained.

Theorem 3. *The verifiable fuzzy keyword search scheme is secure based on the verifiable fuzzy search.*

Proof. To prove the verifiability, we need to prove that the attacker can not forge a valid $proof$. To tamper the search result, the attacker need to forge the $proof$. There are three ways: (1) generate a r_1 with different parameter $path||mem$; (2) randomly generate a r_1 to replace the original one; (3) return the r_1 of another node back to the user.

- For method (1) and (2), due to the collision resistance properties of hash function, each node in G_W has a unique r_1, the attacker can successful cheat with a negligible probability without the secret key k. That is, the attacker can not return part of the search result or some fault one.
- For method (3), According to the construction of G_W, there has a unique $path$ from root to the current node. In other words, the $path$ of any node can be called signature of the node. The r_1 with the different $path$ will be reject by the algorithm $verify$.

Baesd on the above analysis, without the secret key k, the attacker can not construct a valid $prooof$. That is, our proposed scheme is secure based on the assumption of collision resistance of hash function.

7. Performance Analysis

In this section, we analyze the efficiency of the proposed scheme based on simulation. Since the process of file encryption is independent to the process of index construction, we focus on the symbol-tree based search algorithm. Our experiment is simulated on a LINUX machine with Intel Pentium Dual Core E5800 3.2GHz and 2G memory.

Performance of Generating Fuzzy Keyword Set In our experiment, we focus on the wildcard-based fuzzy set construction for all the keywords extracted from the file collection. Fig. 3 shows the fuzzy keyword set construction time with edit distance d=1 and 2. We can see that in both cases, the construction time increases linearly with the number of keywords. The cost constructing fuzzy keyword set under d=1 is much less than the case of d=2 due to the smaller set of possible wildcard- based words.

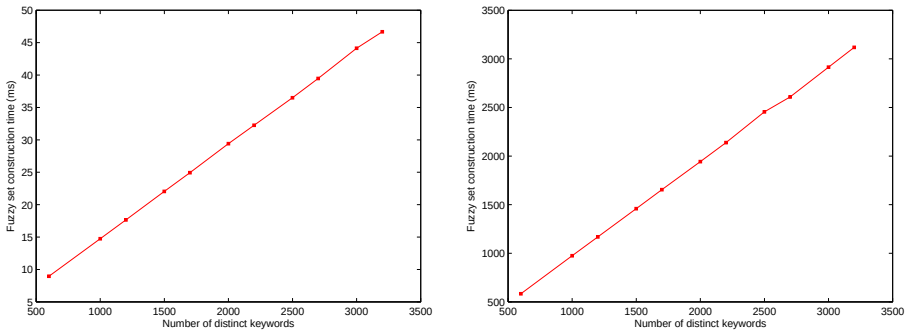


Fig. 3. Fuzzy keyword set construction time using wild-based approach:(a) d=1, (b) d=2.

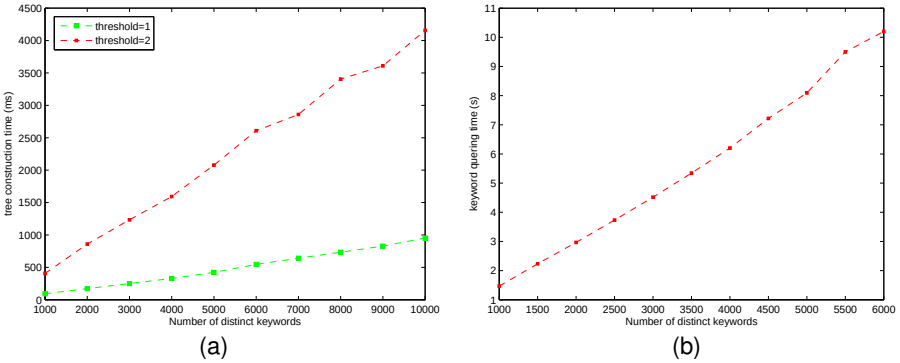


Fig. 4. The evaluation of symbol-based index tree :(a) construction time (b) searching time

Performance of Building Symbol-based Index Given the fuzzy keyword set constructed using wildcard-based approach, we measure the time cost of symbol-based Index construction. Fig. 4(a) shows the time cost of building the symbol-based index tree in the case edit distance d=1 and 2. Although the time cost is not very low,the index construction process can be conducted off-line, it will not affect the searching efficiency. Fig. 4(b) shows the time cost of a single keyword query.

8. Conclusion

In this paper, we investigated the fuzzy keyword search problem in the scenario of a semi-honest-but-curious server, which may execute only a fraction of the search and return part of the searching result honestly. We proposed a new efficient verifiable fuzzy keyword search scheme, which not only supports fuzzy keyword search over encrypted data, but also enjoys the verifiability of the searching result. Though rigorous security and efficiency analysis, we showed that our method is secure and privacy-preserving, while correctly realizing the verifiable fuzzy keyword search.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61272455 and 61100224), major national science and technology projects (No. 2012ZX03002003), the Project Supported by Natural Science Basic Research Plan in Shaanxi Province of China (No. 2011JQ8042) and the Fundamental Research Funds for the Central Universities (Nos. JY10000901034 and K50510010030).

References

1. Abdalla, M., Bellare, M., Catalano, D., Kiltz, E., Kohno, T., Lange, T., Malone-Lee, J., Neven, G., Paillier, P., Shi, H.: Searchable encryption revisited: Consistency properties, relation to anonymous ibe, and extensions. In: Proceedings of Advances in Cryptology - CRYPTO 2005. pp. 205–222 (2005)
2. Agudo, I., Nuñez, D., Giammatteo, G., Rizomiliotis, P., Lambrinoudakis, C.: Cryptography goes to the cloud. In: Lee, C., Seigneur, J.M., Park, J., Wagner, R. (eds.) Secure and Trust Computing, Data Management, and Applications, Communications in Computer and Information Science, vol. 187, pp. 190–197. Springer (2011)
3. Bao, F., Deng, R.H., Ding, X., Yang, Y.: Private query on encrypted data in multi-user settings. In: Proceedings of the 8th International Conference on Information Security Practice and Experience. pp. 71–85. Springer, Sydney, Australia (2008)
4. Behm, A., Ji, S., Li, C., Lu, J.: Space-constrained gram-based indexing for efficient approximate string search. In: Proceedings of the 25th IEEE International Conference on Data Engineering. pp. 604–615. IEEE, Shanghai, China (2009)
5. Bellare, M., Boldyreva, A., O'Neill, A.: Deterministic and efficiently searchable encryption. In: Proceedings of Advances in Cryptology - CRYPTO 2007. pp. 535–552. Springer, Santa Barbara, CA, USA (2007)
6. Boneh, D., Crescenzo, G., Ostrovsky, R., Persiano, G.: Public key encryption with keyword search. In: Proceedings of Advances in Cryptology - EUROCRYPT 2004. pp. 506–522. Springer, Interlaken, Switzerland (2004)

7. Boneh, D., Waters, B.: Conjunctive, subset, and range queries on encrypted data. In: Proceedings of the 4th Theory of Cryptography Conference. vol. 4392, pp. 535–554. Springer, Amsterdam, The Netherlands (2007)
8. Bugiel, S., Nurnberger, S., Sadeghi, A.R., Schneide, T.: Twin clouds: An architecture for secure cloud computing. In: Proceedings of the Workshop on Cryptography and Security in Clouds(WCSC'11). Zurich, Switzerland (2011)
9. Chai, Q., Gong, G.: Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. CACR, University of Waterloo (2011), [Online]. Available: http://www.cacr.math.uwaterloo.ca/techreports/2011/cacr2011-22.pdf (current December 2012)
10. Chang, Y., Mitzenmacher, M.: Privacy preserving keyword searches on remote encypted data. In: Proceedings of the 3rd Applied Cryptography and Network Security. pp. 391–421. New York, NY, USA (2005)
11. Curtmola, R., Garay, J., Kamara, S., Ostrovsky, R.: Searchable symmetric encryption: improved definition and efficient constrcutions. In: Proceedings of the 13th ACM Conference on Computer and Communications Security. pp. 79–88. ACM, Alexandria, Virginia, USA (2006)
12. Goh, E.: Secure indexes. Report 2003/216, Cryptology ePrint Archive (2003), http://eprint.iacr.org/2003/216
13. Ji, S., Li, G., Li, C., Feng, J.: Efficient interactive fuzzy keyword search. In: Proceedings of 18th International World Wide Web Conference. ACM, Madrid, Spain (2009)
14. Li, C., Lu, J., Lu, Y.: Efficient merging and filtering algorithms for approximate string searches. In: Proceedings of the 24th IEEE International Conference on Data Engineering. pp. 257–266. IEEE, Cancun, Mexico (2008)
15. Li, J., Wang, Q., Wang, C., Cao, N., Ren, K.and Lou, W.: Fuzzy keyword search over encrypted data in cloud computing. In: Proceedings of the 29th IEEE International Conference on Computer Communications(INFOCOM'10). pp. 441–445. IEEE, San Diego, CA, USA (2010)
16. Lu, Y., Tsudik, G.: Privacy-preserving cloud database querying. Journal of Internet Services and Information Security 1(4), 5–25 (2011)
17. Poisel, R., Tjoa, S.: Discussion on the challenges and opportunities of cloud forensics. In: Proceedings of the International Cross-Domain Conference and Workshop on Availability, Reliability, and Security(CD-ARES'12). pp. 593–608. Springer, Prague, Czech Republic (2012)
18. Shiraishi, Y., Mohri, M., Fukuta, Y.: A server-aided computation protocol revisited for con-fidentiality of cloud servic. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 2(2), 83–94 (2011)
19. Song, D., Wagner, D., Perrig, A.: Practical techniques for searches on encrypted data. In: Proceedings of the 2000 IEEE Symposium on Security and Privacy. pp. 44–55. IEEE, Berkeley, California, USA (2000)
20. Wang, C., Ren, K., Yu, S., Mahendra, K.: Achieving usable and privacy-assured similarity search over out-sourced cloud data. In: Proceedings of the 31th IEEE International Conference on Computer Communications(INFOCOM'12). pp. 451–459. IEEE, Orlando, USA (2012)

Jianfeng Wang is a graduate student at the Faculty of Science, in Xidian University. His research interests include public key cryptography, cloud computing security and searchable encryption.

Hua Ma received her Masters degree in Applied Mathematics from the Xidian University in 1990. She is currently a Professor at the Faculty of Science, Xidian University. Her research interests include public key cryptography, network security, and electronic commerce.

Qiang Tang is a postdoc researcher in the Interdisciplinary Centre for Security, Reliability and Trust at University of Luxembourg. Before this, he was a postdoc researcher at the Distributed and Embedded Security Research Group in the Computer Science department at University of Twente, the Netherlands. He received his MSc degree from Peking University , China in 2002 and obtained his PhD degree from Royal Holloway, University of London , UK in 2007.

Jin Li He received his B.S. (2002) and M.S. (2004) from Southwest University and Sun Yat-sen University, both in Mathematics. He got his Ph.D degree in information security from Sun Yat-sen University at 2007. Currently, he works at Guangzhou University. His research interests include Applied Cryptography and Security in Cloud Computing (secure outsourcing computation and cloud storage). He served as a senior research associate at Korea Advanced Institute of Technology (Korea) and Illinois Institute of Technology (U.S.A.) from 2008 to 2010, respectively. He has published more than 40 papers in international conferences and journals, including IEEE INFOCOM, IEEE Transaction on Parallel and Distributed Computation, IEEE Transaction on Information Forensics and Security, ESORICS etc. He also served as TPC committee for many international conferences on security. He received a National Science Foundation of China (NSFC) Grant for his research on secure outsourcing computation in cloud computing. He was selected as one of science and technology new stars in Guangdong province.

Hui Zhu associate professor, born in 1981, received the Ph.D. degree in information security from Xidian University in 2009. His current research interests include network security and security authentication protocol.

Siqi Ma is a student at the School of Computer Science and Technology, in Xidian University. Her research interests include cloud computing security and network Security.

Xiaofeng Chen received his Ph.D. in cryptography from the Xidian University in 2003. He is currently a Professor at the School of Telecommunications Enginnering, Xidian University. His research interests include public key cryptography, financial cryptography, and cloud computing security.

Received: November 4, 2012; Accepted: April 1, 2013.

DOI: 10.2298/CSIS120914029K

Traffic Deflection Method for DOS Attack Defense using a Location-Based Routing Protocol in the Sensor Network

Ho-Seok Kang[1], Sung-Ryul Kim[1,*], and Pankoo Kim[2]

[1] Division of Internet and Multimedia Engineering,
Konkuk University, Seoul, Republic of Korea
hsriver@gmail.com, kimsr@konkuk.ac.kr
[2] Department of Computer Engineering,
Chosun University, Gwangju, Republic of Korea
pkkim@chosun.ac.kr
*Corresponding Author

Abstract. As the ubiquitous computing environment gets more attention and development, WSN (Wireless Sensor Network) is getting popular as well. Especially, the development of wireless communication and sensor equipment greatly contributes to the popularization of WSN. On the other hand, the safety and security of WSN attracts lots of attention due to such a development and distribution. The DoS (Denial of Service) attack, which gets more sophisticated and broadens its domain into various services fields, may have negative effects on WSN, making it vulnerable to attacks. Since WSN collects information through sensors that are already deployed, it is difficult to have its energy recharged. When WSN is under a DoS attack, sensor nodes consume lots of energy, bringing about a fatal result to the sensor network. In this paper, we propose a method to efficiently defend against DoS attacks by modifying routing protocols in the WSN. This method uses a location based routing protocol that is simple and easy to implement. In the WSN environment where the location-based routing protocol is implemented, this method disperses the DoS attack concentration of traffic by using the traffic deflection technique and blocks it out before arriving at the target destinations. To find out the number of traffic redirection nodes proper for this method, we have performed a few experiments, through which the number of such nodes was optimized.

Keywords: sensor network, traffic redirection, filtering, location-based routing protocol, Denial of Service

1. Introduction

DoS (Denial of Service) attack is a set of methods that tries to make a target service unusable without actually hacking into the system. DoS attack has gradually developed into a method of using various attack paths as DDoS (distributed denial-of-service) attack and of attacking the entire network to which target belongs. DoS attacks can be classified into flooding, connection, and application attack types. Flooding attack can be divided into SYN/ACK flooding, TCP/IP null, FIN flooding, TCP connection, and HTTP attacks. Application attack is an attack using the characteristics of an application, and the target applications include FTP, VoIP and DNS [1]. Moreover, DoS attacks are increasing three times faster than the other attacks and are more dangerous because their implementation is relatively easier. The methods to protect against DoS attacks can be categorized roughly into four kinds: attack prevention, attack detection, attack source identification, and attack reaction [2]. Among these, the attack-prevention techniques are ways of blocking out DDoS attacks in advance, preventing them and coping with them sufficiently by locating their sources [3, 4].

DoS attacks have been limited to the existing particular networks or servers so far. However, it is recently broadening its domain to small-sized networks, such as AS and VPN, mobile networks, and even sensor networks. In fact, there has appeared a new method of DoS attack using smart phones, which actually threatens the establishment of stable mobile services. Thus, out of all the attack prevention methods, this work intends to design a system apt to defend DoS attacks in the environment of WSN, which is a network that can collect a variety of information in the ubiquitous computing environment.

WSN functions to collect a variety of information measured on particular areas, such as vehicle traffic flow, weather information, and detection systems of military or companies [5, 6]. When such a sensor network is once installed, it works independently and delivers measured information to a few sink nodes by comprising a network through connections with neighboring sensors. Each sensor node consists of a sensing device to obtain information, a processing device to process information collected, a trans-receiver in charge of communication between sensor nodes, and an electric power device to supply power. The problem is that power cannot be supplied to these sensor nodes once they are installed [6]. Therefore, they should work with as little electric power as possible in performing the collection and transmission of data as. In such a sensor network, it becomes a great weakness if each node tries to block out DoS attacks with its own filtering system, because a filtering system requires arithmetic computation which consumes a lot of energy.

In this paper, we applied the concepts of Shield [7] and sShield [8], which are traffic deflection techniques for defending against DoS attacks, to the sensor network environment with limited power. This method is a system which can distribute and block out traffic by using stepwise DoS attack detection. The detection method can be any efficient one and is not in the

range of this paper. By applying deflecting techniques to the location-based routing protocol of sensor network, we attempted to make it possible to flexibly cope with DoS attacks. In this method, it is an important factor how many deflection nodes the administrator chooses. The power consumption of the deflection nodes will be higher than other normal nodes and thus it is important to select the minimum possible number of nodes that may achieve an effective defense. Therefore, in this paper, we conducted an experiment to find out the most appropriate number of nodes.

As for the composition of this paper, chapter 2 explains traffic deflection methods and sensor network protocols as a related work and chapter 3 explains the sensor network protocol suggested by this paper. Chapter 4 examines the number and location of designed systems through experiments, and chapter 5 concludes.

2. Related Works

To apply traffic deflection techniques to WSN, we apply the concepts of Shield [7] and sShield [8] that work for the wired networks. Both methods aim to defend against DoS attacks and in the systems the key point are to determine where to install the filtering nodes since the techniques work with any filtering techniques. Besides, since they change normal traffic paths by force through the modification of protocols, they are helpful in dispersing traffic also.

Both methods operate through the transformation of routing protocols. Therefore, it is needed to examine the routing protocols of WSN. Out of them, the location-based routing protocol was selected for this system because of its ease of implementation, deployment, and addition and deletion nodes.

2.1. Traffic Redirection Methods

Unlike existing DoS defense systems, Shield [7] focuses on not how to identify and block malicious traffic but on where to deploy the defense system. Briefly speaking, it makes legitimate traffic arrive at the destination even under an attack, by controlling, monitoring and even blocking out some of the traffic. Thus, Shield is implemented by using traffic trapping and traffic black-holing that are already widely used. As shown in Fig. 1, traffic trapping and traffic black-holing have a concept of having legitimate users' traffic passed through the shield but attackers' traffic blocked out by the shield, while redirecting traffic with shield nodes.

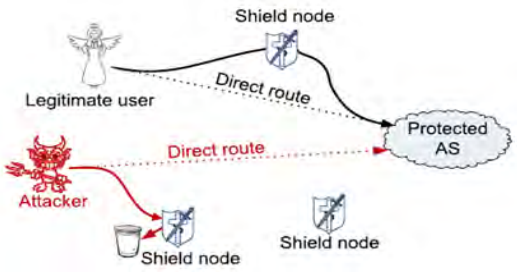


Fig. 1. Diverting traffic flow from a direct route to pass through filtering node (Shield)

There are a few shortcomings of Shield. First, Shield cannot be used inside an AS, because of its use of BGP (Border Gateway Protocol). Therefore, the Shield cannot be deployed inside an AS where small-sized DDoS attacks may occur. Second, under a DDoS attack, the traffic sent by an attacker should be blocked out, but legitimate traffic should arrive at destinations. For this, Shield cannot but provide tunneling or use some kind of source routing. When normal traffic is sent to destinations with such a method, however, there will be an increase in the overhead of a large number of nodes and the network itself.

To supplement these shortcomings, sShield [8] is designed to operate inside an AS by using the concept and definition of Shield. RIP is used as a routing protocol inside an AS. sShield assumes, as it is with Shield, the existence of effective filtering and attack filtering systems and deals with the deployment problem. Since sShield is deployed between two routers and blocks out traffic passing between the two routers, they needed to deploy several sShields and tried to optimize the number and location of sShields. Lastly, to efficiently manage the system, this work systematized sShied to operate in three phases by the riskiness of attacks.

For an efficient management of this system, sShield has three different attack modes, normal routing mode, preventive routing mode and protected routing mode. Each mode is decided by the administrator of AS, depending on attackers' attack phases. Normal routing mode is used for normal operating state where no attack is present and sShield does not do anything and thus no traffic passes through sShield. However, when traffic suspicious of attack is identified, the administrator changes the routing path to sShield, converting normal routing mode to preventive routing mode. In case of an actual attack, the administrator converts preventive routing mode to protected routing mode, in which state traffic caused by attackers is blocked out.

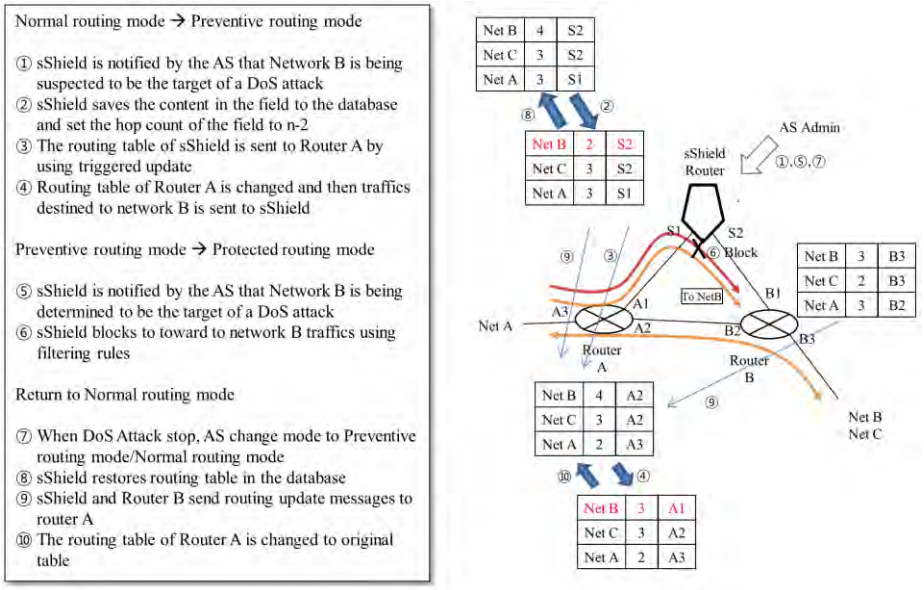


Fig. 2. Changing routing tables of sShield's three routing modes

Fig. 2 shows the process of mode chamges in sShield system. This sShield system also suggested a method to determine the location of attack defense systems deployed, and this structure can be deployed in any position. sShield can be deployed and work inside an AS by using RIP (Routing Information Protocol), which is a typical protocol of IGP (Interior Gateway Protocol) [9]. In other words, it is able to prevent small-scaled DDoS attacks likely to occur inside an AS. Besides, in cooperation with Shield, it is possible to deploy the technique in any position on the internet, and since it uses RIP, traffic unblocked out can reach its destination. More importantly, sShield using the table update of RIP can deflect and block out paths by the unit of networks being attacked, so it is possible to control the flow in a precise way. Lastly, the other existing routers do not have to know the existence of sShield, and routers do not have to be replaced at all, which helps keeping the consistency of a network.

In this paper, we developed a new DoS defense system by applying these traffic deflection ideas to wireless sensor networks.

2.2. Sensor Network Routing Protocols

Routing in WSN is different from ordinary routing protocols used in the existing wired network. Important factors include: nonexistence of infrastructure, unreliability of wireless links, and high probability of errors for many sensor nodes. What matters most is all the routing protocols should save energy to work without being recharged again [5, 6]. To satisfy these

conditions, there have been many different routing protocols made for WSN, and these protocols can be largely classified into six categories as shown in Table 1.

Table 1. Routing protocols for WSN

Category	Representative Protocols
Location-based Protocols	MECN, SMECN, GAF, GEAR, Span, TBF, BVGF, GeRaF
Data-centric Protocols	SPIN, Directed Diffusion, Rumor Routing, COUGAR, ACQURE, EAD, Information-Directed Routing, Gradient-based Routing, Energy-aware Routing, Information-Directed Routing, Quorum-Based Information Dissemination, Home Agent Based Information Dissemination
Mobility-based Protocols	SEAD, TTDD, Joint Mobility and Routing, Data MULES, Dynamic Proxy Tree-Based Data Dissemination
Hierarchical Protocols	LEACH, PEGASIS, HEED, TEEN, APTEEN
Multipath-based Protocols	Sensor-Disjoint Multipath, Braided Multipath, N-to-1 Multipath Discovery
Heterogeneity-based Protocols	IDSQ, CADR, CHR
QoS-based protocols	SAR, SPEED, Energy-aware routing

Such multitude of sensor network protocols has been developed to fit the way and purpose of each different kind of WSN. Data-centric protocols are designed for the purpose of sending data to sink nodes [10, 11, 12]. Hierarchical protocols work by building up sensor nodes in a hierarchical way [5, 6]. Mobility-based protocols are for circumstances when sink nodes move [6]. Multipath-based protocols are to build up several paths from source to sink nodes, not just one single path [6]. Heterogeneity-based protocols are designed for the sensor network environments that combine battery-based sensor networks and power-supplied sensor networks [5, 6]. QoS-based protocols try to minimize the consumption of energy [5, 6]. Lastly, location-based protocols deliver information using the physical locations of nodes as part of routing information, where the physical location of each node is assumed to be known by the nodes [5, 6, 13].

The next explanation is about greedy-based protocol [13], which is the simplest one out of all the location-based protocols. When any sensor network is first deployed, it informs all the nodes of its physical location. All nodes have the physical location table with location information of entire nodes. Besides that, each node should be aware of other nodes within its communication range as well. In this way, when a node attempts to send data to a particular node, it sends data to another node in its communication range, which is located nearest its destination. Fig. 3 shows the process of sending data from node x to node d. Node x sends data to its destination x, through node a. At this point, Node a, having received the data, selects node

b, among b, s, x located within its communication range, to transfer data, which is physically proximate to node the destination node d.

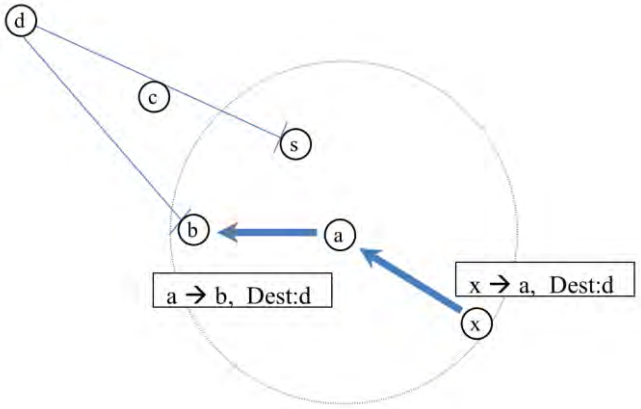


Fig. 3. Location-based routing protocol

2.3. Protection of DoS Attack in WSN

The existing DoS defense methods in WSN were mostly to block the zombie node generation in nodes with authentication by sensor nodes [14, 15, 16], to classify and block services in [17] network traffic, or to select DoS attacks by measuring incoming traffic volume via a sink node [18]. All these methods try to prevent DoS attacks at the source (by authentication or service block) or at the sink and there is no discussion about how to deflect traffic when DoS attacks cannot be prevented that way.

The Shield [7] and sShield [8] are methods to handle exactly that problem in fixed network. However, [7] and [8] work as fixed network routing protocols such as RIP and BGP. In order to apply the idea to WSN, we devised a method for location-based routing protocol, which has the simplest concept among WSN protocols.

3. New DoS Defense System in WSN

In building the suggested protocol, we have made the assumptions listed in the following subsection. Also, we will explain how to distinguish the location where the attacker is located. After explaining all of them, we will explain DoS defense methods in WSN by three phase modes similar to sShield. The three phase modes are normal mode, preventive mode and protected mode.

3.1. Assumptions

- The network of interest is a sensor network structure where free data communication is provided without the concepts of sink and source. However, the system works even in an ordinary sensor network with sink and source nodes.
- The routing protocol for this system is greedy-based routing protocol which is the most basic out of all the location-based protocols.
- Being equipped with GPS, all the nodes know their own locations and each node are also provided with a unique node number. When first deployed, the node numbers are broadcast to the entire nodes.
- The nodes renew information about their neighboring nodes and physical locations on a regular basis. It is to notice any change in neighboring nodes' circumstances. It is possible to have node failures, power exhaustions, and node additions.
- There exists an administrator controlling detection of DoS attacks, selecting nodes, making traffic deflection, and changing modes.
- The nodes in charge of traffic deflection are called ssShield (selected sensors for Shield) nodes in this paper.
- Depending on the riskiness of DoS attack against the sensor network, there exist three modes of operation similar to those of sShield, which are normal, preventive, and protected modes.

3.2. Attacker

The following are the nodes or methods that are likely to lead to a DoS attack in WSN.

- Attacks made through the sink node from an external node.
- Internal sensor nodes deployed inside WSN.
  A new sensor node deployed by an attacker for DoS attack.

3.3. Three Phase Modes

The administrator manages DoS attack situations by dividing them into normal mode, preventive mode, and protected mode. Normal mode indicates a stable state without any attacks. When finding out information suspicious of DoS attack coming through several different paths, the administrator changes normal mode to preventive mode, which redirects traffic to the destination in preparation for a DoS attack. When the deflection traffic turns out to be a DoS attack, the administrator changes the mode to protected mode, which blocks out traffic to the destination.

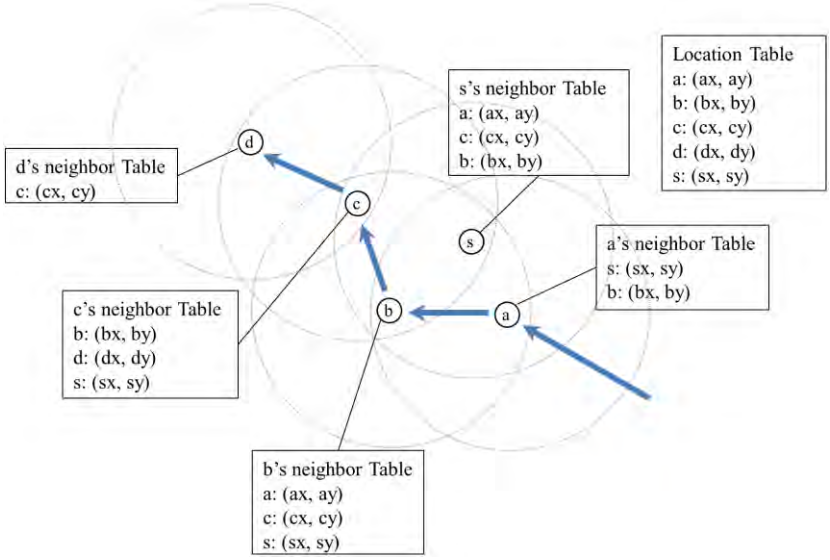


Fig. 4. Sensor network architecture (Normal mode)

Fig. 4 shows a part of the sensor network using the location-based routing protocol. This figure explains how these three modes work. The symbols a, b, c, d, and s indicate each sensor node's number and the grey circle indicates the communication range of each node. Location table is a table with every sensor node's physical coordinates. Moreover, each sensor node has information about its neighboring node to communicate. Fig. 4 shows a state of normal mode and how traffic headed for node d is delivered through other nodes.

At this point, when collecting information suspicious of a DoS attack on node d, the administrator selects some nodes out of all the sensor nodes as ssShield nodes and commands them to change their mode to preventive mode. That is, any node can be an ssShield node, but when they are not selected as ssShield nodes by the administrator, they do not know which nodes are ssShield ones. Fig. 5 shows how node s is selected as ssShield node. The first thing ssShield node does is monitoring all the traffic existing in its communication range. Then, it looks to see if there is traffic heading for node d which is suspected of a DoS attack. If there is traffic heading for node d, it informs that its location has changed in order to bring the traffic to node s. Fig. 6 shows the process of traffic going through node s due to the modified location of s. This Preventive mode works only as distributing suspicious traffic and can be useful for collecting suspicious traffic for further analysis.

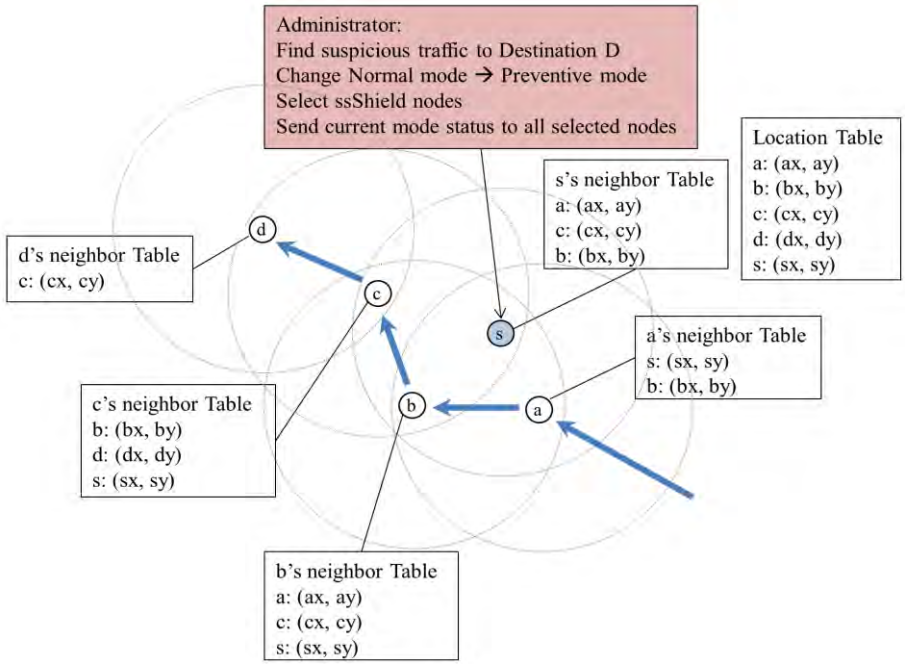


Fig. 5. Change Normal mode to Preventive mode

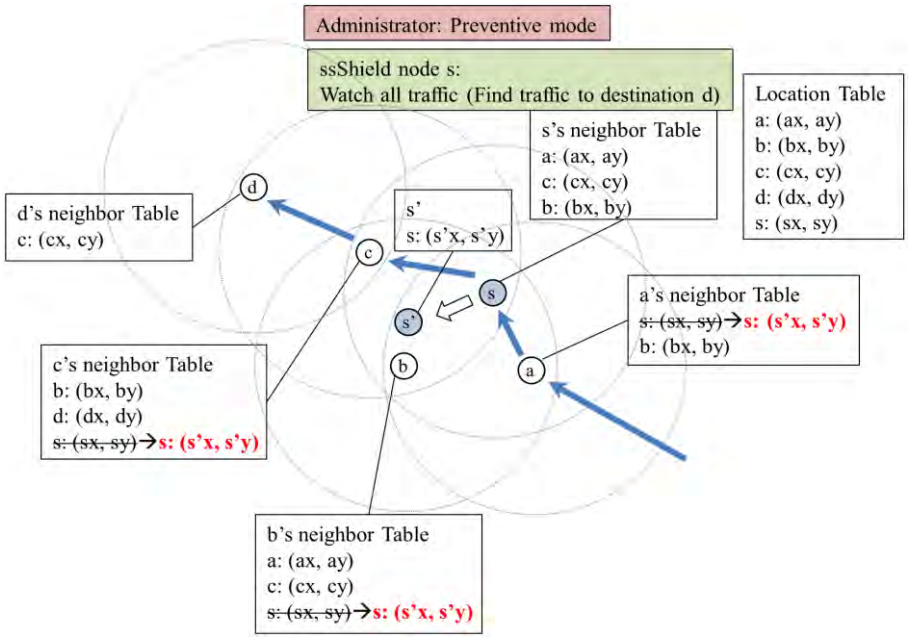


Fig. 6. Traffic redirection by location change

When this traffic is found to be a DoS attack, the administrator commands all the ssShield nodes to block out traffic heading for node d, as shown in Fig. 7. When the DoS attack stops, the administrator makes the ssShield nodes return not to normal mode right away, but to preventive mode first. Then, if there does not occur a DoS attack for a certain period of time, all the ssShield nodes return to normal mode while informing neighboring nodes of their physical locations, as shown in Fig. 8.

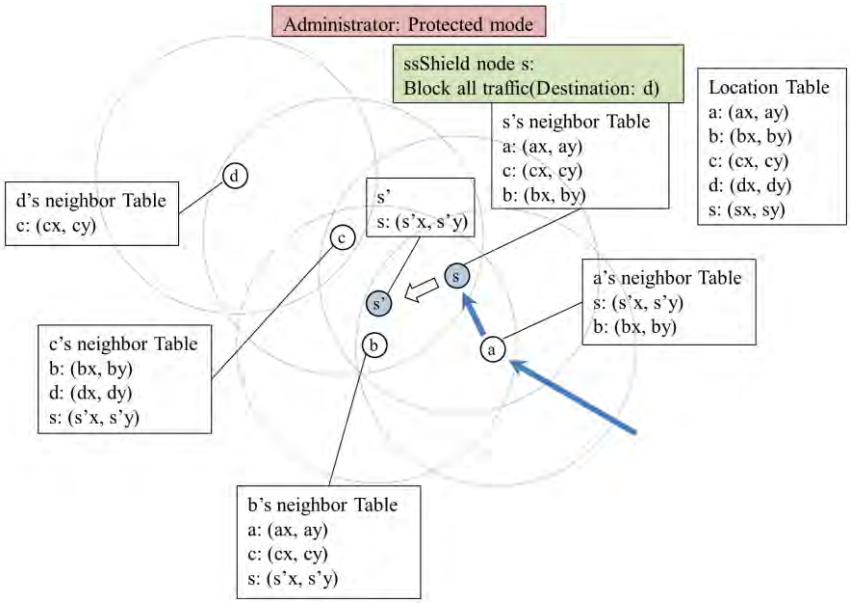


Fig. 7. Protected Mode

3.4. Pros and Cons of this System

ssShield is a system apt to disperse and block out traffic caused by a DoS attack in WSN where the location-based routing protocol is used. While being controlled by the administrator, ssShield can execute the dispersion and block-out of traffic at the same time by deflecting traffic caused by a DoS attack. Some pros and cons of ssShield will be discussed in the following.

As for its advantages, it is possible to efficiently manage energy. First, since a filtering system need not be put on the sensor node, it is possible to prevent waste of energy caused by arithmetic operation. Besides, it is possible to disperse electric power consumption concentrated on particular nodes, caused by a DoS attack. Second, other nodes that are not selected as ssShield nodes do not have to perform such operations as deflection and block-out of traffic. They do not have to be controlled by the administrator,

while keeping the consistency without the need of the existing protocols to be modified.

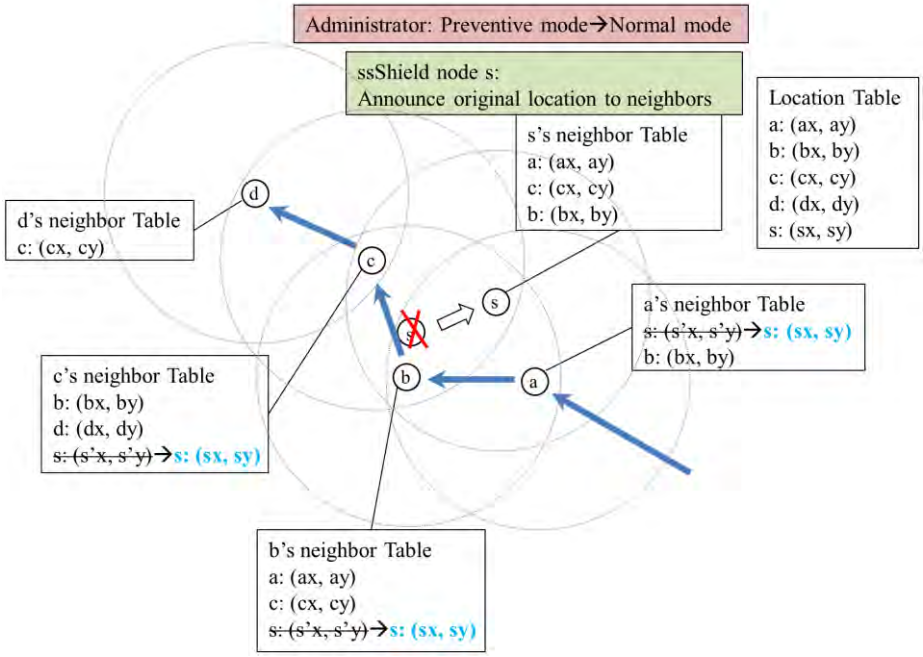


Fig. 8. Return to Normal mode

As for its disadvantages, first, it must redirect and block out both the legitimate and malicious traffic heading for the target nodes of DoS attack. The reason is that there is no way to distinguish the malicious traffic from legitimate traffic unless filtering modules are put on sensor nodes selected as ssShield nodes. Second, sensor nodes selected as ssShield nodes have a higher power consumption rate than other nodes. The energy consumption increases during the process of inspecting all the signals around nodes to make traffic deflection. However, it is still lower than when the WSN is under DoS attack, in terms of the mean energy consumption of sensor nodes. Third, when a traffic redirection should be made in the preventive mode, there may be some cases when traffic cannot be delivered to destinations, depending on the circumstances of sensor nodes. In case when traffic comes to ssShield itself, the ssShield node should cancel path deflection while judging whether its connection is disconnected. Lastly, since modifying physical locations is similar to a malicious activity, it is required to provide a safe key mechanism between the administrator and every sensor node.

4. Experiments

Through an experiment, we attempted to investigate the defense rate of DoS attacks by the number of ssShields. This experiment is to find out how many sensor nodes the administrator should select in order to defend a DoS attack efficiently.

4.1. Experiment Environment

The experimental environment is a simulated one and the topology of sensor nodes was determined randomly. We fixed the number of sensor nodes, the communication range of sensor nodes, and the entire width without any change given. For the experiments, ten times of topology change was made for each round, and twenty times of ssShield selection was made for each topology. At the same time, twenty times of DoS attack was performed while ssShield was arranged one time, through which this system marked the probability that ssShield could defend DoS attacks. For a DoS attack, traffic was arbitrarily caused in randomly selected nodes to attack randomly selected set of nodes. The nodes selected as "ssShield" by Administrator are determined by random selection method. If ssShield nodes are selected by a specific criterion, then there is a high probability of a biased choice which will lead to concentration of energy consumption to specific nodes, which, in turn, will shorten the time the whole network is useful. The possibility remains that some clever mechanism exists that do not have this drawback, but it is left as a future work.

4.2. Result of Simulation

At first, the experiment was performed by increasing by two ssShields nodes to the topology deployed with 100 wireless sensor nodes in WSN environment.

Fig. 9 shows the probability of successfully defending DoS attacks by the ratio of ssShields selected. This experiment used the topology deployed with 100 nodes, and an average of 3.8 nodes existed in the communication range of any single node.

As shown in Fig. 9, when 50% of all wireless sensor nodes were selected, the traffic dispersion and filtering can be provided against DoS attacks most of the time. However, it means that 50% of the entire sensor nodes consume power more than in a normal state. Therefore, it is necessary to select a right number of nodes fit for the sensor administrator's policy and network circumstances.

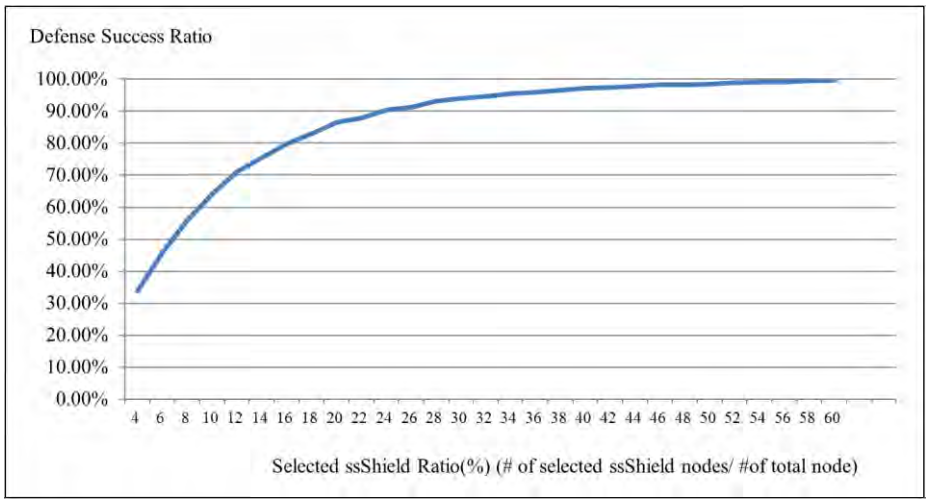


Fig. 9. Probability of defending DoS attack by the ratio of ssShields selected

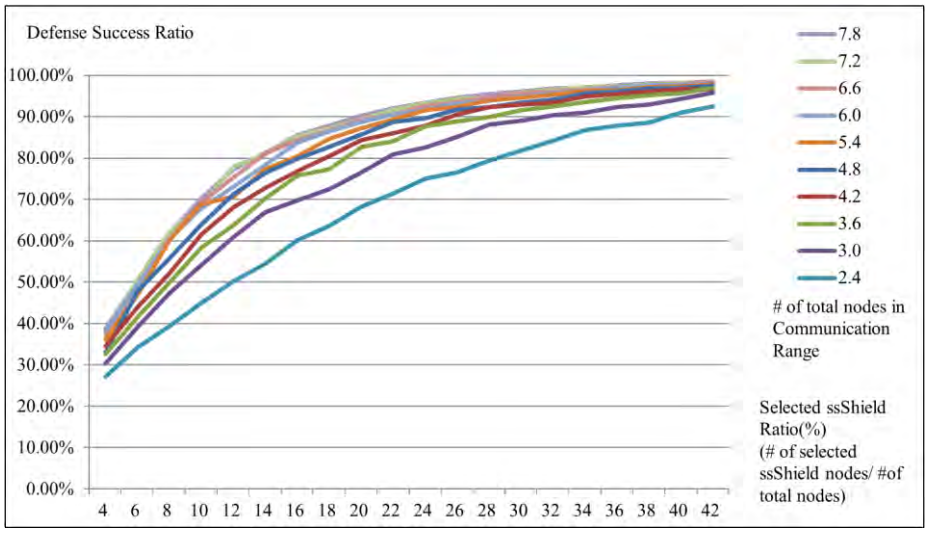


Fig. 10. Probability of defending DoS attacks by an increase of the entire sensor nodes

Fig. 10 is a graph showing how the defense ratio changes against DoS attacks as the number of wireless sensor nodes deployed in the same space increases. As shown in this graph, it is clear that the defense capability depends on the distance between sensor nodes, in other words, how many neighboring nodes there are in one communication range.

5. Conclusion and Future Works

In this paper, we proposed a new method to defend against DoS attacks in WSN using the ideas of Shield and sShield that are DoS attack defense techniques using traffic deflection for the existing wired networks. Out of the wireless sensor network routing algorithms, this paper focused on the location-based routing protocol for the sensor network environment since it is simple and easy to implement and install. The proposed method was systematized with three phases of DoS attacks fit for each risk circumstance, and the administrator was made to select nodes to make traffic deflection when judging a DoS attack. Then, by changing the location parameter of location-based routing protocol only for nodes selected, it is possible to for the selected nodes to make traffic come to themselves, by which a DoS attack could be blocked out. Besides, the number of nodes to deflect path was confirmed by experiments. By using the results, it will be possible to make judgments about the appropriate number of nodes depending on various sensor network conditions. This paper proposed a method to defense DoS attack effectively in WSN (wireless sensor network) environment by using traffic diversion method.

Traffic deflection was applied only to the location-based routing protocol in this paper. Therefore, it is necessary to examine if this method can be applied to other protocols. Moreover, no exact experiment was conducted on the consumption of energy. However, it is expected to measure energy consumed in an entire sensor network by using traffic loaded on the entire nodes in each mode. Sensor nodes were randomly selected for path deflection. This is to prevent energy consumption from concentrating on a specific node, but there may be other methods possible too. Probably, we may consider to use nodes around the sink node, an concentrated spot, or nodes consuming the least amount of energy. All these comprise possible future works.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 20120006492).

References

1. Tao peng, Christopher Lecke, Kotagiri Ramanohanarao.: Survey of Network-Based Defense Mechanisms Countering the DoS and DDoS Problems. ACM Computing Survey(CSUR) vol.39, no.1, Article 3. (2007)
2. K Gar.: Detection of DDoS attack using data mining. International Journal of Computing and Business Research (IJCBR) volume 2 Issue 1. (2011)
3. Trustwave SpiderLabs.: The Web hacking incident database Semiannual report. July to December. (2011)
4. E. Kline, M. Beaumout-Gay, J. Mirkovic, and P. Reiher.: RAD:Reflector attack defense using message authentication codes. In Proceedings of Annual Computer Security Applications Conference(ASAC). pp. 269-278. (2009)

5. Shio Kumar Singh, M P Singh, and D K Singh.: Routing Protocols in Wireless Sensor Networks – A Survey. International Journal of Computer Science & Engineering Survey (IJCSES) Vol.1, No.2. (2010)
6. Kemal Akkaya, Mohamed Younis.: A survey on routing protocols for wirless sensor network. Ad Hoc Networks Vol. 3, Issue 3. (2005)
7. Erik Kline, Alexander Afanasyev, Peter Reiher.: Shield: DoS Filtering Using Traffic Deflecting. 19th IEEE International Conference on Network Protocols (2011).
8. Ho-Seok Kang, Sung-Ryul Kim.: Design and Experiments of small DDoS Defense System using Traffic Deflecting in Autonomous System. Journal of Internet Services and Information Security(JISIS) In proceedings of MIST 2012, Vol.2, no.3,4. pp.43-53 (2012)
9. James F. Kurose, Keith W. Rose, Computer Networking.: A Top-Down Approach Featuring the Internet. Addison Vesley Longman Inc. (2001)
10. C. Intanagonwiwat, R. Govindan, and D. Estrin.: Directed diffusion: A scalable and robust communication paradigm for sensor networks. Proceedings ACM MobiCom'00, Boston, MA, Aug. pp. 56-67. (2000)
11. C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva.: Directed diffusion for wireless sensor networking. IEEE/ACM Transactions on Networking. vol. 11., no. 1. (2003)
12 D. Braginsky and D. Estrin.: Rumor routing algorithm in sensor networks. Proceedings ACM WSNA. in conjunction with ACM MobiCom'02 GA, pp. 22-31. (2002)
13. Brad Karp, H.T. Kung.: GPSR: Greedy Perimeter Stateless Routing for Wireless Network. MobiCom. (2000)
14. Yao Lan, Yu Zhiliang, Zhang Tie, and Gao Fuxiang.: Dynamic window based multihop authentication for WSN. In proceedings of 17th ACM conference on CCS'10. pp. 744-746 (2010)
15. Hao Chen.: Efficient compromising resilient authentication schemes for large scale wireless sensor networks. In proceedingds of 3rd ACM conference on WiSec'10. pp.49-54 (2010)
16. Chakib Bekara, Maryline Laurent-Maknavicius and Kheira Bekara.: H^2BSAP: A hop-by-hop Broadcast Source Authentication Protocol for WSN to mitigate DoS attacks. In proceedings of 11th IEEE Singapore International Conference on ICCS 2008. pp.1197-1203 (2008)
17 Jiang Zhongqiu, Yan Shu, and Wang Liangmin.: Survivability Evaluation of Cluster-Based Wireless Sensor Network under DoS Attacks. In proceedings of 5th International conference on WiCom'09. pp.1-4. (2009)
18 Khusvinder Gill and Shuang-Hua Yang.: A Scheme for Preventing Denial of Service Attacks on Wireless Sensor Networks. In proceedings of 35th IEEE Annual Conference on IECON'09. pp.2603-2609. (2009)

Ho-Seok Kang is a postdoctoral fellowship of the division of Internet and Multimedia Engineering at Konkuk University, Seoul, Korea. He received his Ph.D. degree in computer enginnering at Hongik University, Korea. His recent research interests are in network security, network protocol, mobile security, distributed algorithms and cloud computing.

Sung-Ryul Kim is a professor of the division of Internet and Multimedia Engineering at Konkuk University, Seoul, Korea. He received his Ph.D. degree in computer enginnering at Seoul National University, Korea. His recent research interests are in cryptographic algorithms, distributed algorithms, security in general, cloud computing, and data mining.

Pankoo Kim received the B.S. degree in computer engineering at Chosun University, the M.S. and the Ph.D. degree in computer engineering at Seoul National University in South Korea. He is a professor in the Department of Computer Engineering at Chosun University. His research focuses on Semantic Web Technologies, Ontology, Multimedia, Natural Language Processing, and Data Mining.

Received: September 14, 2012; Accepted: March 12, 2013

DOI: 10.2298/CSIS120922030L

Design and Implementation of E-Discovery as a Service based on Cloud Computing

Taerim Lee[1], Hun Kim[1], Kyung-Hyune Rhee[1], and Sang Uk Shin[1]

[1] Pukyong National University,
Busan, Republic of Korea
{taeri, mybreathing, khrhee, shinsu}@pknu.ac.kr

Abstract. Recently, as IT Compliance becomes more diverse, companies have to take a great amount of effort to comply with it and prepare countermeasures. Especially, E-Discovery is also one of the most notable compliances for IT and law. In order to minimize the time and cost for E-Discovery, many service systems and solutions using the state-of-the-art technology have been competitively developed. Among them, Cloud Computing is one of the most exclusive skills as a computing infrastructure for E-Discovery Service. Unfortunately, these products actually do not cover all kinds of E-Discovery works and have many drawbacks as well as considerable limitations. This paper, therefore, proposes a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS(E-Discovery as a Service) to make the best usage of its advantages and overcome the limitations of the existing E-Discovery solutions. EDaaS enables E-Discovery participants to smoothly collaborate by removing constraints on working places and minimizing the number of direct contact with target systems. What those who want to use the EDaaS need is only a network device for using the Internet. Moreover, EDaaS can help to reduce the waste of time and human resources because no specific software to install on every target system is needed and the relatively exact time of completion can be obtained from it according to the amount of data for the manpower control. As a result of it, EDaaS can solve the litigant's cost problem.

Keywords: E-Discovery, EDRM, Cloud Computing, SaaS.

1. Introduction

Due to the wide distribution of digital devices such as computers, smart phones and rapid advances in various IT technologies, Internet has become a part of our daily life and automated information processing system has been used more and more in our work. As a result of it, electronic documents have been rapidly getting used. This situation has had an impact on the judicial systems and brought big changes on them. In litigation, particularly on civil

litigation in the US Federal Courts, the parties are required, if requested, to produce documents which are potentially relevant to the issues and facts of the matter. This is a part of the process called “Discovery”. When it involves with the electronic documents, or more formally, “Electronically Stored Information (ESI)”, it is called as E-Discovery. Especially nowadays, the growing number of legal cases for civil or criminal trials where critical evidences are stored in digital storages has been submitted as the digital forms of information with a high preference. Moreover, business owners and professional executives are growing more interested in E-Discovery since the number of lawsuits is rapidly increasing among business corporations due to conflicts of interest. And also, many global firms specially aimed at U.S. are reconstructing their business processes and deploying the professional E-Discovery service solution to cope with fast-growing IT compliances effectively apart from ERP (Enterprise Resource Planning) solutions because E-Discovery is also one of the most notable compliances and a specialized field for IT [13]. As IT Compliance becomes more diverse, companies have to take a great amount of effort to comply with it and prepare countermeasures.

The major objective of E-Discovery works is to win a suit. To achieve this goal, the litigants have to secure crucial evidences closely related to litigation issues and apply them to prove their legitimacy. In the E-Discovery procedures, the Potentially Relevant Documents are said to be responsive. The actual E-Discovery works are performed by both jurists and IT experts who are collaborating with each other. When the litigation is filed, an attorney or a legal team hired by the litigant analyzes the contents of the petition and identifies major issues of the litigation at first. Then, they produce a keyword list about evidences which must be secured on the basis of the litigation issues and deliver it to IT experts. By using the generated keywords as well as the specialized tools, IT expert or a special team searches related data as potential evidence and visualizes them for review. After that, attorneys review and analyze again the extracted data from various points of view such as suitability, sensitivity and confidentiality. Finally, all evidences are produced by passing through the procedures mentioned above for a presentation in the trial [1]. Although this procedure sounds easy, it is very complicated works and there are many cases which this procedure is not going well because of several unexpected variables such as system error, data loss, and etc.

When people do an E-Discovery, there are two important factors that have to be obligatorily considered besides winning a suit. One is time and the other is cost. Recently, the volume of ESI that must be reviewed for relevance continues to grow and continues to present a challenge to the parties. So, the cost of E-Discovery can easily be in the millions of dollars. According to some commentators, these costs threaten to skew the justice system can easily exceed the amount at risk. Discovery is a major source of costs in litigation, sometimes accounting for as much as 25% of the total cost. Overwhelmingly, the biggest single cost in E-Discovery is for attorney review time - the time spent considering whether each document is responsive

(relevant) or not. Traditionally, each document or email was reviewed by an attorney. As the volume of ESI continues to grow, it is becoming increasingly untenable to pursue that strategy [7]. In addition, according to FRCP(Federal Rules of Civil Procedure), litigants must submit all evidences within 120 days from the day of lawsuit filed [10]. 120 days seem to be enough time to make evidences but the reality is different. Because that period contains a lot of tasks, such as a checking the litigation issues, a discussion about whole e-Discovery schedule or evidence submission format. If litigants cannot prepare suitable evidence within the fixed period by a law, the case is definitely lost. So, attorneys and their clients are looking for ways to minimize the cost and time of E-Discovery.

To comply with their request, many E-Discovery vendors have competitively developed and released their own service system or software applying the state-of-the-art technologies and Cloud Computing is one of the most exclusive skills as a computing infrastructure for E-Discovery service. Unfortunately, this business is still at a preliminary stage. So, a present level is a simple and partial combination between the existing E-Discovery technologies and Cloud Computing factors for performance enhancement. On the other hand, there are some solutions which implement all E-Discovery functions based on Cloud Computing through a complete platform conversion. However, these products actually do not cover all E-Discovery works and have many drawbacks as well as considerable limitations [3].

In this paper, therefore, we design a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS(E-Discovery as a Service) in order to make the best use of its advantages and overcome the limitations of the existing E-Discovery solutions. The goal of EDaaS is to put all required functions during a whole E-Discovery procedure on the cloud service. This means EDaaS enables E-Discovery participants to smoothly collaborate by removing constraints on working places and minimizing the number of direct contact with target systems. What those who want to use the EDaaS need is only a network device for using the Internet. Moreover, EDaaS can help to reduce the waste of time and human resources because no specific software to install on every target system is needed and the relatively exact time of completion can be obtained from it according to the amount of data for manpower control. As a result of it, EDaaS can solve the litigant's cost problem. Compared to the previous version of this paper appeared in MIST 2012 [2], we improve the EDaaS architecture to expand its functionalities and additionally propose the framework to clarify a configuration of EDaaS. Also, we suggest the way of performance improvement and implement the prototype version of EDaaS.

This paper is organized as follows. Section 2 introduces the background and related work of this study. Section 3 explains how to design and how to use the EDaaS. Section 4 describes three implementation methods for differentiated functions of EDaaS and shows the result of implementation as the prototype. Section 5 then analyzes the practicality of EDaaS to confirm its advantages and limitations. At last, Section 6 presents our conclusion and future work.

2. Background and Related Work

2.1. E-Discovery and EDRM(Electronic Discovery Reference Model)

Electronic discovery (or E-Discovery), first introduced by Federal Rules of Civil Procedure amendments on December 1 2006, refers to Discovery in civil litigation which deals with information in electronic format referred to as ESI (Electronically Stored Information) [10]. This is the result that reflects the modern trend that Discovery's main target is ESI. According to these rules, each company has the responsibility to produce their own evidence for winning the suit, and the use of digital forensic tool is essentially necessary.

EDRM is specified legal requirements of E-Discovery mentioned in U.S. FRCP, and EDRM describes the details about tasks of E-Discovery works. This provides guidelines associated to E-Discovery procedure for standardization and describes functional specification of each phase. This guideline can be recognized as a universal standard because it has been developed in consultation with more than 60 leading E-Discovery-related organizations since 2006. Thus, most of the tools and techniques for E-Discovery are designed on the basis of this model [11]. Fig. 1 shows EDRM diagram which represents a conceptual view of the E-discovery process.

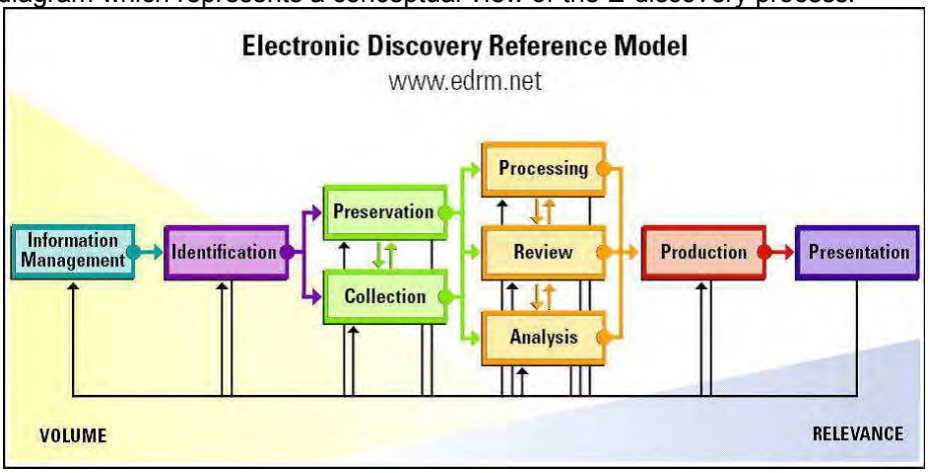


Fig. 1. Electronic Discovery Reference Model

2.2. Major Functions of Existing E-Discovery Service and Solutions

Table 1 shows the phases of e-Discovery and summary from specifications of each phase proposed by EDRM [11].

Most functions of existing E-Discovery service and solutions are focused on the following list of five phases(Collection, Processing, Review, Analysis and Production) because all these phases have a high level of dependence on tool's performance for efficiency improvement of E-Discovery works [3]. The primary technologies for implementing these tools are as follows:

- Document Indexing and Query Processing for an effective search operation
- Classification for removing of duplicated, patent or confidential documents
- Data Format Converting for using of integrated platform, prearranged evidence production and various format compliances
- Data Visualization for a cooperation of review and analysis operation
- Labeling and Tagging for a document selection based on the relevance with litigation issues

Table 1. The phases and the summaries from specifications of each phase proposed by EDRM

Phases	Summary of Specifications
Information Management	Phase to manage their own ESI according to organization's information management policy
Identification	Phase to determine scope of e-Discovery target and identify a real ESI for collecting and preserving
Preservation	Phase to protect ESI from a malicious attack or an intentional destruction
Collection	Phase to collect ESI from various types of storages
Processing	Phase to remove overlapping ESI or unrelated data with lawsuit from collected ESIs and convert the ESI to fit the format for an effective review
Review	Phase to sort sensitive ESI according to privilege, confidentiality, privacy
Analysis	Phase to analyze the collected ESI based on Litigation-related information (Litigation issue, Persons, Keyword, Important documents)
Production	Phase to product ESI with a format negotiated in advance
Presentation	Phase to submit ESI an effective way for being crucial evidence

2.3. The Impact of IT Compliance on the E-Discovery

Generally, GRC (Governance, Risk management, and Compliance) is the umbrella term covering an organization's approach across these three areas. Being closely related concerns, governance, risk and compliance activities

are increasingly being integrated and aligned to some extent in order to avoid conflicts, wasteful overlaps and gaps. While differently interpreted in various organizations, GRC typically encompasses activities such as corporate governance, Enterprise Risk Management (ERM) and corporate compliance with applicable laws and regulations [9]. Among them, Compliance means the conforming to the stated requirements. At an organizational level, it is achieved through management processes which identify the applicable requirements (defined in laws, regulations, contracts, strategies and policies as examples), assess the state of compliance, assess the risks and potential costs of non-compliance against the projected expenses to achieve compliance, and hence prioritize, fund and initiate any corrective actions deemed necessary. Widespread interest in GRC was sparked by the US Sarbanes-Oxley Act and the need for US listed companies to design and implement suitable governance controls for SOX compliance, but the focus of GRC has been shifted towards adding business value through improving operational decision making and strategic planning. It therefore has relevance beyond the SOX world [12]. Especially after the appearance of SOX, many countries and organizations make their own compliance in recent years, such as HIPAA, GLBA, or SB1386. These factors have resulted in the multiple companies demanding on a new type of supporting tool in order to satisfy various requirements of compliance. As a result of that, a large number of E-Discovery technologies related to Digital Forensics have been actively developed and several types of E-Discovery solution have been already released to the market.

2.4. Cloud Computing

Cloud Computing is the most prospective technology for the future of E-Discovery service. A definition of Cloud Computing by NIST(National Institute of Standards and Technology) [5] is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources(e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Cloud Computing includes various types of services such as: Infrastructure as a Service(IaaS), where a customer makes use of a service provider's computing, storage or networking infrastructure; Platform as a Service(PaaS), where a customer leverages the provider's resources to run custom applications; and finally Software as a Service(SaaS), where customers use software that is run on the providers infrastructure.

Cloud computing has the five essential characteristics; rapid elasticity, measured service, on-demand self-service, ubiquitous network access, resource pooling. Cloud Computing structure consists of applications, servers, distributed file systems, distributed databases, caches, and cloud storage, mass data analysis, cluster management, server virtualization, etc. The user connects to the cloud service by using the web browser or the

dedicated client, and uses the provided application. Fig. 2 shows a simple SaaS structure of cloud computing system.

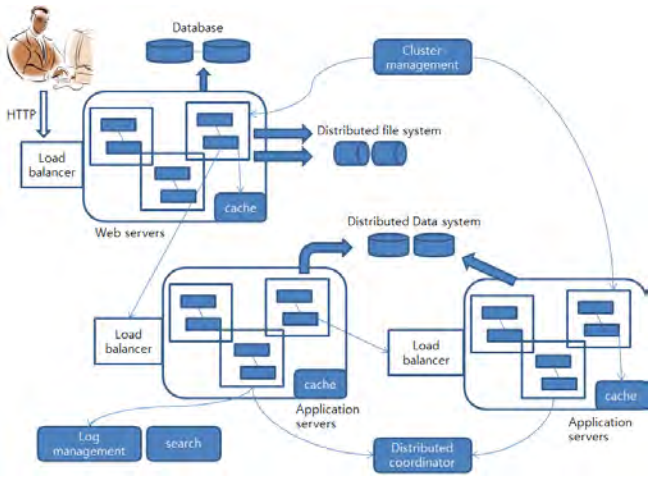


Fig. 2. A Simple SaaS Structure of Cloud Computing System

2.5. E-Discovery Market and Trend of Solution Development

Fig. 3 first introduced in GARTNER 2012 Report shows the famous vendors' position or role in E-Discovery market [3].

Fig. 3. Magic Quadrant for E-Discovery Software

This report was compiled based on the investigation of functionality and characteristics of various E-Discovery software and introduces about each vendor's strengths and cautions. The market covered by this Magic Quadrant contains vendors of e-discovery software solutions for the Identification, Preservation, Collection, Processing, Review, Analysis and Production of ESI in support of the common-law discovery process for litigation, regardless of delivery method. Among them, a vendor who belongs to the group of leaders and visionaries similarly has a clear intention to develop E-Discovery software based on Cloud Computing in a form of SaaS although there are some differences between vendors.

In general, the convergence is made by a partial phased combination and this kind of E-Discovery service consists of two software parts; one is an installation type which was developed at first to deal with many tasks from Collection to Processing phase and the other is cloud server which was implemented Review and Analysis platform. Using the first type software, E-Discovery specialists or hands-on workers can select potentially relevant documents from target system, and convert some documentary format for suitable to the integrated Review and Analysis platform and transfer them to cloud server. After that, various E-Discovery participants, especially company's legal team or attorneys from the external law firm, can review and analyze a relevance of documents as evidence at the same time with no limitations of place. This is an attempt to reduce wasted cost for Review and Analysis phase by improving work efficiency because this phase requires a lot of collaboration among various participants.

AccessData and Guidance Software are representative vendors who make this kind of product. The reason why they are all belong to the leaders group and choose the way of partial convergence is that they already have a powerful software with similar to the first type and they want to keep using and selling that. However in the real litigation cases, cooperation is required through the entire procedure of E-Discovery as well as Review and Analysis. Accordingly, it is necessary to combine additional phases from Identification to Production or to implement all functions on the complete Cloud Computing platform. At this point, vendors such as Xerox Litigation Service, Integreon are continually trying to develop solutions which implement a considerable portion of E-Discovery procedure by using Cloud Computing technologies. Unfortunately, they have not produced a noticeable outcome yet, so they are classified as the Niche Players Group.

Therefore, differentiation factors of our research as follows; the goal of our research is to suggest a new type of E-Discovery service by using Cloud Computing technology. As far as we know, there are no studies related to this goal. Thus, we will compare with famous commercial solution. Considering the trend of E-Discovery solution development, all vendors above mentioned are on the same page, but our attempts and methods to develop a solution are totally different. Simply put, our design and framework is to implement all functions which were required during a whole E-Discovery procedure on the Cloud Computing platform and our methods to conduct them have a distinctive differences from the methods of existing vendors. It means our

development result is the complete convergence of E-Discovery and Cloud Computing beyond the present level of convergence.

3. Design of EDaaS(E-Discovery as a Service)

3.1. Convergence of E-Discovery Solutions and Cloud Computing

In recent years, the quantity of a company's data which may become an object of E-Discovery potentially is growing larger day after day and E-Discovery participants are becoming more diverse. Especially, E-Discovery participants may include company's legal team, general employees, staffs, managers in each department, external law firm, or outsourcing company specialized in E-Discovery, etc. They are people who were closely related with litigation, E-Discovery works or litigant parties. So, nothing is more important than smooth cooperation among participants for the success of E-Discovery works. To reflect this circumstance, the recent trend of technical development for E-Discovery is the convergence of existing services or solutions with Cloud Computing. But even if a lot of famous vendors have been released a new convergence type of solution competitively, serious challenges still remain. Top priority challenge is the complete convergence of E-Discovery with Cloud Computing.

Before attempting to combine E-Discovery Solution with Cloud Computing, most of tools for E-Discovery were developed in a general form called installation type software. It means these kinds of tools must be installed at target system for use. So, E-Discovery participants need extra time for software installation beyond the total time required for E-Discovery works. In order to reduce waste time like this, pre-installing of an E-Discovery tool on every in-house system is time and cost consuming and obviously inefficient. Moreover, installation-oriented software can usually give no guarantee of steady operation pace because its operating efficiency definitely depends on the performance of system where it was installed. With all its faults, vendors don't make an effort to change a principle of their development method because they already have powerful software and they want to sell it consistently. However, it is time for a change. Therefore, we design a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS in order to make the best use of its advantages and overcome the limitations of the existing E-Discovery solutions.

3.2. EDaaS Architecture

The goal of EDaaS is to provide for all functions required during a whole E-Discovery procedure on the cloud service. That is, EDaaS is composed in the

manner of SaaS. To do this, each function will be implemented in the form of application, and each application will interoperate with separated cloud storages based on its purpose and E-Discovery work schedule. Fig. 4 shows the overview of EDaaS architecture.

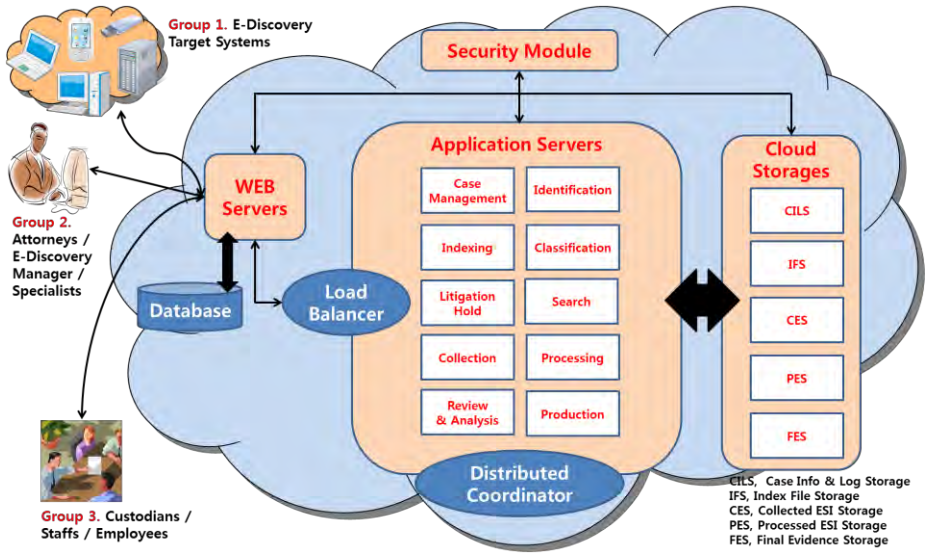


Fig. 4. The Overview of EDaaS Architecture

This architecture is a PIM(Platform-Independent Model). Generally, business applications from various problem domains usually comprise complex functionalities. If such functionalities would not be embedded into the PIM of a software system being designed, a programmer has to create latter a program code of such functionalities, or at least has to amend a generated program code, "by hand" [4].

Users of EDaaS can be divided into three groups. The first group 1 includes E-Discovery target systems which were identified that potentially relevant documents were stored and these systems will be connected for indexing and collection. The second group 2 includes those who have a responsibility to do an E-Discovery works because they were hired as specialists by a litigant such as attorneys in law firm, staffs in outsourcing company specialized in E-Discovery. Of course, if a litigant is a company and the company has a legal or E-Discovery team, these people also belong to the second group. The last group 3 includes those who are related to the litigation issues and have a duty to interview for Identification.

EDaaS consists of 4 parts for the E-Discovery service operation(WEB Servers, Application Servers, Cloud Storages, Security Module) and 2 parts for the system resource management(Load Balancer, Distributed Coordinator). Blocks depicted in Application Servers section of Fig. 4 are

service applications of EDaaS. The name and purpose of each application is shown at the next Table 2.

Table 2. The name and purpose of each application for EDaaS

Name	Target Users	Interoperated Storages	Purpose
Case Management	Group 2	CILS	Saving and managing the all information about case and E-Discovery works (litigation issue, participants, the progress of work, the people concerned, E-Discovery target systems, etc.)
Identification	Group 3	CILS	Providing a specific protocol and reply forms for interview to identify E-Discovery target systems
Indexing	Group 1 and 2	CILS and IFS	Creating index files of each target system for classification and search
Classification	Group 2	IFS	Classifying documents according to contents and updating index files by using the result
Litigation Hold	Group 2	N/A	Ordering target system to prevent users from modifying or deleting important data as potential evidence
Search	Group 2	CILS and IFS	Search for potentially relevant documents related with litigation issue and saving the search result (the path of document)
Collection	Group 1 and 2	CILS and CES	Making a copy of the relevant documents and creating hash values for file integrity
Processing	Group 2	CES and PES	Converting a document file format suitable for integrated Review and Analysis platform
Review and Analysis	Group 2	PES and FES	Providing an integrated platform, visualizing the contents of document, tagging relevant documents as evidence and moving them to FES
Production	Group 2	FES	Convert a document file to the negotiated evidence format and making a final report

In addition to applications, there are essential parts for EDaaS and Fig. 5 shows the entire framework of EDaaS.

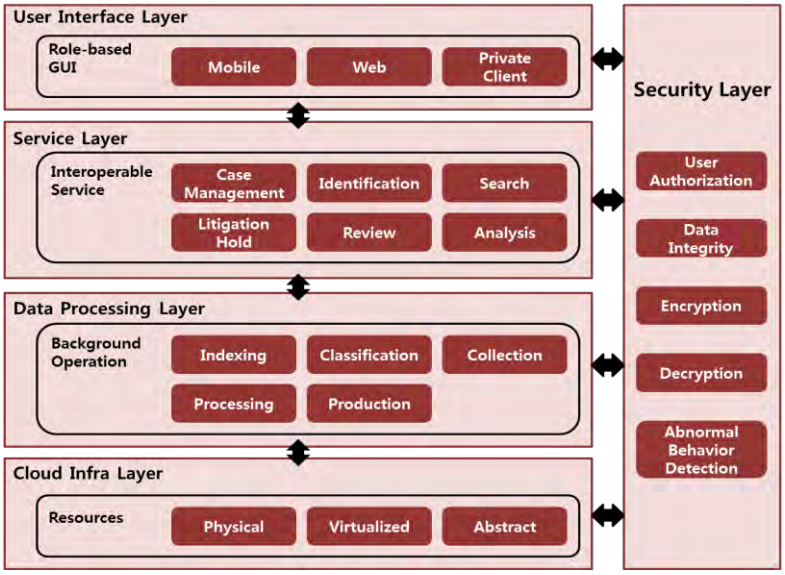


Fig. 5. The Framework of EDaaS

The framework consists of 5 Layers. To the exclusion of Service and Data Processing Layer which was composed of applications described in Table 2, there are 3 more parts; User Interface Layer, Security Layer and Cloud Infra Layer. Each Layer's role is as the following:

- User Interface Layer: Role-based GUI identifies a client's device type, such as Mobile, Desktop or a special device which was made for using a EDaaS only and provides an appropriate GUI for each device.
- Security Layer: It provides a series of functions based on cryptographic technique for user authorization, data integrity, etc. Particularly, this part can be implemented by using a special hardware as well as software. Also, it monitors a state of each layer from user's abnormal behavior.
- Cloud Infra Layer: This layer is for physical hardware of EDaaS. It's a basis part of networks, storages, virtual servers. EDaaS can provide an actual service based on these devices.

3.3. Use Scenario

In order to use the functions of EDaaS, all participants and target systems of E-Discovery have to connect the WEB Servers by using a browser. According to the WEB Server's request, Load Balancer assigns an available Application Server and then WEB Server sends a user's request to the Application Server. After that, Application Server executes a specific application corresponding to the user's request. Fig. 6 shows the mutual

relation between Applications and User Groups with E-Discovery Procedure of EDRM as the center.

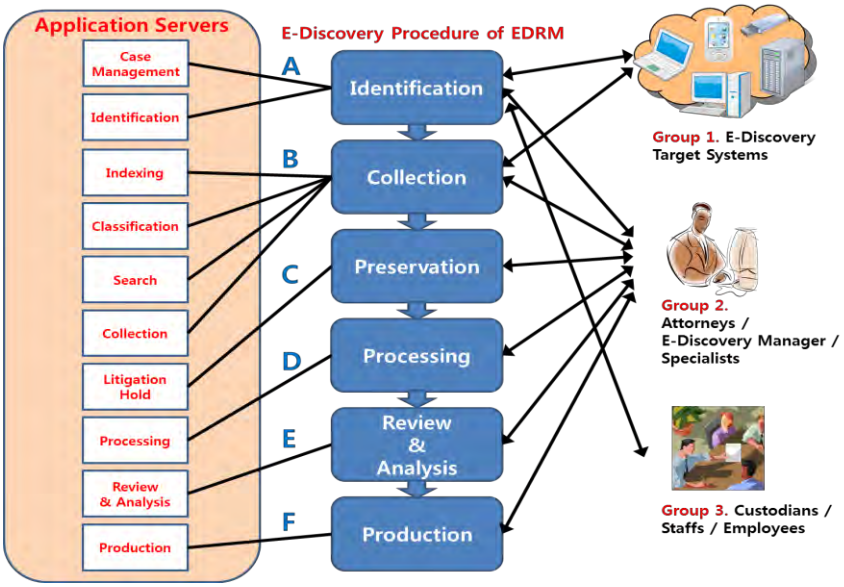


Fig. 6. The mutual relation between Applications and User Groups with E-Discovery Procedure of EDRM

Relation A to F means a bundle of EDaaS Applications to do an essential works for each process of EDRM. These relations reflect the realities of E-Discovery work flow. Full details are as follows:

- A : Once the litigation is occurred, the chief of E-Discovery team creates a database in CLIS and E-Discovery participants record all the information about the litigation and E-Discovery works by using the Case Management application. People those are involved in the litigation have to connect and give an interview personally according to the procedures of Identification. This can make the participants identify E-Discovery target systems.
- B : Identified systems are indexed by the Indexing application. Using an index, participants can search the potentially relevant documents for the future review of suitability as evidence, and the information produced by a Classification application can be used during this process. Because this application enables to remove duplicated documents and identify sensitive documents which are not supposed to make public such as patent or business secret. Classification result can be saved by updating index files with no extra storage. If target documents for review are decided, Collection application can be used to make a copy of each original document and save them to the CES.

- C : By using an Litigation Hold Application, E-Discovery manager have to protect original ESIs from potential threats, such as an intentional Digital Forgery and an accidental loss of data, etc.
- D : Copied files are converted their format suitable for integrated Review and Analysis platform and then they are saved to the PES by using the Processing application.
- E : Attorneys can review and analyze the processed documents and sort out them for the final submission of evidence.
- F : Before the submission, selected documents have to be converted to the negotiated evidence format by using the Production application.

In order to increase work efficiency, various participants can progress this whole process at the same time, regardless of sequence. Also, if the participants know that there are unintended mistakes, errors or failings by the evaluation of each Application's result, they can go back anytime to the troubled part for reworking.

4. Implementation of EDaaS Prototype

Despite the large number of methodologies, standards, and tools, development of large-scale information systems remains a challenging task. The percentage of unsuccessful development projects in terms of exceeding time and/or budget is constantly between 50% and 70%, from the early 80's to the late 90's. Thirty percent of all projects never reach deployment. Prototype-based methods intended to correct these shortcomings and to bring a software project closer to its users [6]. So, we first developed the prototype version of EDaaS which has basic functions with our proposed methods. The development environment for EDaaS is as follows:

- Operating System: Windows 7 Professional K Service Pack 1, IIS 7
- Integrated Development Environment: Microsoft Visual Studio 2010, .NET Framework 4.0
- Database: Microsoft SQL
- Open source library: Apache Lucene 3.1.0 (Indexing/Search), Apache Mahout 0.5 (Classification)

The implementation of Load Balancer and Distributed Coordinator for large-scale service was excluded from the EDaaS architecture because our focus is to develop the basic functions above all. So, we just implemented core parts for 3 components of EDaaS(Web Servers, Application Servers, Cloud Storages) as one in server computer and prepared a local network which was set for Network File System(NFS) to test. Each PC as a target system of EDaaS on this network was assigned static IP address.

4.1. Implementation Methods for Core Functions of EDaaS

In order to differentiate EDaaS from the existing E-Discovery service and solutions, we suggest the following three implementation methods:

- Remote Indexing: The most straightforward method to create index files at the cloud server-side is storing all of original documents in the cloud storage. Considering the amount of company's data is rapidly increasing, this method is very inefficient from the perspective of storage efficiency and making backup every day is also inefficient because people cannot expect when the E-Discovery work will be needed. Remote Indexing is an alternative to solve these problems. At the beginning, Indexing application of EDaaS creates a new user account which is equivalent to the administrator on target system. This function can be implemented in the form of web browser's plug-in. When this plug-in is installed with user's agreement once, it can start to create a new account by modifying the Windows Registry. Using this account, the application makes a reconnection with target system, and start creating index files by using OS dependent functions such as Network File Sharing or File System. Naturally, developers have to prepare additional methods to deal with communication errors for the stability of indexing operation.
- Classification: Making a dictionary of terms which were made up documents and vectorizing is required prior to create index files. The function for the automated document classification based on its contents can be implemented by using the information produced through these kinds of operations. To do this, developers can use the machine learning algorithms as the case may be. If the E-Discovery participants can decide categories of documents and prepare appropriate learning samples in advance, supervised-learning algorithms like Support Vector Machine will be useful. Were it otherwise, unsupervised-learning like K-means will be more useful [8]. In addition, using a distributed processing system like Hadoop [14] enables to reduce the entire operation time.
- Collection: The function for collection can be implemented in a similar way to Remote Indexing. Using an account created for Remote Indexing, all files in target system can be shared over the networks. The work necessary for collection is only copying files what user want. Above this, hash algorithms can be used to verify the originality and integrity of files. To do this, the application has to get hash values of files before making a copy and compare those values after copy operation.

4.2. Website for EDaaS

This website provides various interfaces. To use the EDaaS, all users have to register and log-in at first page. Through this site, administrator of EDaaS can manage all the information of users and create groups to authorize each user based on his or her grade. According to this grade, available functions of

EDaaS are decided and each user can identify these functions at second page after log-in. For example, second web page for Group 1 and 2 includes menus to request applications for Case Management, Indexing, Search, Review/Analysis and the other page for Group 3 to start an interview process for Identification. Fig. 7 shows the status of webpage when administrator logs to the EDaaS for the management of user's information and rights.

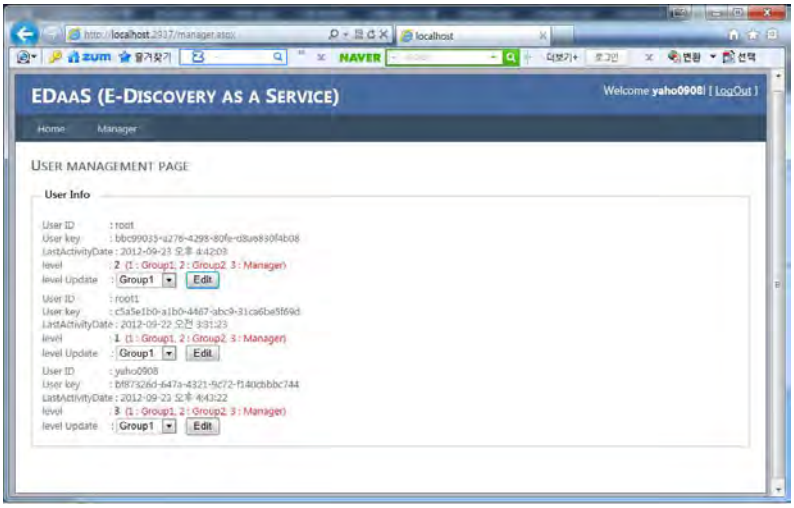


Fig. 7. EDaaS webpage for the management of user's information and rights

4.3. Basic Functions for Application Server of EDaaS

- Indexing and Search : These functions were implemented by using the Apache Lucene Library based on Java. In this prototype, we restricted the document format of E-Discovery target to .TXT text file and applied a simple Boolean search method. Indexing application was made with the C# Thread to run in the background.
- Classification : This function was implemented by using the Apache Mahout Library based on Java and Hadoop Map-Reduce. The biggest reason why we use the Mahout is the interoperability with Lucene. Generally, extra methods for vectorization of each document are required prior to perform a classification. However, Lucene index file is what Mahout only needs for vectorization. Also, it provides various algorithms for document classification, but we choose a K-means clustering method first because it enables to classify documents automatically without training set.[1]

[1] A training set is a set of data used in various areas of information science to discover potentially predictive relationships. Training sets are used in artificial

– Collection : EDaaS prototype runs on network which was set for NFS. So, EDaaS can collect potentially relevant documents by making copy of them.

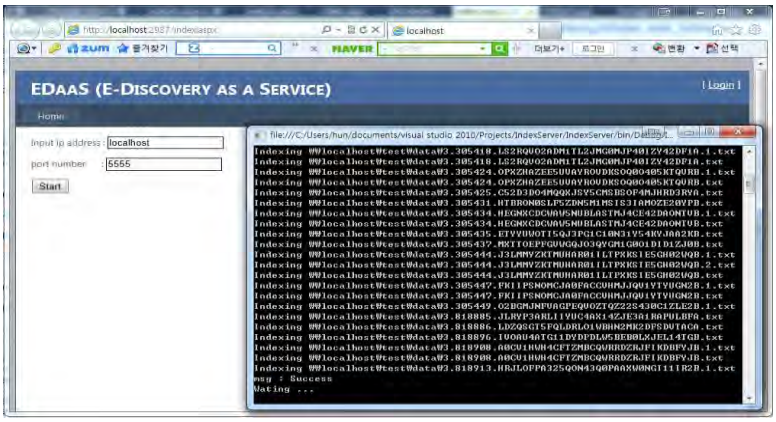


Fig. 8. The Capture of Remote Indexing Operation

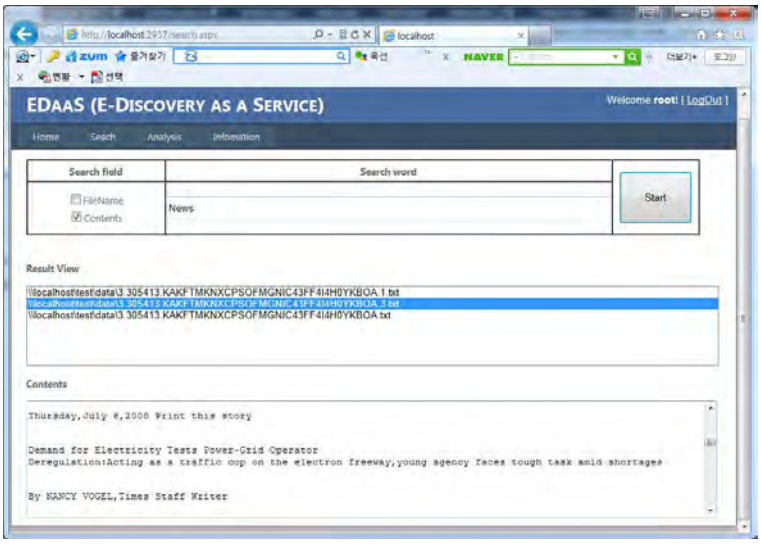


Fig. 9. The Capture of Search and Review Operation

Fig. 8 is the capture of Remote Indexing operation and the console in the right side is to check the logs. Also, Fig. 9 is the capture of Search and Review operation. If the user clicks the one of search result in the middle of page, he can identify the contents of each document.

---

intelligence, machine learning, genetic programming, intelligent systems, and statistics.

5. Analysis about Practicality of EDaaS

Until a recent date, Information Retrieval to find evidence used to be considered the most important function of E-Discovery solution, so evaluation methods for the performance of solution mostly focus on this kind of function. Considering the object of E-Discovery solution, that is quite natural, but it does not fit for informing advantages of EDaaS because it was designed from the another viewpoint. In addition, as far as we know, there are no studies related to this purpose of EDaaS. Therefore, we explain its advantages through the comparison with a typical existing solution.

5.1. Advantages of EDaaS

E-Discovery participants can use EDaaS anytime and anywhere if they have a device for using the Internet. This means no specific software to install on every target system is needed. Especially, the more E-Discovery target systems, the better EDaaS is; it can reduce the waste of time and human resources for the software installation. Moreover, it is difficult to get an estimated time of completion in the case of using the installation type software because its operating efficiency definitely depends on the performance of system where it was installed. If the litigant has to hire persons to the number of target systems for the rapid progression of E-Discovery work, it will cost a huge amount of money. On the other hand, EDaaS can give a relatively exact time of completion according to the amount of data. This information is very useful for the placement of human resources. For this reason, EDaaS can solve the litigant's cost problem. With these advantages, EDaaS enables for participants to collaborate smoothly by removing constraints on working place and minimizing the number of direct contact with target systems. Table 3 shows a comparison with AccessData Summation to explain advantages of EDaaS based on Cloud Computing. Founded in 1987, AccessData Group is a privately held company, with a workforce of over 450, that has addressed the E-Discovery market since 2008 and it has been most famous vendor recently. Also, Summation is the integrated solution which was redesigned to run on the powerful and proven AccessData technology core in 2010 [3].

5.2. Limitations of EDaaS

There are two considerations for practical use of EDaaS. The first is the performance of indexing. The biggest influence is the read/write time for the physical storages on the local system indexing, but remote indexing of EDaaS is additionally influenced by the communication time. So, it is necessary to verify whether or not this tradeoff is tolerable through the experiment. The second is the OS function of Network File System for

Remote Indexing and Collection. Windows OS uses 4 static ports(137, 138, 139, 445) for the sharing service of file and printer. The problem is most ISPs and companies prevent using these ports for security reasons. Furthermore, private local network continues to increase, using the function of NFS as it is with systems on the external network is becoming more difficult. It means additional actions like port forwarding are required to implement Remote Indexing of EDaaS.

Table 3. A comparison with AccessData Summation

Phases	EDaaS	AccessData Summation
Software Installation	N/A	All the target systems of E-Discovery
Extra burden on Installation	N/A	Time, cost, and human resources
Concurrent Users	No limitation	Only one user per system
Working Place	No limitation	Only the place where the system installed it is
Performance	Stable and Predictable (except for network)	Unstable and Unpredictable (It depends on the performance of each system installed it)

5.3. The Future Development Direction for Improving EDaaS

For the performance enhancement of Remote Indexing function, we will bring a Hadoop Map-Reduce technique and implement that function in the form of distributed processing. It is capable of solving the potential Big Data problem. Also in order to prepare when the Remote Indexing is not available because of the network configurations such as the restriction of service port, IP sharing router or VPN, we will develop the additional software in installation type. The ultimate goal of this software is to enable the sharing of file system through the specific port. After expansion of Remote Indexing is complete, experiment for performance evaluation will be done by comparing with local indexing method. Above these works, we will implement the rest of EDaaS architecture and update EDaaS prototype by adding useful techniques for search and review to make it suitable for real E-Discovery business.

6. Conclusion

In this paper, we designed a new type of E-Discovery Service Structure based on Cloud Computing called EDaaS in order to make the best use of cloud computing advantages and overcome the limitations of existing E-

Discovery service or solutions. And then, we explained the structure and framework of EDaaS and suggested a series of a use scenario. Also, we introduced the prototype of EDaaS which was implemented by using three implementation methods; Remote Indexing, Classification and Collection. Finally, we analyzed the practicality of EDaaS and talked about the considerations for the way of improvement.

From now on, complete realization of EDaaS and upgrading its functions based on the study for the improvement of performance will be our future work. It will be performed in accompaniment with the suggestion of a better method for Remote Indexing and the expansion of target OS in order to overcome the limitations of EDaaS prototype.

Acknowledgement. This research was partly supported by Basic Science Research Program(No. 20110006097) and by Next-Generation Information Computing Development Program(No.2011-0029927) through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(Corresponding author: shinsu@pknu.ac.kr).

References

1. Cohen A.I., Kalbaugh G.E.: ESI Handbook : Sources, Technology and Process. Aspen Publishers, Mckinney, USA (2010)
2. Lee T., Kim H., Rhee K.H., Shin S.U.: A Study on Design and Implementation of E-Discovery Service based on Cloud Computing, Journal of Internet Services and Information Security, Vol.2, No.3/4 (MIST 2012 Volume 2), 65-76. (2012)
3. Logan D., Childs S.: Magic quadrant for E-Discovery software. AccessData Company, USA (2012). [Online]. Available: http://accessdata.com/gartner-2012 (current September 2012)
4. Luković I., Popović A., Mostić J., Ristić S.: A Tool for Modeling Form Type Check Constraints and Complex Functionalities of Business Applications. Computer Science and Information Systems, Vol. 7, Issue 2, 359-385. (2010)
5. Mell P., Grance T.: The NIST definition of cloud computing. National Institute Standards and Technology, USA (2011). [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (current September 2012)
6. Milosavljević G., Perišić B.: A Method and a Tool for Rapid Prototyping of Large-Scale Business Information Systems. Computer Science and Information Systems, Vol. 1, Issue 2, 57-82. (2004)
7. Roitblat H.L., Kershaw A., Oot P.: Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review. Journal of the American Society for Information Science and Technology, Vol. 61, Issue 1, 70-80. (2010)
8. Sebastiani F.: Machine Learning in Automated Text Categorization. Journal of ACM Computing Surveys, Vol. 34, No. 1, 1-47. (2002)
9. Tarantino A.: Compliance Handbook(Technology, Finance, Environmental, and International Guidance and Best Practices). Wiley, USA (2007)
10. The Committee on the Judiciary House of Representatives.: Federal Rules of Civil Procedures, USA (2010). [Online]. Available:

http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/2010%20Rules/Civil%20Procedure.pdf (current September 2012)
11. The Electronic Discovery Reference Model.: EDRM Framework Guides, USA (2009) [Online]. Available: http://www.edrm.net/resources/guides/edrm-framework-guides (current September 2012)
12. The Free Encyclopedia Wikipedia.: Governance, Risk Management, and Compliance (2011). [Online]. Available: http://en.wikipedia.org/wiki/Governance,_risk_management,_and_compliance#Integrated_governance.2C_risk_and_compliance (current September 2012)
13. Volonino L., Redpath I.J.: e-Discovery For Dummies. Wiley, USA (2009)
14. White T.: Hadoop: The Definitive Guide 1st Edition. O'Reilly, USA (2009)

Taerim Lee received his Bachelor and Master of Engineering degrees from Pukyong National University, Busan Korea in 2008 and 2010, respectively. He is currently doing a Ph.D. program in Department of Information Security, Graduate School, Pukyong National University. His research interests include digital forensics, e-Discovery, cloud computing, and machine learning.

Hun Kim received his B.S. degree in Major of Computer and Multimedia Engineering from Pukyong National University, Busan, Korea in 2012. He is currently pursuing his master's degree in Department of Information Security, Graduate School, Pukyong National University. His research interests include digital forensics, e-Discovery, Cloud System and security.

Kyung-Hyune Rhee received his M.S. and Ph.D. degrees from Korea Advanced Institute of Science and Technology (KAIST), Daejon, Korea in 1985 and 1992, respectively. He worked as a senior researcher in Electronic and Telecommunications Research Institute (ETRI), Daejon, Korea from 1985 to 1993. He also worked as a visiting scholar in the University of Adelaide in Australia, the University of Tokyo in Japan, the University of California at Irvine in USA, and Kyushu University in Japan. He has served as a Chairman of Division of Information and Communication Technology, Colombo Plan Staff College for Technician Education in Manila, the Philippines. He is currently a professor in the Department of IT Convergence and Application Engineering, Pukyong National University, Busan, Korea. His research interests center on multimedia security and analysis, key management protocols and mobile ad-hoc and VANET communication security.

Sang Uk Shin received his M.S. and Ph.D. degrees from Pukyong National University, Busan, Korea in 1997 and 2000, respectively. He worked as a senior researcher in Electronics and Telecommunications Research Institute, Daejeon Korea from 2000 to 2003. He is currently an associate professor in Department of IT Convergence and Application Engineering, Pukyong National University. His research interests include digital forensics, e-Discovery, cryptographic protocol, mobile/wireless network security and multimedia content security.

Received: September 22, 2012; Accepted: March 08, 2013

DOI: 10.2298/CSIS120714031H

A Topographic-Awareness and Situational-Perception Based Mobility Model with Artificial Bee Colony Algorithm for Tactical MANET

Jinhai Huo[1], Bowen Deng[1], Shuhang Wu[1], Jian Yuan[1], Ilsun You[*, 2]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
[2]School of Information Science, Korean Bible University, Korea
horkin0056@126.com
{dbw10, wsh05}@mails.tsinghua.edu.cn
jyuan@tsinghua.edu.cn
isyou@bible.ac.kr

Abstract. A topographic-awareness and situational-perception based mobility model with path optimization for tactical MANET is proposed in this paper. Firstly, a formalized process is constructed to generate a random acceleration on nodes as the disturbance caused by small-scale topographic factors in the battlefield. Secondly, a path optimization method with the artificial bee colony algorithm is introduced to mimic the trace planning when the nodes possess the terrain information of battlefield. Thirdly, a topographic-awareness based bypass strategy is proposed to simulate the action of nodes facing large-scale terrain factors in the case when the terrain information is lacking. Finally, a situational-perception based avoidance strategy is built to simulate the process of cognition and decision when there is an encounter with the enemies on the march. The mobility model consists of the four parts above and imitates the dynamic characteristics of tactical nodes in military environment.

Keywords: mobility model, tactical MANET, the ABC algorithm, bypass strategy

1. Introduction

A Mobile Ad-hoc Network (MANET) [1] is a collection of mobile facilities and nodes which can build a communication network in the absence of infrastructure. Due to this characteristic, the MANET is widely utilized in military environment and is considered as the foundation of Network-Centric Warfare (NCW) [2]. Mobility models play an essential role in MANET simulation since the trajectories of nodes have an influence on the routing protocol performance[3, 4, 5] and topology algorithm design. The mobility features of nodes separate from each other in various scenarios. Thus, it is necessary to build a model facing military demands which can reproduce the dynamic characteristics in the battlefield environment.

[*]Corresponding Author

Most of the existing models are constructed for social network[6,7] or urban cellular mobile network[8], but military application has its own features. The terrain factors such as torturous roads in small-scale and mountains in large-scale will force the nodes to adjust the path in the battlefield. The perception of enemies will also influence the trajectories of tactical nodes.

In the last paper [9], we had considered the impacts of terrain factors on the dynamic characteristics of mobile nodes in tactical MANET. We established a formalized process to imitate the disturbance of small-scale terrain factors and proposed a terrain-awareness based bypass strategy to imitate the nodes perceiving and bypassing the large-scale terrain factors. However, we did not notice the troops might often acquire the terrain information of the battlefield before the military mission. Meanwhile, the influence that enemy situation might have on the mobility feature was not considered. These omissions had required an improvement in our work.

With the consideration above, a topographic-awareness and situational-perception based mobility model is designed with path optimization for tactical MANET in this work. As is shown in Fig.1, we separate the elements that influence the mobility of nodes into two parts: the static element of terrain factors and the dynamic element of enemy situations. As to the static element, there are small-scale and large-scale topographic factors. Furthermore, when the nodes bypass the large-scale topographic factors, they may have acquired the terrain knowledge or not. Thus, we firstly generate a colored noise with first-order Markov property as the random disturbance on nodes caused by small-scale topographic factors. Secondly, we utilize the artificial bee colony (ABC) algorithm[10] to simulate the path optimization when bypassing large-scale topographic factors if the nodes possess terrain information of the battlefield. Thirdly, if the information is lacking, we propose a topographic-awareness based bypass strategy to mimic the judgment process of nodes about the path selection. Fourthly, we build a situational-perception based avoidance strategy to simulate the process of cognition and decision when facing the dynamic element of enemy situations.

The rest of this paper is organized as follows. A survey of related works in mobility modeling is reviewed in Section 2. The description of the Markov random disturbance on the acceleration is given in Section 3. The procedure of path optimization based on terrain information is proposed in Section 4. The topographic-awareness based bypassing strategy is described in Section 5. The principle of situational-perception based avoidance strategy is given in Section 6. The simulation is shown in Section 7. The conclusion and future works are given in Section 8.

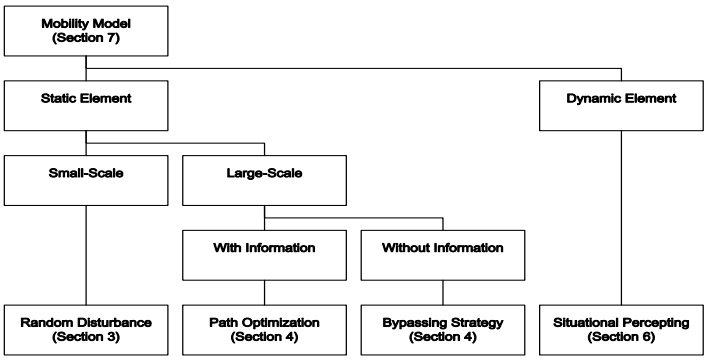


Fig. 1. The structure of mobility model in this work

2. Related Works

Many mobility models have been proposed to simulate the movement of nodes in MANET. Most classical mobility models are based on randomly walking of nodes while some are based on the collection and abstraction of entity movement trajectories in real world. There are also some models aimed at describing mobility feature in specific scenarios. According to the different principles above when designing the mobility models, we can classify them into three categories: synthetic mobility models, realistic trace mobility models and specific scenario mobility models, as is shown in Fig. 2.

2.1. Synthetic Mobility Model

The synthetic mobility models attempt to present the movement of entities based on random walking. In these models, nodes can move freely in the simulation area at a randomly chosen speed. The trajectories of movements are straight lines towards each destination ignoring the influence of the environment. These mobility models are the most commonly utilized and can be classified into entity mobility models and group mobility models based on whether clusters of nodes moving together[11]. The authors of [12] have given a detailed description and comparison of five classical synthetic mobility models: Random Walk Mobility Model, Random Waypoint Mobility Model, Random Direction Mobility Model, City Section Mobility Model, and Reference Point Group Mobility Model. In recent years, researchers also proposed some improvements on these models. For instance, the authors of [13] focus on the boundary of a group. They prefer an elliptic scope with the reference point at the center to imitate the group mobility in a battlefield. The authors of [14] centralized in the direction of node movements and specified them with given angles instead of random directions. The authors of [15] proposed a trust-based framework to support evaluation of information in a VANET.

The synthetic mobility models are simple and ubiquitous, but the random movement brings the problems that the nodes fail to capture the normal human mobility characteristics[16]. Furthermore, the nodes can move anywhere in the area along a straight line instead of being restricted by the environments.

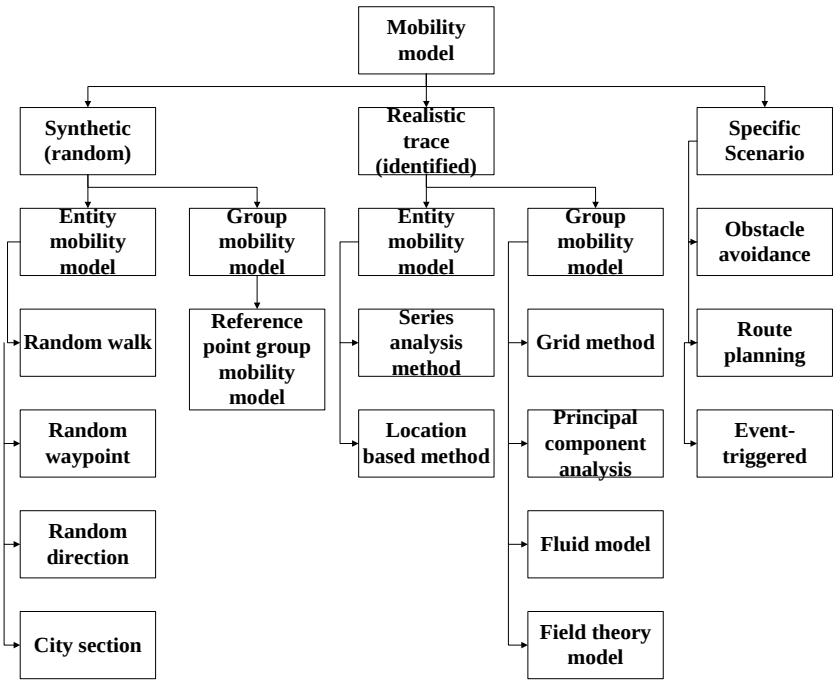


Fig. 2.The classification of mobility model

2.2. Realistic Trace Mobility Model

In the consideration of movement in real world, some researchers attempt to build a mobility model based on realistic trajectories of nodes. They collect the accurate mobility traces of UAVs, vehicles, and humans to abstract the information such as velocity and angle. With the data they build mobility models that reproduce the scene of experiment. The mobility models in [17] and [18] are good examples. The authors in [17] use the trace data collected from a military experiment carried out in Lakehurst, N.J., U.S.A. The distribution of absolute relative direction angle during the experiment is depicted to build the trace model. In [18], more characteristics are collected such as distribution of distance between the leader and followers, distribution of speed for movement duration and probability on pause and move.

These mobility models are extraction of realistic traces, thus truly describe the feature of movement in real world. However, these scene reconstruction

based models are strictly restricted in single scenarios. The dependence on experiment data has confined the application since the experiments in the battlefield are difficult to achieve [19].

2.3. Specific Scenario Mobility Model

Specific Scenario mobility models are presented for special applications with temporal and spatial dependency. The movement of nodes is correlated in time and commonly has correlations with topographic factors[20] or clear destinations[21]. The action of nodes may be triggered by some events and the route is probably planned to avoid some obstacles[22]. For example, entities in Obstacle Avoidance Mobility Model (OAM) [23] need to avoid some obstacles on their way to the target. The Smooth-turn Mobility Model[24] for Airborne Networks captures the correlation of acceleration for airborne vehicles as they cannot make sharp turns as easily as ground vehicles do.

It is a good attempt to take the geographic restrictions into account. However, current route planning algorithms are based on the awareness of limited path selections which are not suitable for real world with intricate terrain factors.

Currently, barely any models could achieve path optimization based on the cognition of topographic factors. Furthermore, hardly any mobility models have considered the dynamic feature when the entities encounter enemies. Therefore, a topographic-awareness and situational-perception based mobility model with path optimization is required.

3. The Formalized Process of Markov Random Disturbance

The random disturbance on the mobile entity is caused by the present small-scale terrain factors like muddy paths and tortuous roads instead of the former ones. This can be modeled as a Markov process. Thus in the work [9], we established a formalized process as shown in Fig.3 to generate the random disturbance on the acceleration of nodes. And the noise is superimposed on the entity as a disturbance of movement.

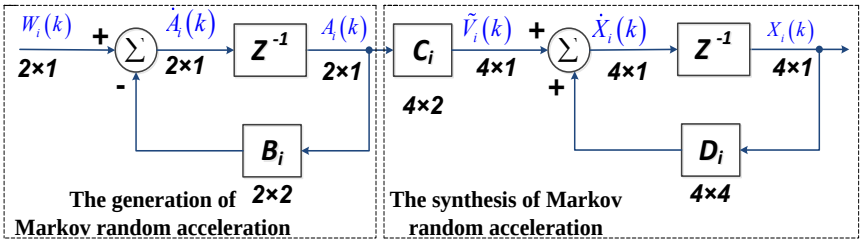


Fig. 3. The state space model of Markov random mobility

3.1. The Generation of Colored Noise

As is shown in Fig.4, the node N_i departs from (x_{i0}, y_{i0}) and marches along the direction α_{Ci} at the constant velocity $\overline{v}_{Ci}$. The node is disturbed by the uncertainty of terrain factors $\vec{a}_R$ on direction θ_i. $x_i(t)$, $y_i(t)$ are the coordinate of N_i. $\dot{x}_i(t)$, $\dot{y}_i(t)$ are the constant velocity of N_i whist $v_{x_i}(t)$ and $v_{y_i}(t)$ are the speed caused by random disturbance $a_{x_i}(t)$ and $a_{y_i}(t)$. ΔT_i is the very short acceleration time on N_i.

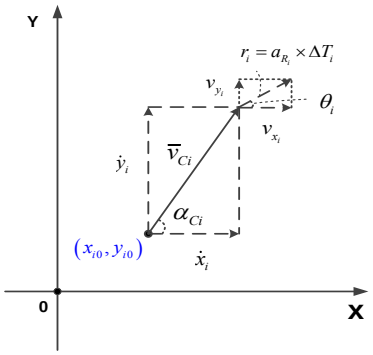


Fig. 4. The speed decomposition of nodes

The Markov random acceleration is generated by the first part of the discrete time system in Fig.3. The state equation is:

$$A_i(k+1) = -B_i A_i(k) + W_i(k) \tag{1}$$

Where $A_i = \left[a_{x_i} \quad a_{y_i} \right]'$ is the random acceleration vector, and

$B_i = \begin{bmatrix} b_{i1} & 0 \\ 0 & b_{i2} \end{bmatrix}$, $\|B_i\| < 1$ effects the correlation of output A_i.

$W_i(k) = \begin{bmatrix} a_{R_i}(k)\cos\theta_i(k) \\ a_{R_i}(k)\sin\theta_i(k) \end{bmatrix}$ is a Gaussian white noise signal whose margin submits the Rayleigh distribution and the phase submits the uniform distribution. A colored noise with first-order Markov property is the output.

3.2. The Synthesis of Random Acceleration

In the process, the constant velocity of N_i remains the same whist the random disturbance on velocity changes over time. The trace of nodes is determined by the combination of the two velocities shown in the second part of Fig. 3. The state equation of the recursive process is:

$$\begin{cases} X_i(k+1) = D_i X_i(k) + \tilde{V}_i(k) \\ \tilde{V}_i(k) = C_i A_i(k) \end{cases} \quad (2)$$

$X_i = [x_i \quad \dot{x}_i \quad y_i \quad \dot{y}_i]'$ is the vector of position coordinates and velocities.

$D_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ T_R & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & T_R & 1 \end{bmatrix}$, $C_i = \begin{bmatrix} T_R\Delta T_i & 0 \\ 0 & 0 \\ 0 & T_R\Delta T \\ 0 & 0 \end{bmatrix}$ T_R is the interval of generating random acceleration and ΔT_i is the interval between different disturbance.

Then the expressions of the coordinates and the velocities are as follows:

$$\begin{cases} x_i(k+1) = x_i(k) + T_R\dot{x}_i(k) + T_R\Delta T_i a_{x_i}(k) \\ y_i(k+1) = y_i(k) + T_R\dot{y}_i(k) + T_R\Delta T_i a_{y_i}(k) \end{cases} \quad (3)$$

$$\begin{cases} \dot{x}_i(k+1) = \dot{x}_i(k) \\ \dot{y}_i(k+1) = \dot{y}_i(k) \end{cases} \quad (4)$$

Eq.3 gives the recurrence relation between the coordinates and the random accelerations of N_i. And Eq.4 shows that the constant velocity remains unchanged during the moving process.

4. The Terrain Information Based Path Optimization

In this section, we propose a terrain information based path optimization method with artificial bee colony (ABC) algorithm to imitate the behavior when the nodes avoiding large-scale terrain factors. In the battlefield environment,

the path is often the optimization considering a variety of factors. We abstract the tactical node path planning as a constrained optimization of a multi-dimensional function[25]. The path selection is abstracted as a solution while the terrain factors construct the parameters of the utility function.

4.1. The Theory of Artificial Bee Colony (ABC) Algorithm

The Artificial Bee Colony (ABC) algorithm proposed by Karaboga in 2005 is a branch among the attempts such as ant colony optimization, particle swarm optimization[26] and bird flocking which employ insect behavior to solve optimization problems[27,28,29]. The ABC algorithm is designed based on two fundamental concepts: self-organization and division of labor. The main feature of the ABC algorithm is the collective global optimization can be achieved through individual partial optimization[30], so the convergence rate is faster.

The honey observation process of the ABC algorithm consists of three components[31]: food sources, employed foragers and unemployed foragers. The algorithm defines two leading modes of behavior: the recruitment to a nectar source and the abandonment of a source.

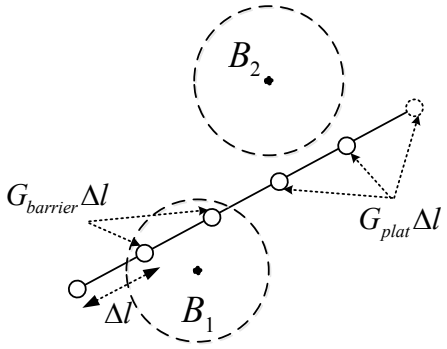


Fig. 5. The process of path optimization

1) Food sources: every food source has a value depends on the parameters of the function.

2) Employed forgers: the bees associated with a particular food source are regarded as “being employed”. They carry the information about this particular source and share them with unemployed foragers.

3) Unemployed foragers: They continually look for a food source to exploit and search the environment surrounding the food source shared by employed foragers to optimize the solution.

4.2. The Process of Path Optimization

As is shown in Fig.5, the path selecting problem is turned into a constrained optimization[32]. Set the line from departure to destination as x-axis and the former coordinate is converted by the equation below:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} + \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \theta = \arg\tan \frac{y_t - y_0}{x_t - x_0} \quad (5)$$

$S(x_0, y_0)$ is the departure point. We divided the path into N_t pieces and calculate the synthesis of every piece. Define

$$J = \sum_{j=1}^{Nt} \sum_{k=1}^{M} \omega_k(x_j, y_j) \sqrt{(y_j - y_{j-1})^2 + (x_j - x_{j-1})^2} \quad (6)$$

J is the utility function with the concept "Road Loss". $\omega_k(x_j, y_j)$ represents the parameter that influence node mobility and M is the number of parameters taken into account. In this paper, the only ω represents the obstacle. And the value of ω represents the impact level on node mobility. In this paper, we assign $\omega = 1000$ in obstacle area and $\omega = 10$ in plain area. The integration of $\sum_{j=1}^{Nt} \sqrt{(y_{k-1,j} - y_{k-1,j-1})^2 + (x_{k-1,j} - x_{k-1,j-1})^2}$ represents the length of path. Thus the entity can minimize the Road Loss by avoiding the barriers with high value of ω while reducing the length of the path.

The procedure is shown in flow chart Fig.6:

1) Initialize the parameters. Divided the x-axis into N_t pieces and every probable solution is a trial of ordinates.

2) Generate some initial path randomly as the food source and calculate the function J of them.

3) Unemployed foragers search for a better path near the food source shared by the employed foragers. If any, replace the former path and record.

4) If there is no better path near the food source, the employed foragers will leave for a new food source.

5) If the number of iterations reaches the limit, then convert the coordinate back and terminate the calculation. Otherwise go to step 2).

6) End.

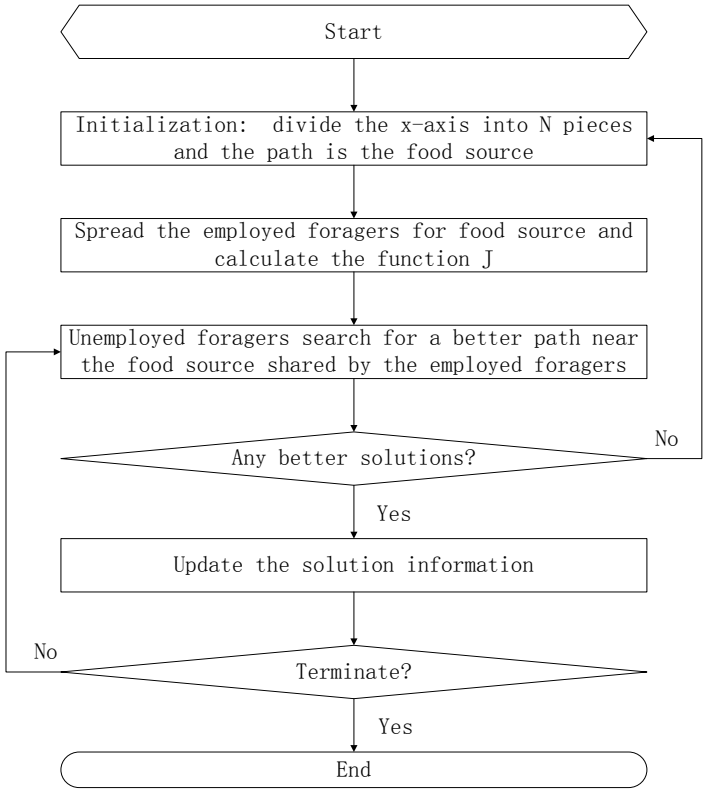


Fig. 6. The flow chart of path optimization with artificial bee colony algorithm

5. The Topographic-Awareness Based Bypass Strategy

Sometimes the entities cannot acquire the terrain information when planning the path. In this situation, we propose a topographic-awareness based bypass strategy to imitate the behavior when the node faces obstacles. The entity can detect the terrain factors within a certain range of distance. Once a barrier is observed on the way to the destination, the entity takes the strategy to bypass the obstacle. The whole procedure can be divided into two parts: the judgment of obstacles and the strategy of avoidance.

5.1. The Judgment of Obstacles

There may be several barriers observed by the entity in the same time. Thus the entity has to judge which barrier will constitute the obstruction along the path.

Take Fig.7 for example. There are three barriers in Fig.7: B_1, B_2, B_3 and the entity CH_i is marching to the target T. In the triangle constructed by CH_i, T, and B_n, define $d_{(CH_iB_n)}$ as the distance between CH_i and B_n, if

$$d^2_{(CH_iB_n)} + d^2_{(CH_iT)} \geq d^2_{(B_nT)}, r_n \leq R_n \quad (7)$$

Then the nearest B_n is the obstacle that will be bypassed in the next step.

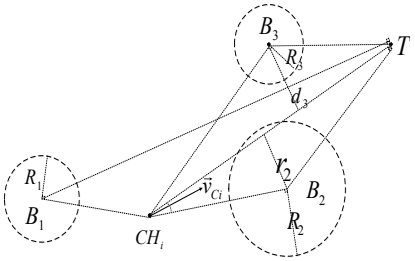


Fig. 7. The judgment of obstacles

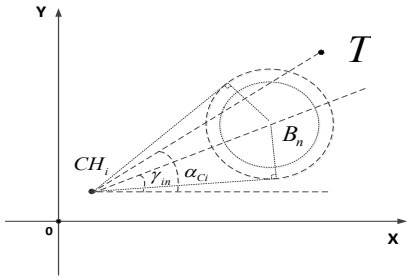


Fig. 8. The selection of bypass direction

5.2. The Bypass Strategy

There are two parameters in the strategy: the bypass direction and the bypass acceleration. As is shown in Fig.8, the bypass direction is judged by the relative location of the entity, the destination and the barrier. If $\alpha_{Ci} \geq \gamma_{in}$, the node will take the clockwise route. Otherwise, the counterclockwise route will be chosen.

In order to minimize the path, the synthesis velocity $\vec{v}_{Ri}$ should be along the boundary of the barrier B_n. If the acceleration on the entity is large enough, the situation is shown in the first figure in Fig.9. $\vec{v}_{Ci}$ is the constant velocity of entity CH_i before observing the barrier. The direction of acceleration $\vec{a}_{Ti}$ is chosen randomly from the possible range Φ_{Ci} .And ΔT is the interval between detections.

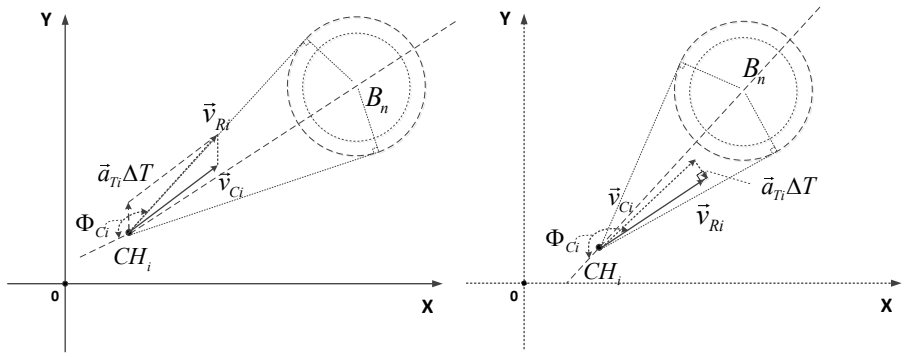


Fig. 9. The selection of bypass acceleration

As is shown in the second figure in Fig.9, once the mobility cannot turn in the necessary direction, the direction of acceleration is perpendicular to the $\vec{v}_{Ci}$ with the maximum value to avoid running into the barrier area.

6. The Situational-Perception Based Avoidance Strategy

The tactical entities may encounter enemies in the battlefield. The process of enemy avoidance brings uncertainty to the dynamic feature of entity mobility. This uncertainty will influence the performance of evaluating MANET. Therefore, it is necessary to take the situational-perception based avoidance strategy into account when designing the mobility model.

We regard the enemies on patrol as a trail of barriers. The difference from Section 5 is that the entities can acquire the velocity of enemies as well as the location. This brings two changes in the avoidance strategy: the judgment of enemy situation and the direction of avoidance.

6.1. The Judgment of Enemy Situation

As is shown in Fig.10, we can estimate the future position of enemies by assuming the velocity of enemies remain the same between detection interval

ΔT . The zone between current position and future position is also a dangerous area that could not be passed through.

In the case $\angle B_{11}CH_iT$ and $\angle B_{12}CH_iT$ locate at both sides of the line CH_iT , if

$$d^2_{(B_{11}CH_i)} + d^2_{(TCH_i)} \geq d^2_{(B_{11}T)} \text{ or } d^2_{(B_{12}CH_i)} + d^2_{(TCH_i)} \geq d^2_{(B_{12}T)} \quad (8)$$

Then the avoidance strategy should be started.

In the case $\angle B_{11}CH_iT$ and $\angle B_{12}CH_iT$ locate at same side of the line CH_iT , if

$$d^2_{(B_{11}CH_i)} + d^2_{(TCH_i)} \geq d^2_{(B_{11}T)}, r_{11} \leq R_1 \text{ or } d^2_{(B_{12}CH_i)} + d^2_{(TCH_i)} \geq d^2_{(B_{12}T)}, r_{12} \leq R_1 \quad (9)$$

Then the avoidance strategy should be started.

6.2. The Direction of Avoidance

To avoid the enemies, the entity should move to the opposite direction of the enemy. Thus the direction should be selected based on the current position instead of the future one. As is shown in Fig.10, $\angle\theta_v = \angle\vec{v}_{Ci}CH_i\vec{v}_{Ei}$, if $0 \leq \angle\theta_v \leq \pi$,then the counterclockwise direction should be taken. Otherwise if $-\pi \leq \angle\theta_v \leq 0$ then the clockwise direction should be taken.

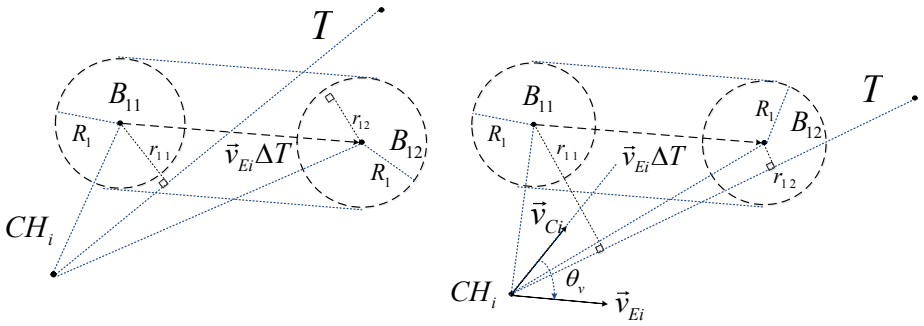


Fig. 10. The judgment of enemy situation

7. Simulation

In order to evaluate our proposal, we performed some scenes to simulate the four parts of the mobility model.

7.1. The Simulation of Random Disturbance

The simulation of the formalized process of Markov random disturbance on the acceleration is given in Fig.11. In a 3500m$\times$3500m simulation area, there are 6 nodes marching towards the destination in 500s while suffering the disturbance which changes every second. b_i=0.9, and $\sigma = 0,1,2,3,4,5$ in the Rayleigh distribution contributing to the Markov disturbance.

It can be seen that the movement of entities are disturbed and the tiny random acceleration changes like the effect of uncertain small-scale terrain factors have on the node. Different values of parameter bring different intensities of disturbance. The higher σ is, the severer disturbance is. This process mimics the influence of small-scale terrain factors on the movement. The uncertainty of topographic feature can be adjusted by the parameter σ.

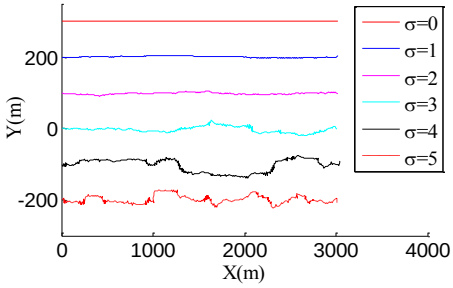


Fig. 11. The mobility trace of nodes with Markov random disturbance

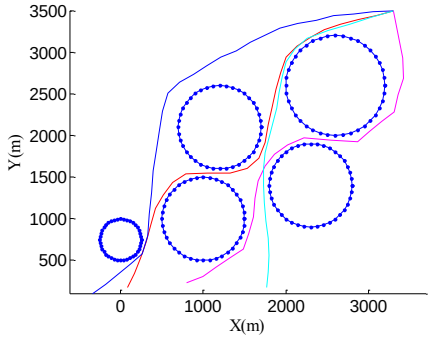


Fig. 12. The terrain information based path optimization with the ABC algorithm

7.2. The Simulation of Path Optimization

The simulation of the terrain information based path optimization is given in Fig.12. In a 4000m$\times$4000m simulation area, there are 4 nodes marching towards the destination. With the terrain information of the battlefield, the

nodes can plan and optimize the path before moving. The simulation parameters are given in Tab. 1 as below:

Table 1. The parameters of the ABC algorithm for path optimization

Parameter	Value	Parameter	Value
ω (barrier/plain)	1000/ 10	Interval of optimization (m)	200(x-axis)
Employed foragers	30	Number of iteration	2000
Unemployed foragers	30	Number of search	30

It can be seen that the ABC algorithm realize the path optimization with the terrain information. The entities bypass the obstacles and select a path that takes the length as well as the terrain into account. This process mimics the path planning of nodes with the topographic information in the battlefield.

7.3. The Simulation of Bypass Strategy

The simulation of the topographic-awareness based bypass strategy is given in Fig.13. In order to contrasting the bypass strategy with path optimization, the simulation scene remains the same with Section 7.2. The detection range of entities is 200m.

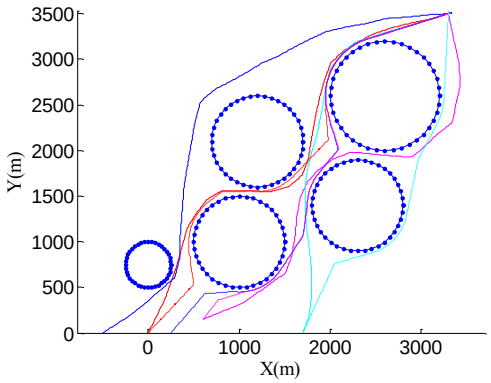


Fig. 13. The mobility route of bypass strategy comparing with path optimization

The solid line is the path optimized by the ABC algorithm in Section 6 while the dotted line is the mobility route of bypass strategy. However, the nodes will not take the bypass strategy until an obstacle is observed in a near distance. This leads to a result that the path is not the optimal selection. The phenomenon mimics the realistic situation when the troops marching in a strange environment. The entities without terrain information can only rely on the observation and awareness of topography to make the decision.

7.4. The Simulation of Avoidance Strategy

The simulation of the situational-perception based avoidance strategy is given in Fig.14 and Fig.15. In the same simulation area with Section 7.2, there are 3 nodes marching towards the destination and 5 enemies patrolling in the simulation area. The simulation parameters are given in Tab. 2 as below.

Table 2. The parameters of the entities and enemies for avoidance strategy

Parameters of entities	Value of entities	Value of enemies
Detection range (m)	200	40
Velocity (m/s)	8-16	5-7
Acceleration limit (m/s^2)	5	0 (constant patrol)

The situation of the entities and enemies in the battlefield is given in Fig.14. There are three enemies patrolling in line and two in circle at a constant velocity. Since the enemies and entities are both moving, we record the distance between one entity and five enemies by time to evaluate if the strategy could allow the entities to avoid their enemies. The five lines in Fig.15 represent the distance between one entity and the five enemies. X-axis represents the time while y-axis represent the distance. It can be seen that the entities will approach the enemy during the movement. But the avoidance strategy will be triggered to guarantee the entity does not run into the enemy.

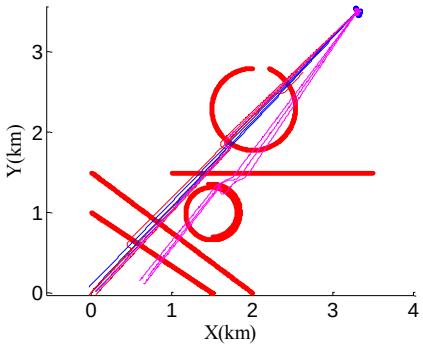


Fig. 14. The situation between the entities and enemies

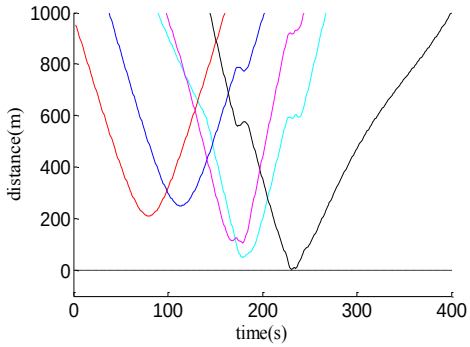


Fig. 15. The distance between one entity and the enemies

7.5. The Simulation of the Synthetic of the Mobility Model

The synthetic of random disturbance, path optimization, bypass strategy and avoidance strategy forms the integral mobility model. Firstly, the node is constantly disturbed by the random terrain factors in small-scale during the whole mobility process. Secondly, if there is a priori knowledge of terrain

information, the node will optimize the path with the ABC algorithm before moving. Thirdly, if the information is lacking, the node will utilize the topographic-awareness based bypass strategy to adjust the path during the movement. Fourthly, the node will observe the enemy situation and make a decision on movement to avoid running into the enemies. The simulation of the synthetic mobility model is given in Fig.16 as follow.

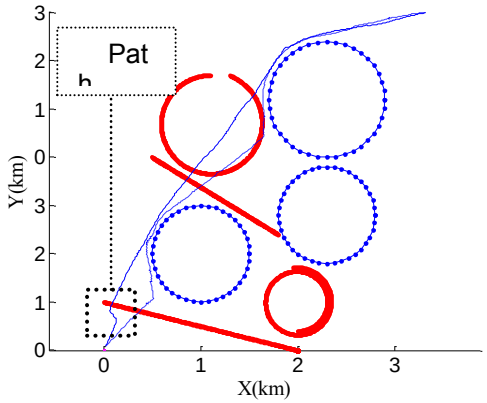


Fig. 16. The trace of the node in the synthetic of the mobility model

The solid line is the path optimized by the ABC algorithm in Section 6 while the dotted line is the mobility trace of bypass strategy. It can be seen that the node will follow the path planned by the optimization or adjusted by the bypass strategy. When an enemy is observed, the node will utilize the avoidance strategy. This phenomenon entirely mimics the realistic situation when the node marching in the battlefield.

7.6. The Parametric analysis

There are five parameters that may have significant influence on the simulation. They are the σ, the entity detection internals, the entity detection range, the entity acceleration and the enemy detection range. In order to evaluate the selection of these parameters, we introduce a concept as "successful bypass rate" which represents the probability that the entity avoids all the barriers and enemies during the movement. The higher the successful bypass rate is, the more appropriate the parameter is within a reasonable range. The simulation results are given in Fig.17 as follow.

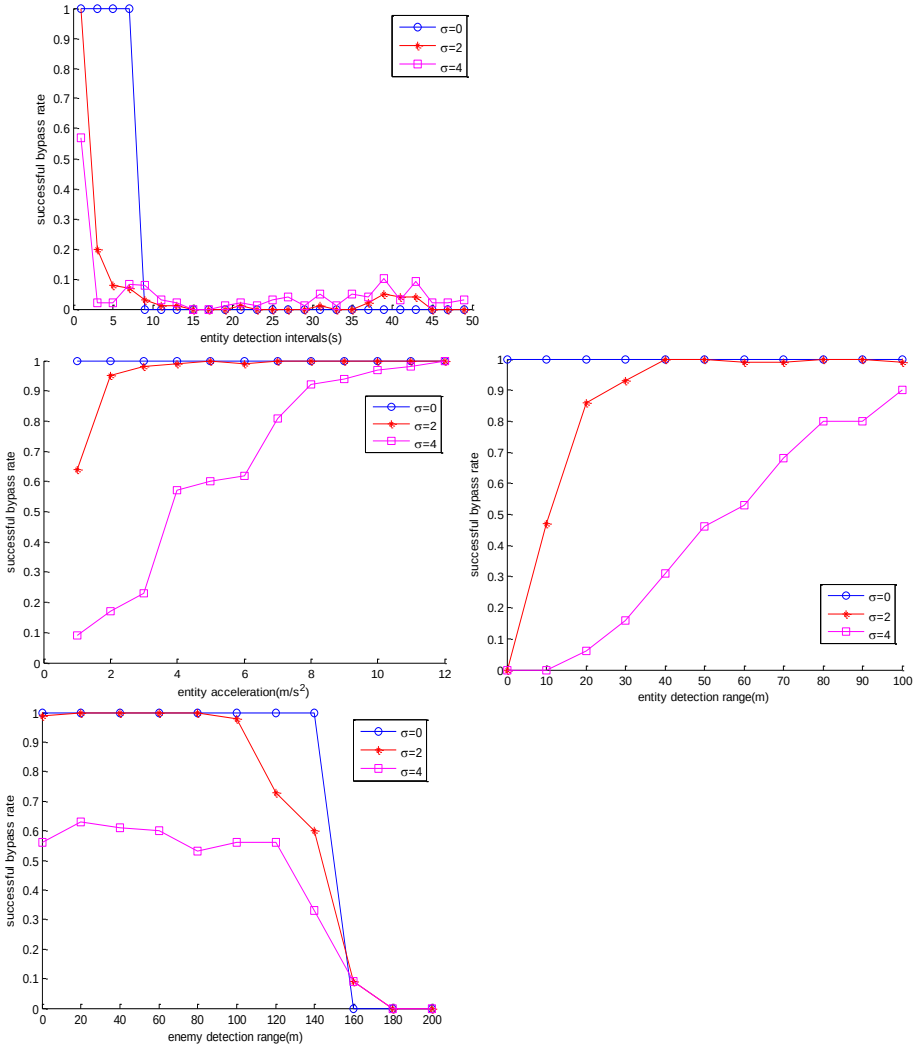


Fig. 17. The analysis of parameters

Firstly, the higher the σ is, the lower the successful bypass rate is acquired. The phenomenon implies that severe disturbance caused by small-scale terrain factors may lower the successful bypass rate of nodes. Secondly, a longer detection internal of entities may lower the successful bypass rate. It suggests that the node will encounter the obstacles or enemies without a frequent detection. Thirdly, high entity accelerations will improve the successful bypass rate which means a rapid adjustment on mobility helps avoiding the dangerous zone in the battlefield. Fourthly, enhance the entity detection ability or weaken the enemy detection ability will bring an

improvement on successful bypass rate. This result suggests the importance of observation in the battlefield environment.

8. Conclusion

In this paper, we propose a topographic-awareness and situational-perception based mobility model with path optimization for tactical MANET to imitate the influence which terrain factors and enemy situations have on the dynamic characteristics of mobile entities.

Firstly, we constructed a formalized process to generate a random acceleration on nodes as the disturbance caused by small-scale topographic factors in the battlefield.

Secondly, we propose a terrain information based path optimization method with the ABC algorithm. We abstract the path planning as a constrained optimization of a multi-dimensional function to mimic the plan when the entities face large-scale terrain factors.

Thirdly, we improve the topographic-awareness bypass strategy to imitate the behavior when the entities without the terrain information face large-scale obstacles in the battlefield.

Fourthly, we establish an enemy situational-perception based avoidance strategy to simulate the behavior when the entity encounters enemies in the battlefield.

The synthesis of the four parts above comprehensively represents the dynamic characteristics of mobile nodes according to the awareness of the terrain factors and enemy situations in the tactical environment. In the future, we plan to evaluate different routing protocols with this mobility model so as to select and improve a protocol for military application in MANET. Our ultimate goal is to realistically imitate the mobility of nodes in tactical environment and precisely simulate the performance of MANET.

Acknowledgement. This work is supported by the National Basic Research Program of China (973 Program) under Grants 2013CB329100 and 2013CB329105，as well as the National Nature Science Foundation of China (Grant No. 61273214).

References

[1] F. Bai, N. Sadagopan, and A. Helmy, “The Important Framework for Analyzing the Impact of Mobility on Performance of Routing for Ad Hoc Networks”, AdHoc Networks Journal, Elsevier Science, 2003, Vol. 1, No. 4, pp. 383-403.

[2] Jack L. Burbank, Philip F. Chimento, Brian K. Haberman, and William T. Kasch, “Key Challenges of Military Tactical Networking and the Elusive Promise of MANET Technology”, IEEE Communications Magazine, 2006, Vol. 44, pp.39-45.

[3] Megat Zuhairi, David Harle, “A Simulation Study on the Impact of Mobility Models on Routing Protocol Performance with Unidirectional Link Presence”, ICOIN 2011, pp.335-340

[4] Ahmed F, Sajjadur R M, “Performance Investigation on Two classes of MANET Routing Protocols Across various Mobility models With QoS Constants”, Computer Networks & Communication, 2011, vol.3, pp.197-215.
[5] Arafatur MRFA, Naeem J, and Sharif MMA, “A simulation based performance comparison of routing protocol on mobile Ad-Hoc Network”, Conference on Computer and Communication Engineering, Kuala Lumpur, 2010, pp.11-13.
[6] Leila Harfouche, Selma Boumerdassi, and Eric Renault, “Towards a Social Mobility Model”, Personal, Indoor and Mobile Radio Communications, IEEE, 2009, pp.2876-2880.
[7] Musoles M, Hailes S, and Mascolo C, “An ad hoc mobility model founded on social network theory”, The 7th ACM international symposium on modeling, analysis and simulation of wireless and mobile systems, 2007, pp.20–24.
[8] Jingbo Sun, Yue Wang, Hongbo Si, Jian Yuan, and Xiuming Shan. “Aggregate Human Mobility Modeling Using Principal Component Analysis”, Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications, 2010, Vol. 1, pp.83-95.
[9] Bowen Deng, Yujia Zhai, and Yue Wang, “A Terrain-Awareness based Mobility Model with Markov Random Disturbance for Tactical MANET”, The Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2012), 2012, pp.224-231.
[10] Karaboga D, “An idea based on honey bee swarm for numerical optimization”, Technical Report 06 Erciyes University, 2005.
[11] Arvind Kumar Shukla, CK Jha, and Deepak Sharma, “The Efficiency Analysis of Mobility Model using Routing Protocols”, International Conference on Advances in Computer Applications (ICACA) 2012, 2012, pp.6-10.
[12] Tracy Camp, Jeff Boleng, and Vanessa Davies, “A Survey of Mobility Models for Ad hoc Network Research”, Special Issue on Mobile Ad hoc Networking: Research, Trends and Applications, Journal of Wireless Communications and Mobile Computing, 2002, pp483-502.
[13] Wenjie Chen, Lihua Dou, Yujin Li, Yunchuan Wei, “Ellipse Group Mobility Model for Tactical MANET”, Control and Decision Conference, 2010, pp.2288-2293.
[14] N Vetrivelan, A V Reddy, “Modeling and Analysing a Novel Restricted Angle Scenario Model in MANET”, TENCON 2010, pp.263-268.
[15] Jie Zhang, Chen Chen, and Robin Cohen. “A Scalable and Effective Trust-Based Framework for Vehicular Ad-Hoc Networks”, Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications, 2010, Vol. 1, pp.3-15.
[16] A. Jardosh, E. M. Belding-Royer, K. C. Almeroth, and S. Suri, “Towards Realistic Mobility Models for Mobile Ad hoc Networks”, ACM MobiCom, 2003, pp.217-229.
[17] Xiaofeng Lu, Yung-chih Chen, Ian Leung, Zhang Xiong, and Pietro Liò, “A Novel Mobility Model from a Heterogeneous Military MANET Trace”, ADHOC-NOW 2008, 2008, LNCS 5198, pp.463–474.
[18] Hyun Seo, Sunghun Kim, Joongsoo Ma, “A Novel Mobility Model for the Military Operations with Real Traces”, ICACT 2010, 2010, pp.129-133.
[19] Tracy Camp, Jeff Boleng, and Vanessa Davies, “A survey of mobility models for ad hoc network research”, Wireless Communications and Mobile Computing, 2002, Vol.2, pp.483-502.
[20] Sudip Misra, Prateek Agarwal, “Bio-inspired group mobility model for mobile ad hoc networks based on bird-flocking behavior”, Soft Computing - A Fusion Of Foundations, Methodologies And Applications, Vol.16, 2012, pp.437-450.
[21] Kashyap Merchant, Wei-jen Hsu, Haw-wei Shu, Chih-hsin Hsu, and Ahmed Helmy, "Weighted Way Mobility Model and its Impact on Ad Hoc Networks", ACM MobiCom, 2004.

[22] Sharma RK, Ghose D, “Collision avoidance between UAV clusters using swarm intelligence techniques”, Journal of Systems Science, 2009, Vol.40, pp.521–538.
[23] Chenchen Yu, Xiaohong Li, and Dafang Zhang, “An Obstacle Avoidance Mobility Model”, ICIS 2010, 2010, pp.130-134.
[24] Yan Wan, Kamesh Namuduri, Yi Zhou, Dayin He, and Shengli Fu, "A smooth-turn mobility model for airborne networks", Airborne '12 Proceedings of the first ACM MobiHoc workshop on Airborne Networks and Communications, pp. 25-30
[25] Karaboga D, Gorkemli B. “A combinatorial Artificial Bee Colony algorithm for traveling salesman problem”, 2011 International Symposium on Innovations in Intelligent Systems and Applications, 2011, pp.50-53.
[26] Chazelle B, “The convergence of bird flocking”, Proceedings of ACM-SIAM symposium on discrete algorithms (SODA09), 2009, pp.422–431.
[27] Akay B, Karaboga D. “A modified Artificial Bee Colony algorithm for real-parameter optimization”, Information Sciences, 2012, Vol. 192, pp.120-142.
[28] Karaboga N, Cetinkaya MB, “A novel and efficient algorithm for adaptive filtering: Artificial bee colony algorithm”, Turkish Journal of Electrical Engineering and Computer Sciences, 2011, Vol.19, pp.175-190.
[29] Karaboga D, Ozturk C. “A novel clustering approach: Artificial Bee Colony (ABC) algorithm”, Applied Soft Computing, 2011, Vol.11.
[30] HU C L, SUN T Y, “Reliable multi-goal route planning for vehicle using skeletonization and genetic algorithms”, Proc 2008 CACS International Automatic Control Conference, 2008, pp.159-263.
[31] Ozturk C, Karaboga D, Gorkemli B. “Artificial bee colony algorithm for dynamic deployment of wireless sensor networks”, Turkish Journal of Electrical Engineering and Computer Sciences, 2012, Vol. 20, pp.255-262.
[32] Karaboga D, Akay B. “A modified Artificial Bee Colony (ABC) algorithm for constrained optimization problems”, Applied Soft Computing, 2011, Vol.11, pp.3021-3031.

Jinhai Huo received the M.E degree in communication and information systems from PLA University of Science and Technology, China, in 2003. He is currently a Ph.D. candidate in the Department of Electronic Engineering at Tsinghua University, Beijing, China. His main research interest is in key technology of mobile ad hoc networks.

Bowen Deng received the B.S. degree in electronic information engineering from the Harbin Institute of Technology, China, in 2010. He is currently a postgraduate student in the Department of Electronic Engineering at Tsinghua University, Beijing, China. His main research interest is in cognitive networks.

Shuhang Wu received the B.S. degree in basic science from the Tsinghua University, Beijing, China, in 2005. He is currently a Ph.D. candidate in the Department of Electronic Engineering at Tsinghua University. His main research interest is in cognitive networks.

Jian Yuan received his Ph.D. degree in electrical engineering from the University of Electronic Science and Technology of China, in 1998. He is currently a professor in the Department of Electronic Engineering at Tsinghua

University, Beijing, China. His main research interest is in complex dynamics of networked systems and dependability of mobile networks.

Ilsun You received his M.S. and Ph.D. degrees in Computer Science from Dankook University, Seoul, Korea in 1997 and 2002, respectively. He is now working as an associate professor at Korean Bible University, South Korea. Dr. You has chaired or is currently chairing international conferences and workshops such as IMIS, MobiWorld, MIST, AsiaARES, and so forth. He is in the editorial board for International Journal of Ad Hoc and Ubiquitous Computing (IJAHUC), Computing and Informatics (CAI), and Journal of High Speed Networks. Also, he has served as a guest editor of several journals such as Information Sciecnes, Journal Computers & Mathematics with Applications, Wireless Communications and Mobile Computing, Mobile Information Systems, and so so. His main research interests include internet security, authentication, and formal security analysis.

Received: July 14, 2012; Accepted: January 21, 2013

DOI: 10.2298/CSIS120711032J

A Real-time Location-based SNS Smartphone Application for the Disabled Population

Hae-Duck J. Jeong[1], Jiyoung Lim[1], WooSeok Hyun[1], and Arisu An[2]

[1] Department of Computer Software
Korean Bible University
Seoul, South Korea
{joshua, jylim, wshyun}@bible.ac.kr
[2] Development G-Team of International Business Division
Mappers
Seoul, South Korea
hiphopbob@naver.com

Abstract. Smartphone usage and data consumption have been sharply rising, and the disabled have also become smartphone users as the number of users of these phones has exponentially increased in recent years. The theme of this paper is how to create a better world for the disabled using the information that people want to exchange with each other between the disabled and the general population. The main goal is also to provide the information that they need from each other in a way that can be displayed on the map in real-time. We propose a new location-based SNS application for the disabled population (except those who are visually impaired or the disabled who are not able to use a smartphone) with three major characteristics of this application to be considered as follows: (i) the person uses a Social Networking Service (SNS) by constructing a friend matching system such as Facebook or Twitter, which are the most widely-used SNS in the world; (ii) the general population registers real-time information for a specific location on the map for the disabled population using SNS. This information with photos and messages is given and evaluated by users; and (iii) this system makes it easier to see that the menu in the GUI was implemented.

Keywords: location-based SNS, Android, smartphone, GPS, disabled population, real-time system.

1. Introduction

According to smartphone usage statistics 2012, AnsonAlex.com [6] announced that Go-Gulf.com recently published an infographic containing worldwide smartphone usage statistics in 2012. 80 percent of the world's population now has a mobile phone. Out of the five billion mobile phones in the world, 1.08 billion are smartphones. With increasing prevalence of smartphones, not only people's life styles, but also trends in the future of information technology have been changed. Many applications that have a wide variety of different functions to

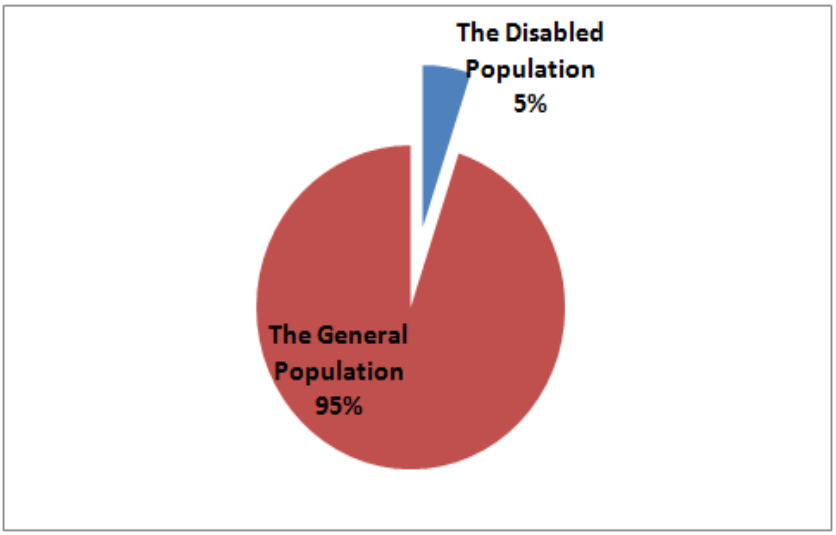


Fig. 1. Registration status of the Disabled Population in South Korea as of December 2011

access easily for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, have been released [15], [16], and telecommunication companies also provide their products at an affordable price [8]. Use of smart technology in education for the one million disabled by including the disabled is currently being conducted in the Seoul city government office [13].

The statistics of Ministry of Health and Welfare of South Korea show that the registered number of the disabled population was five percent of South Koreans as of December 2011 (i.e., 2,519,241 disabled out of 49,779,440 South Koreans), as shown in Fig. 1 [22]. Table 1 also shows the registered numbers for different types of the disabled in South Korea as of December 2011.

National Information Society Agency (NIA) investigated the information gap index and real condition in 2011. Figure 2 shows that, according to the study, the smartphone supply ratio of both the disabled population and the average of socially disadvantaged class (including the disabled, low income users, aged people, and farmers and fishermen) in South Korea was 8.6%. The smartphone supply ratio of the general population was 39.6% as of October 2011 [25].

According to the Organization for Economic Cooperation and Development (OECD) report published in July 2012, South Korea's high-speed Internet penetration rate topped 100 percent for the first time among the group's 34 nations. Application developers around the world are targeting Korea as their gateway into Asia as they strive to integrate globalization with social media. Figure 3 provides quick insight on the current leading SNS applications in South Korea [24].

Table 1. The registered numbers for different types of the disabled in South Korea as of December 2011 [22]

Types of the Disabled	Numbers
Liver	8,145
Epilepsy	8,950
Brain Lesions	260,718
Visual	251,258
Kidney	60,110
Heart	9,542
Face	2,715
Language	17,463
Autism	15,857
Intestinal Fistula	13,098
Psychiatric	94,739
Intellectual	167,479
Physical	1,333,429
Hearing	261,067
Respiratory	14,671
Total	**2,519,241**

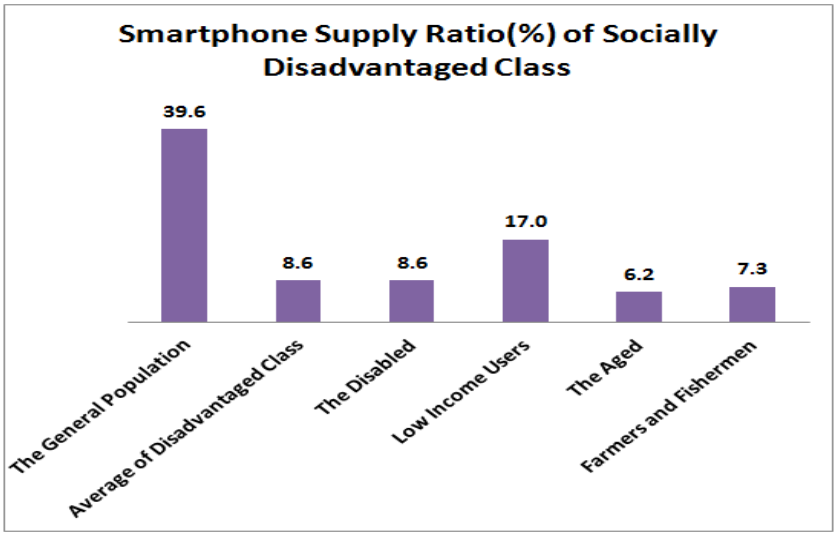


Fig. 2. Smartphone supply ratio(%) of socially disadvantaged class in South Korea in 2011. Sample sizes are 1,500 general people, and 3,700 for each of the disabled, low income users, the aged, and farmers and fishermen. The survey was investigated from August 2011 to November 2011

As of May 2012, Cyworld has the largest user base in the South Korean SNS applications with just a shy of 20 million users. Kakao Talk, a mobile-

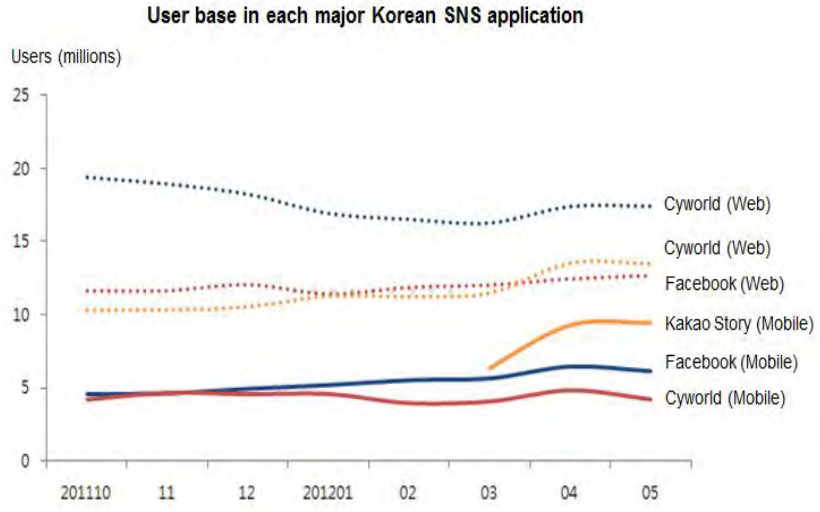


Fig. 3. The chart shown depicts the size of the user bases of the top SNS applications in South Korea as of May 2012 [24]

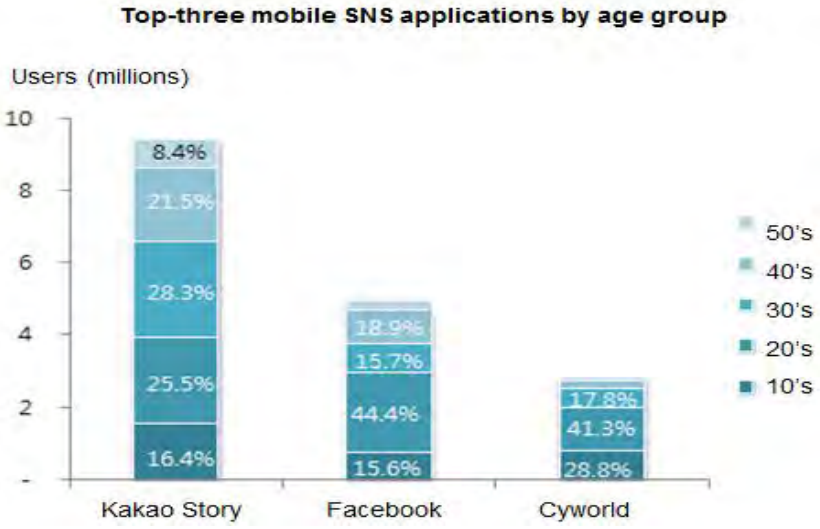


Fig. 4. Of the top three mobile SNS applications, the chart above separates user volume by age group [24]

based messenger, is the second most-used SNS application in South Korea. Facebook, the global leader in social media, ranks third in South Korea.

Figure 4 shows that Kakao Story ranks as the most popular mobile SNS application with its most active user base in their 30's. It is interesting to note that much of Kakao Story's user base is evenly distributed among users in their 20's, 30's and 40's. On the other hand, Facebook and Cyworld have the most dominant user base in their 20's.

Thereby the frequencies of use of SNS have also increased exponentially. Generally people send and receive messages in real-time through smartphones using Facebook and Twitter, use main services via registering friends and matching friends, or easily use location-based SNS such as Foursquare. Almost all the SNS smartphone applications have been developed for the general population, while ones for the disabled have not much been considered and developed yet. Thus, we propose a new system for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, to give the information flow of such interests and activities to the disabled, and the information is evaluated by several consensuses being served. It is designed and implemented based on Android [3]. Section 2 addresses related technologies such as Android, global positioning system, and social networking system. In Section 3, the proposed location-based SNS for the disabled is discussed. Section 4 presents implementation and results of the proposed application, and finally the conclusions are described in Section 5.

2. Related Technologies

A few related technologies of the proposed SNS smartphone application for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, are Android, global positioning system (GPS), and SNS as follows.

2.1. Android

Android is a Linux-based open mobile platform for mobile devices such as smartphones and tablet computers. It is composed of not only an operating system, but also middleware, user interface (UI), browser, and application. It also includes C/C++ libraries that are used in components of various Android systems. Android architecture is divided into five hierarchical categories: applications, application framework, libraries, Android runtime, and Linux kernel [2], [5], [18]. The proposed application was designed and developed on Android.

2.2. Global Positioning System (GPS)

Basic Concepts of GPS The GPS is a space-based satellite navigation system that provides location and time information in all weather, anywhere on or near the Earth. The current GPS is composed of three major segments. These are a space segment, a control segment, and a user segment. The United

States Air Force develops, maintains, and operates the space and control segments. GPS satellites broadcast signals from space, and each GPS receiver uses these signals to calculate its three-dimensional location (latitude, longitude, and altitude) and the current time [1], [12], [21]. The space segment consists of 24 to 32 satellites in medium earth orbit and also includes the payload adapters to the boosters required to launch them into orbit. The control segment is composed of a master control station, an alternate master control station, and a host of dedicated and shared ground antennas and monitor stations. The user segment is made of hundreds of thousands of the US and allied military users of the secure GPS precise positioning service, and tens of millions of civil, commercial, and scientific users of the standard positioning service. The GPS will be used to mark user's current location in the proposed application.

Navigation Equations of GPS The receiver uses messages received from satellites to determine the satellite positions and time sent. The x, y, and z components of satellite position and the time sent are designated as $[x_i, y_i, z_i, t_i]$, where the subscript i denotes the satellite and has the value $1, 2, \cdots, n$, where $n \geq 4$. When the time of message reception indicated by the on-board clock is t_r, the true reception time is $t_r + b$, where b is receiver's clock bias (i.e., clock delay). The message's transit time is $t_r + b - t_i$. Assuming the message traveled at the speed of light, c, the distance traveled is $(t_r + b - t_i)c$. Knowing the distance from receiver to satellite and the satellite's position implies that the receiver is on the surface of a sphere centered at the satellite's position. Thus, the receiver is at or near the intersection of the surfaces of the spheres. In the ideal case of no errors, the receiver is at the intersection of the surfaces of the spheres. The clock error or bias, b, is the amount that the receiver's clock is off. The receiver has four unknowns, the three components of GPS receiver position and the clock bias $[x, y, z, b]$. The equations of the sphere surfaces are defined by

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = ([t_r + b - t_i]c)^2, i = 1, 2, \cdots, n \quad (1)$$

or in terms of pseudoranges, $p_i = (t_r - t_i)c$, as

$$p_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - bc, i = 1, 2, \cdots, n. \quad (2)$$

These equations can be solved by algebraic or numerical methods such as Bancroft's algorithm, trilateration, multidimensional Newton-Raphson calculations, and additional algorithms for more than four satellites below [10], [12].

– **Bancroft's algorithm**
  Bancroft's algorithm involves an algebraic as opposed to numerical method and can be used for the case of four satellites or for the case of more than four satellites. If there are four satellites then Bancroft's method provides one or two solutions for the four unknowns. If there are more than four satellites then Bancroft's method provides the solution which minimizes the sum of the squares of the errors for the over determined system [7], [12].

– **Trilateration**
The receiver can use trilateration and one dimensional numerical root finding. Satellite position and pseudorange determines a sphere centered on the satellite with radius equal to the pseudorange. Trilateration is used to estimate receiver position based on the intersection of three sphere surfaces so determined. In the usual case of two intersections of three sphere surfaces, the point nearest the surface of the sphere corresponding to the fourth satellite is chosen. Let d denote the signed distance from the current estimate of receiver position to the sphere around the fourth satellite. The notation, d(correction) denotes this as a function of the clock correction. The problem is to determine the correction such that d(correction) $= 0$. This is the familiar problem of finding the zeroes of a one dimensional non-linear function of a scalar variable. Iterative numerical methods, such as those found in the chapter on root finding in Numerical Recipes [23] can solve this type of problem.

– **Multidimensional Newton-Raphson calculations**
Alternatively, multidimensional root finding methods such as the Newton-Raphson method can be used [23]. The approach is to linearize around an approximate solution, say $[x^{(k)}, y^{(k)}, z^{(k)}, b^{(k)}]$ from iteration k, then solve the linear equations derived from the quadratic equations above to obtain $[x^{(k+1)}, y^{(k+1)}, z^{(k+1)}, b^{(k+1)}]$. Although there is no guarantee that the method always converges due to the fact that multidimensional roots cannot be bounded, when a neighborhood containing a solution is known as is usually the case for GPS, it is quite likely that a solution will be found. It has been shown that results are comparable in accuracy to those of Bancroft's method [17]. This algorithm was used to implement our proposed system.

– **Additional methods for more than four satellites**
When more than four satellites are available, the calculation can use the four best or more than four, considering number of channels, processing capability, and geometric dilution of precision (GDOP). Using more than four is an over-determined system of equations with no unique solution, which must be solved by least-squares or a similar technique. If all visible satellites are used, the results are as good as or better than using the four best. Errors can be estimated through the residuals. With each combination of four or more satellites, a GDOP factor can be calculated, based on the relative sky directions of the satellites used. As more satellites are picked up, pseudoranges from various 4-way combinations can be processed to add more estimates to the location and clock offset. The receiver then takes the weighted average of these positions and clock offsets. After the final location and time are calculated, the location is expressed in a specific coordinate system such as latitude and longitude, using the WGS 84 geodetic datum or a country-specific system [9].

2.3. Social Networking Service (SNS)

The number of SNS users in South Korea is on the sharp rise. This is in line with the SNS craze that is taken over the entire world. In 2010, a report estimated that one out of every 14 people in the world were SNS users. In South Korea alone, there were 24 million SNS users reported in July of 2010 [19]. SNS is not just something that is in with the younger generation, either. More and more older people are following in the footsteps of their younger counterparts. Gartner indentifies that one of the top 10 consumer mobile applications for 2012 is social [11]. The keyword called social is one of the hot issues and furthermore it has seriously to be considered as one of the important information technologies. As applications have been developed for mobile devices, a new paradigm has come with social relations among people (namely, social SNS becoming more important). Who the experts or issue makers are and who are connected with them have become important in the current SNS age [20].

An SNS is an online service, platform, or site that focuses on building and reflecting social networks or social relations among people who share interests and/or activities. An SNS consists of a representation of each user, his/her social links, and a variety of additional services. Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Online community services are sometimes considered as a social network service, though in a broader sense, because social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks [14]. Some examples of these relations-centered popular social networking websites are Facebook, LinkedIn, MySpace, Twitter, and Cyworld. Many other social networking websites are slowly gaining ground with more users signing up every month. Particularly, Facebook is one of the most widely-used SNS in the world. Currently, there are over 800 million users worldwide. It was started by a Harvard student, Mark Zuckerberg, in 2004. It originally started as a site for university students. It's now for anyone over the age 13 with a valid email address. It helps you stay connected with your friends. You can post pictures, video clips, status updates, and wall messages to your friends. There are many game applications and a chatting feature as well. With smartphones, you can access your Facebook account wherever you are [19].

3. Proposed Location-based SNS for the Disabled

Almost all the SNS smartphone applications have been developed for the general population, while ones for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, can hardly be found. Thus, we proposed a new SNS smartphone application for the disabled. Major concepts for the proposed system include the following.

3.1. Major Concepts of the System

The proposed location-based SNS system under Android includes 10 features for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone [4].

- Function of displaying RSS (Really Simple Syndication or RDF Site Summary) information supported by Seoul city with public toilet information installed for the disabled on the map. It consists of receiving RSS information in real-time and easily seeing near public toilets.
- Function of board and public notice. Anyone can write a message on a board and administrative committee members upload public notices for users without any difficulties.
- Function of registering membership and automatic login. Users easily register membership through GUI and can upload one's photo. When a user first logs in, one can then login the system through the automatic login function without the typing effort of login ID and password.
- Function of marker registration on the map. As one of the most important functions in the proposed system, users register a marker at a current position. Users can see a marker of a particular location to be able to register it, and an event in real-time on the map will be able to be given when passing near the place. Real-time information such as photos and messages is registered.
- Function of evaluating real-time information. The disabled and the general population evaluate for markers. The advantage of the evaluation system is that their accumulated points obtained from markers, which were registered positions, were evaluated with points that would become reliable information. The range of evaluation points is from 1.0 to 5.0.
- Function for management of friends. It is an important social regional part in SNS such as searching friends, adding friends, requiring friends and complying with their requests, etc. It shows who requests me as a friend through a specific column in a database, and it is implemented to be able to know peer-to-peer relationships using identifiers in a database.
- Function of showing one's opinion and thought such as News Feed of Facebook.
- Function of TTS (Text To Speech). It provides automatic conversion of text streams to voice on a smartphone giving what the meaning is for a touched location. The disabled and/or the general population are useful if they use Text-To-Speech (TTS) function for the purpose of double-checking texts.
- Function of voice recognition. Using voice-recognition engine supported by Android, we use services supported that call words converted from one's voice instead of buttons, and recognize them.
- Function of navigating in an area. It can only see markers within a 500m radius to reduce much overhead because there are too many markers displayed on the map. One of major concepts of the system has function of navigating in an area that we can easily find where the nearest restroom is

and estimate how far the distance is from the current place to the nearest restroom within a 500m radius on the map.

Thus, the implemented system provides 10 major functions plus additional ones such as modification of personal information, editing photos, member's management, and so on.

3.2. Administration Mode

The management for the proposed system must be through the administration mode. The administrator maintains the system to bridge between the disabled and the general. People have access to all features except they cannot create new additional applications and define specific operations. The administrator also has all the authority including uploaded comments, and managing members' information. On the other hand, the administrator manages each member's information and has the right of access for members to provide security protection. Furthermore, every member including the disabled and the general population can become this administrator by himself/herself, because of the way this proposed system was particularly designed and implemented. There are two aspects in this system: while one administrator covers the general (for example, the general population provides information on the map that the disabled need.), the other manages the disabled (for example, the disabled give the general information to upload on the map.).

3.3. System Configuration and Action

Configuration and action of the proposed system is explained in this section. As shown in Figure 5 its abstract structure consists of Android 2.1, PHP, MySQL, XML, and Apache, and the proposed system was implemented through data transmission between a server and a terminal. There are several processing steps of the data transmission as follows: Send information from an Android terminal and receive it in the post method of PHP. Mainly three types of libraries, AdoDB 5 library for accessing databases, nusoap library for using public information in Seoul, and GD library for importing related images, have been used in this application. Then process query statements in a MySQL database and receive the result in PHP, and then return back to the Android. When sending information that is required in PHP, it is sent to the Android terminal which uses it after constructing it in XML method.

Figure 6 shows a sequence diagram of these steps for a basic message information exchange between users and the abstract structure. When requesting information, one sends it to a database through a few steps and takes it from the database via the mediator PHP. For example, when sending data of a photo type, after storing a photo in a specific path of a server in PHP, an address of the server with the specific path that the photo is stored in the database is combined. A string of a linked type is stored to reduce overhead in a database.

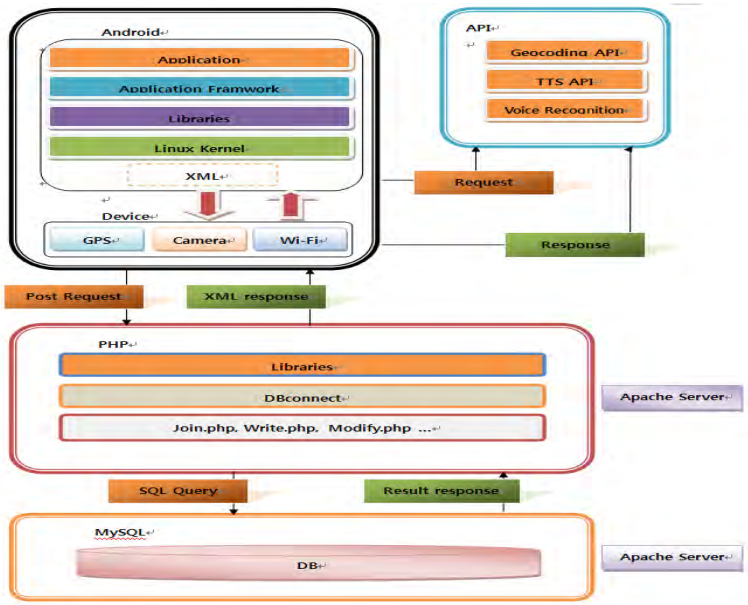


Fig. 5. The abstract structure of the proposed system

Thus, a string of the photo's link path only has to be sent to the terminal instead of bring the whole photo file in a database or send it back to the Android terminal.

One of several non-commercial DBMS products such as MySQL, SQLite and PostgreSQL was taken into account rather than commercial ones because they are free or opensource. MySQL was considered when designing a physical data model in this application. Figure 7 shows a physical data model for the proposed system using the MySQL database management system.

Figure 8 shows the whole state diagram of the proposed system. The structure is composed of the initial main menu that is divided into all submenus between start and end of the system. The system considers that configuration of the initial main menu for the disabled is very important. Fig. 9 also shows a flowchart of writing scenarios which sends the data, including coordinate values and information of photos from input fields, to the database using PHP. Figures 10 and 11 show flowcharts for friend matching and inserting coordinate values, respectively.

4. Implementation and Results of the System

Figure 12 and Figure 13 show implementation and results of the proposed application including 10 functions as mentioned in Section 3. There are six selected menus in Figure 12 and Figure 13 such as main menu, registration, on TTS,

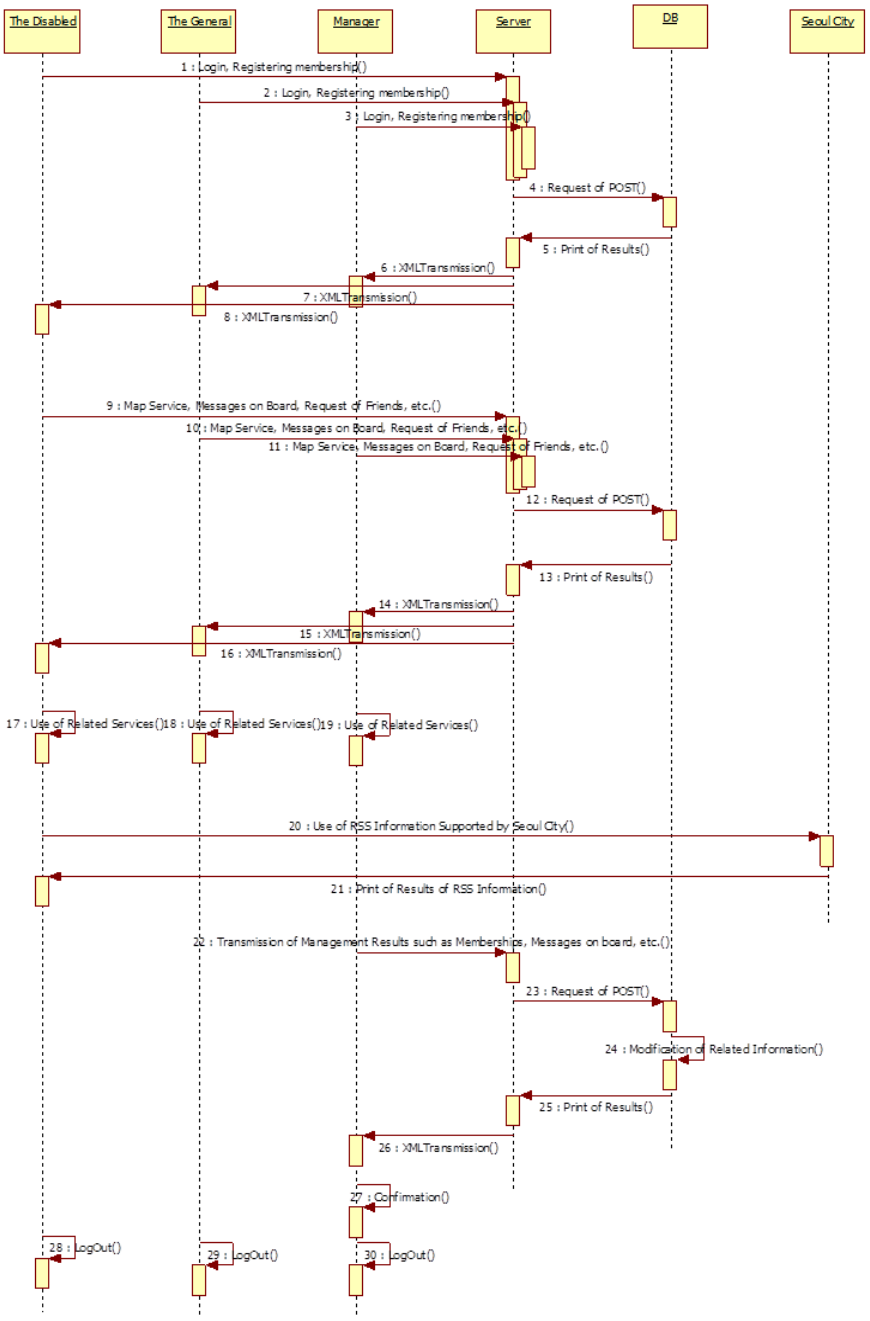


Fig. 6. A sequence diagram for a basic message information

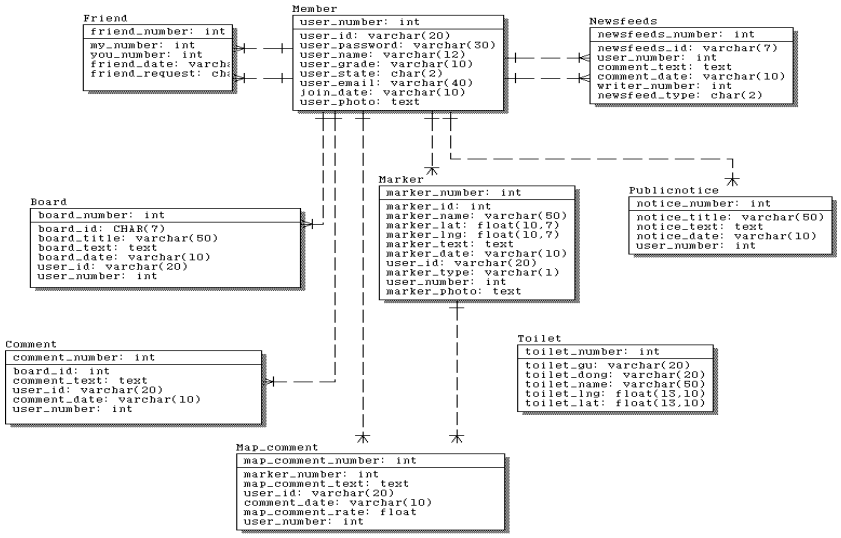


Fig. 7. A physical data model for the proposed system

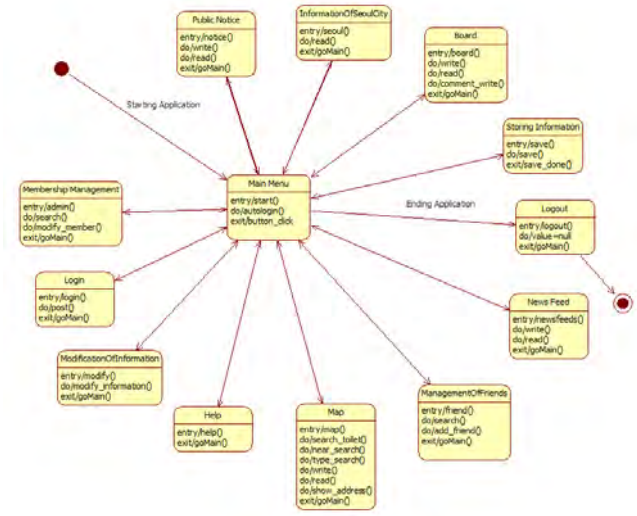


Fig. 8. The proposed state diagram

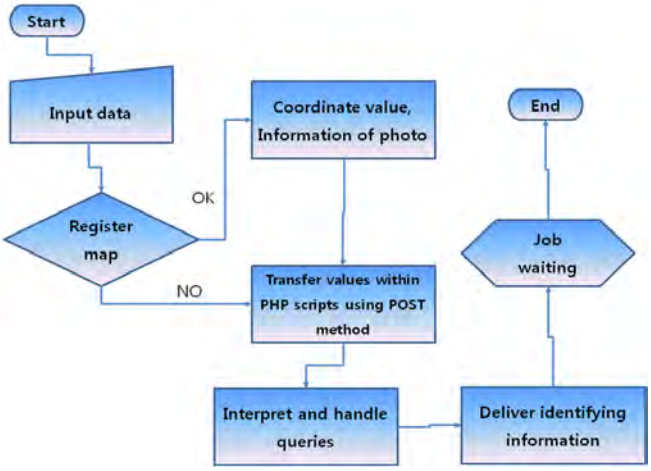


Fig. 9. A flowchart for writing scenarios

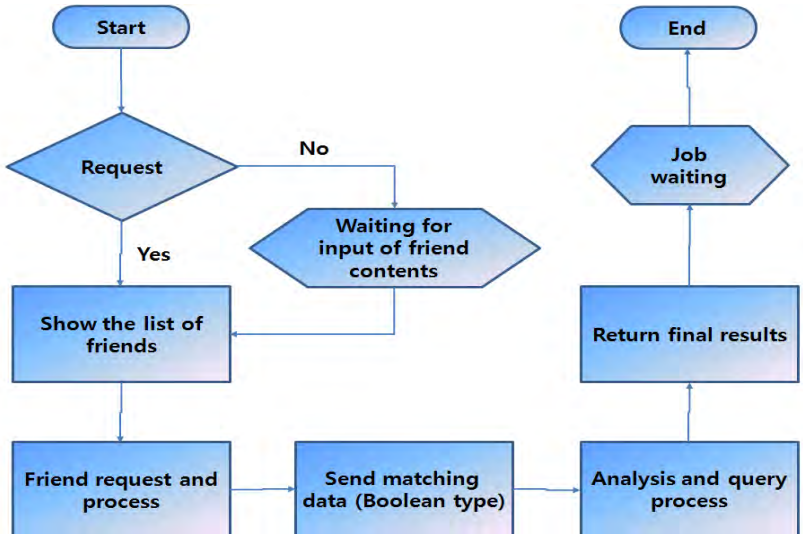


Fig. 10. A flowchart for friend matching

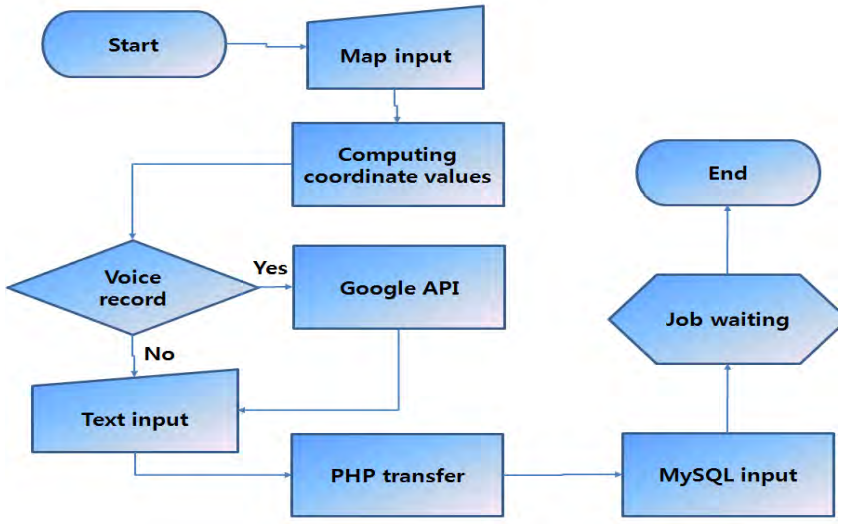


Fig. 11. A flowchart for inserting coordinate values

(a) Main menu (b) Registration (c) On TTS

Fig. 12. Some of selected menus

(a) User's current location (b) Nearby public toilet (c) Toilet information

Fig. 13. Some of selected menus (continued)

user's current location, nearby public toilets, and information about the toilets. The disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, can not only easily use this system with 10 functions plus other ones because big identifiers have been added, but also it will be handy to find near public toilets installed for the disabled on the map. The GUI-based main menu was designed and implemented to easily access because the main menu plays an important role in the disabled and is seriously taken into account and is linked to all other submenus [4]. In addition, soft colors were used to make it feel comfortable and peaceful because the disabled generally handle applications and computer systems more difficulties than the general population.

5. Conclusions and Future Works

A location-based SNS system for the disabled, except people who are visually impaired or the disabled who are not able to use a smartphone, was proposed and implemented in this paper. The system was based on GUI overall and was easily constructed. Furthermore, soft colors were used to make it feel comfortable and peaceful. The proposed application has taken into account ways to best organize, post and get messages in real-time, even though it consists of so many complicated functions. Our results with a real terminal showed that users can get nearby information wherever GPS information is received. According to the need, users can also use the information about public toilets basically provided by Seoul city. Another advantage of the proposed system is that without visiting a specific location in person, users can see and find all events that happen in that place in real-time and a variety of information in advance. Further study to improve performance will be considered to get faster response speeds. This application may cause bottlenecks in the data flow since there is no function to manage photos and data that both require quick process within a short

period of time. On top of that, more complex menus by using hypertext will be considered in the future work. Even though this application was designed and implemented with the help and support of special education professionals for the disabled, the experimental results will also be done by asking the disabled population to use the proposed system later.

Acknowledgments. The authors would like to thank to funding agency for providing the financial means. Parts of this work were supported by a research grant from Korean Bible University, South Korea. We also thank Robert Hotchkiss and reviewers for his invaluable and constructive remarks and suggestions.

References

1. Difference between Tokyo Datum and WGS-84 Datum. http://kin.naver.com/qna/
2. Aguero, J., Rebollo, M., Carrascosa, C., Julian, V.: Does Android Dream with Intelligent Agents? Advances in Soft Computing 50, 194–204 (2009)
3. An, A., Jeong, H.D., Lim, J., Hyun, W.: Design and Implementation of Location-based SNS Smartphone Application for the Disabled Population. In: The Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing: Future Internet and Next Generation Networks (FINGNet-2012). pp. 365–370. Palermo, Italy (2012)
4. An, A., Jeong, H.D., Lim, J., You, I.S., Hyun, W.: The Real-time Interactive Information Supporting System for the Disabled and General Population (2011)
5. Android: Android Operating System, Wikipedia. http://en.wikipedia.org/wiki/Android OS
6. AnsonAlex: Smartphone Usage Statistics 2012. http://ansonalex.com/infographics/smartphone-usage-statistics-2012-infographic/ (2012)
7. Bancroft, S.: An Algebraic Solution of the GPS Equations. IEEE Transactions on Aerospace and Electronic Systems AES-21, 56–59 (1985)
8. Choi, K.S.: Coming out the Cost of a Smartphone for the Disabled and the Elderly. Digital Times (2011)
9. Dana, P.: Geometric Dilution of Precision (GDOP) and Visibility. University of Colorado at Boulder
10. Darwin, I.: Android Cookbook. O'Reilly Media, Inc. (2012)
11. Gartner: Gartner Identifies 10 Consumer Mobile Applications to Watch in 2012. http://www.gartner.com/it/page.jsp?id=1544815
12. GPS: Global Positioning System, Wikipedia. http://ko.wikipedia.org/wiki/GPS)
13. Jeong, D.L.: Smartphone Education for the One Million Disabled in Seoul. Welfare-news.net (2011)
14. Kim, J.Y., Son, D.H., Kim, H.J.: Trend of SNS Technology. Korean Institute of Information Scientists and Engineers 29(11), 9–16 (2011)
15. Kim, S.I.: Meaning and Prospect of Siri. http://www.digieco.co.kr/KTFront/report/ (2011)
16. Kim, T.K.: Design and Implementation of Korean TTS Service based on Android OS. Master's thesis, Department of Computer Engineering, Wonkwang University (2011)

17. Kumar, B.H., Reddy, K.C., Namassivaya, N.: Determination of GPS receiver position using Multivariate Newton-Raphson Technique for over specified cases. International Journal of Applied Engineering Research 13(11), 1457–1460 (2008)
18. Lee, C.Y., An, B., Ahn, H.Y.: Android based Local SNS. Institute of Webcating, Internet Television and Telecommunication 10(6), 93–98 (2010)
19. Lee, H.S., Clyde, J.: English Speaking Professional Program of EBS FM Radio. Doosan-donga, Korea (2011-2012)
20. Lim, S.S., Kim, H.A., Heo, W.R., Jung, K.: A study of Optimization of Spread of Information and Phenomena of Phase Transition in a Social Network. Korean Institute of Information Scientists and Engineers 29(11), 37–43 (2011)
21. Liu, H., Xia, F., Yang, Z., Cao, Y.: An Energy-Efficient Localization Strategy for Smartphones. Computer Science and Information Systems 8(4), 1117–1128 (2011)
22. Ministry of Health and Welfare of South Korea: The Registered Numbers for Different Types of the Disabled in South Korea as of December 2011 (2012)
23. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge (1986)
24. SeoulSpace: Quick Statistics on the Leading SNS Applications in South Korea. http://seoulspace.co.kr/2012/07/30/quick-statistics-korea-sns-applications/ (2012)
25. Yun, H.W.: Smartphone Supply Ratio 8.6% of Socially Disadvantaged Class. Weekly Kyunghyang, No. 973, http://weekly.khan.co.kr/khnm.html?mode=view&artid=201204251118011&code=116 (May 2012)

Hae-Duck Joshua Jeong (hdjjeong@gmail.com) is an associate professor in the Department of Computer Software at Korean Bible University, Seoul, South Korea. He received his Ph.D. in Computer Science and Software Engineering from the University of Canterbury, New Zealand. He is the author or co-author of more than 80 research publications, including more than twenty patents. He has also been the co-chair of NeoFusion'2012-2013, IMIS'2009-2013, FINGNet'2011-2013 and MUE'2007 as well as a member of numerous technical program committees for international conferences. Dr. Jeong is on the editorial board and a reviewer for various domestic and international journals. He is the corresponding guest editor or guest editor of IJICIC, ComSIS and MCM. His research interests include teletraffic modeling, stochastic simulation, multimedia telecommunication networks, intrusion detection system, social networking service, and real-time system. Member of IEEE NZ, KIPS, KSII, and ORSNZ.

Jiyoung Lim (jylim@bible.ac.kr) received her B.S. and M.S. degrees in Computer Science at Ewha Womans University, South Korea, in 1994 and 1996, respectively and received her Ph.D. degree in Computer Science and Engineering from Ewha Womans University in 2001. She is currently an associate professor of Computer Software at Korean Bible University, Seoul, South Korea. Her research interests include wireless/sensor network security, and M2M network security.

Wooseok Hyun (wshyun@bible.ac.kr) is an associate professor in Computer Software at Korean Bible University, Seoul, South Korea. She received her

Ph.D. in Computer Science from Gyeongsang National University, South Korea. She is the author or co-author of more than 30 research publications, including five patents; organization chair of MUE'2007; reviewer of various domestic and international journals. Her research interests include ubiquitous computing, intelligent system, fuzzy system, information retrieval system, and artificial intelligence. Member of KIISE, KIPS, KMMS.

Arisu An (hiphopbob@naver.com) is a researcher in the Development G-Team of International Business Division at Mappers, South Korea. He received his B.E. in Computer Software at Korean Bible University in 2012. His research interests include social networking service, real-time system, digital map, GIS solution, and GPS navigation map.

Received: September 30, 2012; Accepted: January 31, 2013.

DOI: 10.2298/CSIS121101033H

Activity Inference for Constructing User Intention Model

Myunggwon Hwang, *Do-Heon Jeong, Jinhyung Kim, Sa-kwang Song, and Hanmin Jung

Korea Institute of Science and Technology Information (KISTI)
Daejeon, Republic of Korea
{mgh, heon, jinhyung, esmallj, jhm}@kisti.re.kr
*Corresponding Author

Abstract. User intention modeling is a key component for providing appropriate services within ubiquitous and pervasive computing environments. Intention modeling should be concentrated on inferring user activities based on the objects a user approaches or touches. In order to support this kind of modeling, we propose the creation of object–activity pairs based on relatedness in a general domain. In this paper, we show our method for achieving this and evaluate its effectiveness.

Keywords: Ubiquitous and pervasive computing, context awareness, user intention modeling, lexical cohesion.

1. Introduction

Ubiquitous and pervasive computing (UPC) is a post-desktop model of human–computer interaction in which information processing has been thoroughly integrated into everyday objects and activities[1]. The fundamental basis of UPC is context awareness, which makes it possible for computers to both sense and react to user behaviors based on the user's environment. By context we mean "any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object) that are considered relevant to the interaction between a user and an application, including the user and the application themselves [1]." The aim of UPC is to understand the context, grasp the user's intention, and provide suitable services in proper time. Moreover, understanding the context means to know where the user is and what the user does. To this end, many studies have dealt with user intention modeling to construct rules or relatedness between contexts and user intentions.

Various methods of user intention modeling have been proposed and been shown to perform well under specific conditions [2, 3, 4, 5, 6]. Unfortunately,

[1] Ubiquitous computing - Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Ubiquitous_computing, 22 Sep., 2012.

good performance has been largely domain-bound and dependent on narrow and/or artificial assumptions about intended actions.

User intention modeling must contend with the complexity of *n*:*m* mapping between contexts and intentions. For example, when a subject uses a computer at his desk in an office, a UPC system cannot infer his exact intentions, because the context is related to many intentions. The focus of this study is to improve the effectiveness of multiple mappings for user intention modeling. Assuming that most user intentions involve engagement of an object, suitable services can be provided if the UPC system can determine the activities related to the various objects the user approaches and touches. In a previous paper [7], we proposed a method to overcome the limitations faced in earlier works. In this paper, we employ similar techniques to support more flexible and reliable modeling of user intentions, based on the lexical cohesion (similarity measurement) between nouns and verbs.

Text strings are used to represent all tangible and intangible objects (nouns) as well as human activities (verbs). To collect human activities appropriate for given objects, we employ WordNet, Google n-gram data, and the Dice coefficient. Specifically, WordNet is used for selecting verbs to describe main activities, while Google n-gram data, a massive corpus of collective intelligence, is used to calculate lexical cohesions. Our evaluation of this method suggests that it can provide great convenience in preparing object–activity pairs, through processing of collective intelligence data.

This paper[2] is organized as follows. Section 2 describes work related to our research. In section 3, we explain our proposed similarity measurement between objects and activities. In section 4, we evaluate our method and assess its contribution. Finally, in section 5, we summarize our findings.

2. Related Works

The main of UPC is to detect user intentions accurately and to provide services appropriate to these intentions. Assuming that understanding user intention is essential, research on UPC must correlate strongly with that on human–computer interaction (HCI) and Ambient Intelligence [17]. In addition, to understand user intention, it is mandatory to perform interrelation modeling between user activities and surrounding objects. User intention can vary according to places and objects. Various methods have been proposed to recognize place- and object-dependent intentions, such as data mining methods based on a large corpus [8], machine learning methods [9, 10], ontology-driven methods [11], and information retrieval-based methods [7]. We extended some of the information retrieval-based techniques and applied them to a domain of 17 different activities common to an office area In the previous work [7]. We measured similarities between these intentions and

[2] This paper is an extension of [15] presented at the IMIS 2012 conference. It contains additional content such as more examples for easy understanding and new experimental results with abundant test sets.

surrounding objects based on the n-gram dataset from Google using Bayesian probability. Figures 1 and 2 provide an example output from the application we have developed.

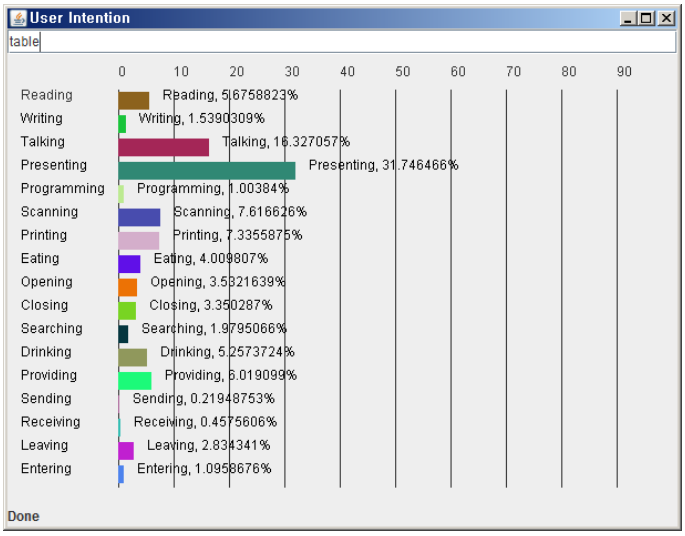


Fig. 1. An application for user intention modeling (object: 'table')

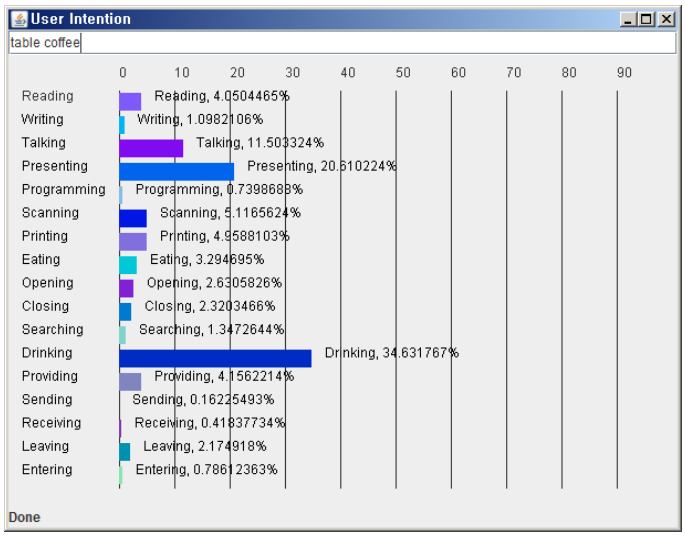


Fig. 2. An application for user intention modeling (objects: 'table' and 'coffee')

Figure 1 shows the relatedness between the single object 'table' and each of our 17 activities. The graphed output shows that when a user approaches or touches a table, a UPC system can expect certain activities more than

others, namely, ‘presenting,’ ‘talking,’ ‘reading,’ ‘providing,’ ‘scanning,’ and so on, in order of relatedness. Figure 2 shows the results when two objects, ‘table’ and ‘coffee,’ are approached at the same time. The addition of the ‘coffee’ object to the ‘table’ object provides a further clue for assessing the relatedness of activities, so that in this case the application shows ‘drinking’ as the most likely user intention. Based on this intention, a UPC system can provide appropriate services related to ‘water’ and ‘cup,’ for example.

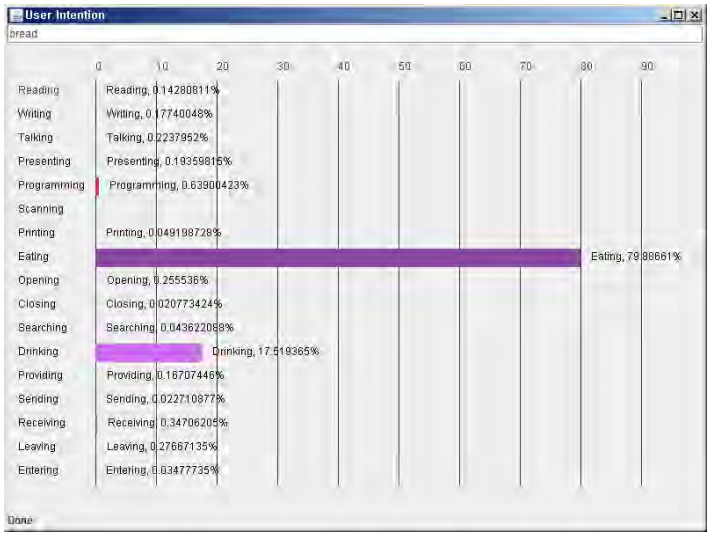


Fig. 3. An application for user intention modeling (object: ‘bread’)

Figure 3 shows a case in which a single approached object, ‘bread,’ relates not only to the likely activity of ‘eating’ but also to the associated activity of ‘drinking’. Thus, a UPC system can expect that someone with a piece of bread likely intends to eat the bread and possibly drink something soon after. Based on results from the previous work, we have found that these kinds of linkages between objects near the user and possible activities involving those objects can be obtained through analysis of large sets of data. However, the work restricted the domain of user intentions to 17 commonplace activities. Therefore, we concentrate on making unrestricted matrix between activities and objects in this research.

3. Inferring Human Activities

At the core of our system, and of future UPC environments, there must be a base of fundamental data for user intention modeling. Figure 4 illustrates the proposed architecture for context-aware applications in UPC. The components presented in this paper are shown in grey.

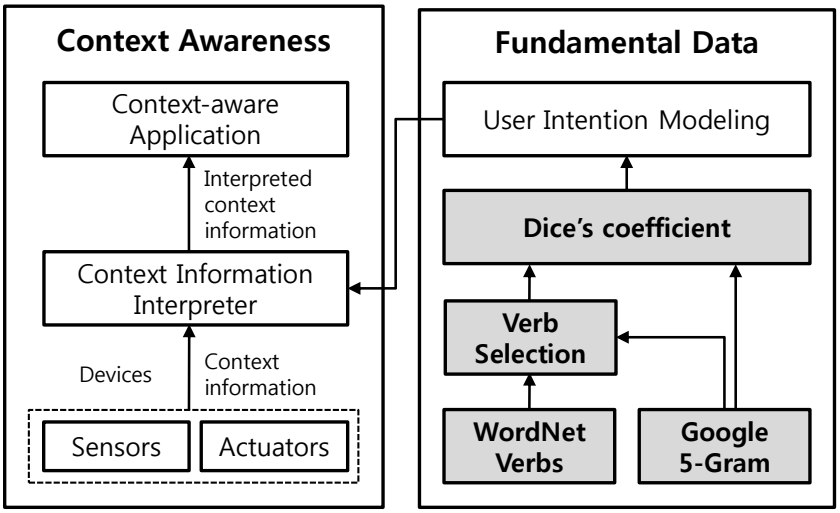


Fig. 4. Architecture for proposed context awareness

3.1. Data Resources

As mentioned in section 1, we used two primary data resources: WordNet and N-Gram. Here we describe these resources in greater detail.

WordNet. We used WordNet[3] to select verbs that describe the human activities needed for modeling most user intentions. Developed at Princeton University, WordNet is a lexical database of English based on psycholinguistics. It has been continuously expanded since 1985 [12, 13] and contains nouns, verbs, adverbs, adjectives, and their definitions, along with semantic networks and subject categories. Though our original work in this area [15] used both nouns and verbs from WordNet, the system proposed here uses only the verbs describing major human activities. This was done because such human activities form a relatively stable set compared to the set of nouns, which is constantly expanded by new products, new issues, social phenomena, etc [18, 19].

[3] WordNet (A lexical database for English): http://wordnet.princeton.edu/

Google N-Gram. We used the 5-gram data from Google n-gram as a broad corpus for similarity measurements between objects and activities. Because n-gram was made by the collective intelligence of people all over the world, we could be confident of a certain degree of objectivity to our measurements. Google n-gram provides token types ranging from a 'unigram' to a '5-gram' of constituent tokens extracted from a huge amount of documents on January, 2006 [14]. Some statistics and examples from N-Gram are listed in Table 1. The data is provided by LDC (Linguistic Data Consortium)[4] [7]. For our system, we chose the 5-gram as the best collection to represent lexical cohesions and specific contexts with semantics. We considered both the words in each token and the count of those words to be important factors for activity inference.

Table 1. Statistics and examples of Google n-gram

Token types	Number of tokens	Token examples (count)
Unigram	13,588,391	Mobile (94,162,727) Phone (151,683,102)
Bigram	314,843,401	Mobile phone (9,414,886) Smart phone (965,621)
Trigram	977,069,902	talking about phone (326) talking on phone (4,144)
4-gram	1,313,818,354	talking on mobile phone (3,888) talks with mobile phone (57)
5-gram	1,176,470,664	protection for your mobile phone (4,372) talking on mobile phone with (271)

3.2. Selection of Activities

WordNet contains 11,488 verbs in total. As a fundamental design principle, we wanted to employ all of these verbs in our system. However, we recognized that this amount of data would likely cause large processing delays. For this reason, we focused on a small subset of verbs describing common human activities. If certain verbs were used to describe human activities frequently, they were also expressed in text. Based on this assumption, we wrote a simple program to calculate the frequencies of verbs in a Google 5-gram. The basic steps of this program are as follows:

Step 1. All WordNet verbs are prepared. Each verb is used for

Step 2. In order to minimize word variances between WordNet and Google 5-gram, all the words are stemmed by Porter stemmer[5].

Step 3. Stop words (including 'be' verbs) are removed.

Step 4. Document frequencies (*df*) are calculated based on the total counts of tokens for each verb. For example, in table 1, the verb 'talk'

[4] LDC (Linguistic Data Consortium): http://www.ldc.upenn.edu/

[5] Porter Stemming Algorithm: http://tartarus.org/martin/PorterStemmer/

also implies 'talking' and 'talks' and thus yields a total count of 5 in the example data.

Step 5. Verbs are filtered by their *df*. We set a threshold value for *df* at 10,000 to remove less common human activities. This filters out all but the most frequently used 20% of activities.

When executed, this program returned a result set of 2,727 WordNet verbs: $vs = \{v_i, 0 < i <= n\}$, where *vs*, *v*, and *n* indicate verb set, a verb, and |*vs*| (2,727 in this work).

3.3. Similarity Measurement

This section describes how we measure similarity as a basis for calculating relatedness between our set of human activities and world objects. An object can be related to various activities in different degrees. For example, the single object 'table' might relate to activities such as 'presenting,' 'talking,' 'reading,' and so forth (as in figure 1), whereas the two objects 'table' and 'coffee' together might relate more strongly to a smaller range of activities (as in figure 2). Google 5-gram contains almost all verbs and nouns, and provides occurrence counts for every token. All world objects and all human activities are expressed in nouns and verbs found in text documents, and if an object is related to an activity deeply, the pair will occur with greater frequency than others will. Therefore, we can enumerate the relatedness by measuring similarities (lexical cohesion) between nouns and verbs using Dice's coefficient.

$$relatedness(v_i, n_j) = \frac{2 \times Occ(v_i \cap n_j)}{Occ(v_i) + Occ(n_j)} \quad (1)$$

where, v_i and n_j represent a verb and a noun, respectively, and *Occ* is the occurrence of the token. To illustrate, we provide a sample measurement of relatedness in table 2.

Table 2. Examples of similarity measurement for the noun 'Book'

v_i	$Occ(v_i)$	n_j	$Occ(n_j)$	$Occ(v_i \cap n_j)$	relatedness
Read	317,579,294	Book	541,966,689	5,216,567	0.01214
Search	721,112,089			4,532,558	0.00718
Print	177,593,953			1,921,587	0.00534
Write	170,529,264			810,037	0.00227

Table 2 shows examples of calculating the relatedness of the verbs 'read,' 'search,' 'print,' and 'write' to the noun 'book.' From these results, we can conclude that people generally intend to read when they access or touch a book. Based on the full set of such measurements, we constructed a

relatedness matrix between 2,727 verbs and 755,312 nouns (the resulting file size exceeds 4 GB).

4. Evaluation

User intention modeling attempts to identify the activities in which a user might engage by analyzing the nearby objects in which he/she shows interest. To accomplish this, our research focuses on modeling user intention based on the relatedness of nearby nouns a set of verbs. This section provides some experimental data showing how reliably our system selects verbs appropriate to given nouns.

Figure 5 shows the output of an application we designed to evaluate reliability. The application receives one or more nouns as an input and returns the top 20 verbs according to relatedness. In the case of multiple nouns, the application calculates relatedness for each verb. The output verbs are then evaluated by their appropriateness to the input noun. Table 3 shows a sample evaluation of the top 20 verbs related to the single noun 'door' extracted from our system.

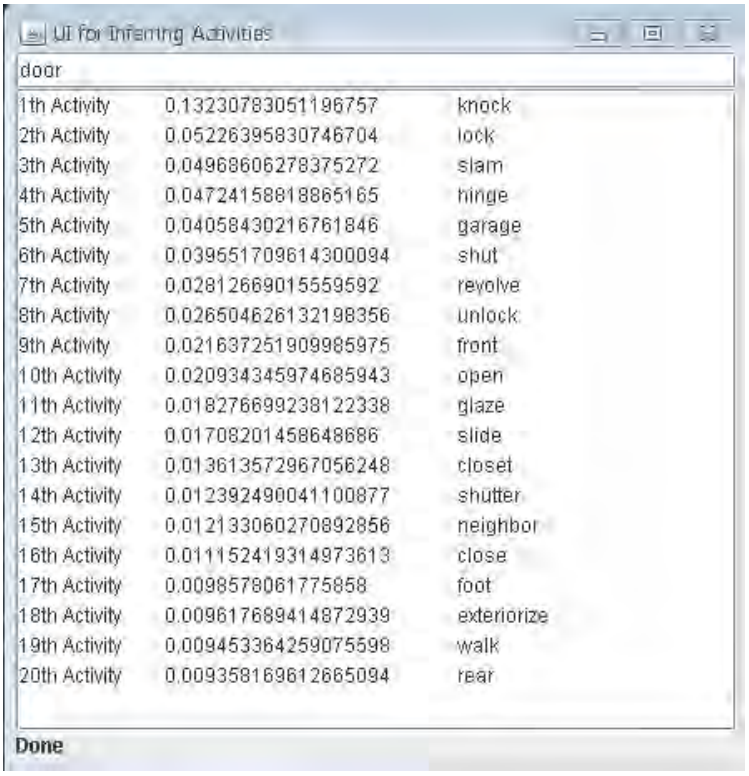


Fig. 5. Application output of the top 20 activities for the object 'door', according to the proposed system.

Table 3. Top 20 verbs for 'door' and their appropriateness (1 means appropriate and 0 means inappropriate).

Verbs	Eval.	Verbs	Eval.	Verbs	Eval.
Knock	1	Unlock	1	Neighbor	0
Lock	1	Front	0	Close	1
Slam	1	Open	1	Foot	0
Hinge	1	Glaze	0	Exteriorize	0
Garage	0	Slide	1	Walk	0
Shut	1	Closet	0	Rear	0
Revolve	0	shutter	0		

To evaluate our system for single noun inputs, we chose 200 nouns that represent tangible and intangible objects familiar in daily life. To evaluate our system for multiple noun inputs, we chose an additional 100 objects. Multiple noun input sets were created by combining a noun from our set of single noun inputs (e.g., 'computer java' and 'book bookshelf') with specific statuses or compound nouns (e.g., 'copy machine,' 'opened window' and 'table' + 'copy machine'). Table 4 provides a few examples.

Table 4. Examples of object selection

Count of clues	Examples of clues
Single object	Computer, printer, bookshelf, water, scanner, server, door, coffee, money, java, …
Multiple objects	Computer java, opened door, copy machine, table printer, coffee water, apache server, received email, … table copy machine,

To calculate the reliability of our verb output sets, we use equation (2):

$$reliability = \frac{cnt(appr.)}{cnt(appr.) + cnt(inappr.)} \quad (2)$$

Table 5 provides the results of our evaluation:

Table 5. Evaluation results on data reliability

	Evaluation for single nouns	Evaluation for multiple nouns
Appropriate	1,888	1,284
Inappropriate	2,112	716
Reliability (%)	47.2	64.2

From these results, it is clear that multiple noun inputs yield higher reliability than do single noun inputs by a difference of 17%. The reason for this is that single nouns have more ambiguity. For example, the noun 'window' has at least two possible meanings: (1) a tangible, transparent opening in a wall or

door and (2) an intangible object offered by graphical user interfaces. Even if we use the branded, capitalized term 'Windows' to indicate the latter of these two meanings, the stemmer used by our system converts the term to the basic and ambiguous form 'window'. Another example is the term 'java', which may refer to coffee or to a popular programming language.

Such ambiguities have a negative impact on reliability and overall results for single noun inputs are somewhat discouraging. However, we believe the system can be significantly improved by limiting the domain of activities based on location. For example, if someone touches a car key in an office, his/her immediate intention is probably to leave the office, but if he/she touches a car key inside a parking garage, his/her intention is probably to unlock a car and drive away. Such place-bound improvements have already been demonstrated in previous work [7]. We believe that this kind of domain restriction can help to construct a far more useful relatedness matrix between our 2,727 verbs (main activities) and 755,312 world objects. We intend to make these improvements in the future and to base further applications on a more refined relatedness matrix.

5. Conclusion

Inferring human activities is essential to the future of UPC and supporting a well-structured user intention model is the natural starting point to make such inference possible and reliable. To this end, we have leveraged the collective intelligence represented in WordNet and Google 5-gram to find and measure similarities between objects and activities. To test the reliability of our inferences, we created an application and used 200 single objects and 100 multiple objects as test inputs. Even though the overall evaluation was somewhat not satisfied to our expectation, we were able to identify clear paths for improvement using place-bound restrictions of activity sets. We expect that the result is useful for preparing ground data for actual user intention modeling in order to choose wide-ranging and reliable object-activity pairs.

References

1. Dey, A. K. and Abowd, G. D.: Towards a Better Understanding of Context and Context Awareness. In Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing, pp. 304-307, 1999.
2. Gu, T., Pung, H. K. and Zhang, D. Q.: Toward and OSGi-Based Infrastructure for Context-Aware Applications. IEEE Pervasive Computing, Vol. 3(4), pp. 66-74, 2004.
3. Huang, P. C. and Kuo, Y. H.: A Reliable Context Model for Context-aware Application. Systems Man and Cybernetics, pp. 246-250, 2008.
4. Gui, F., Guillen, M., Rishe, N., Barreto, A., Andrian, J. and Adjouadi, M.: A Client-Server Architecture for Context-Aware Search Application. Network-Based Information Systems, pp. 539-546, 2009.

5. Gu, T., Pung, H. K. and Zhang, D. Q.: A Middleware for Building Context-Aware Mobile Services. Vehicular Technology Conference, Vol. 5, pp. 2656-2660, 2004.
6. Zhan, Y., Wang, S., Zhao, Z., Chen, C. and Ma, J.: A Mobile Device Oriented Framework for Context Information Management. Information, Computing and Telecommunication, pp. 150-153, 2009.
7. Hwang, M., Choi, D. and Kim, P.: Information Retrieval Techniques to Grasp User Intention in Pervasive Computing Environment. In Proceedings of International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 186-191, 2011.
8. Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W. Y. and Wenyin, L.: User Intention Modeling in Web Application Using Data Mining. World Wide Web: Internet and Web Information Systems, Vol. 5, No. 3, pp. 181-191, 2002.
9. Jansen, B. J., Booth, D. L. and Spink, A.: Determining the User Intent of Web Search Engine Queries. In Proceedings of the 16th international conference on World Wide Web, pp. 1149-1150, 2007.
10. Jung, S., Lee, C., Kim, K. and Lee, G. G.: Hybrid approach to user intention modeling for dialog simulation. In Proceedings of the ACL-IJCNLP Conference Short Papers, pp. 17-20, 2009.
11. Jeon, H., Kim, T. and Choi, J.: Ontology-based User Intention Recognition for Proactive Planning of Intelligent Robot Behavior. In Proceedings of International Conference on Multimedia and Ubiquitous Engineering, pp. 244-248, 2008.
12. Fellbaum, C.: WordNet: An Electronic Lexical Database, MIT Press.
13. Hwang, M., Choi, C. and Kim, P.: Automatic Enrichment Method of Semantic Relation Network and its Application to Word Sense Disambiguation. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 6, pp. 845-858, 2011.
14. Brants, T. and Franz, A.: Web 1T 5-gram Corpus Version 1.1 (LDC2006T13), April 2006.
15. Hwang, M., Jeong, D. H., Kim, J., Song, S. K. and Jung, H.: Similarity Measurement between Objects and Activities for User Intention Modeling. In Proceedings of 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 103-107, 2012.
16. Jung, J. J.: Contextualized mobile recommendation service based on interactive social network discovered from mobile users. Expert Systems with Applications, Vol. 36, No.9, pp. 11950-11956, 2009.
17. Augusto, J. C. and J. M. Paul.: Ambient Intelligence: Concepts and applications. Computer Science and Information Systems, Vol. 4, No. 1, pp. 1-27, 2007.
18. Jung, J. J.: Boosting Social Collaborations Based on Contextual Synchronization: An Empirical Study. Expert Systems with Applications, Vol. 38, No. 5, pp. 4809-4815, 2011.
19. Eagle, N. and Pentland, A.: Reality mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing, Vol. 10, No. 4, pp. 255-268, 2006.

Myunggwon Hwang received the B.S. degree in Computer Engineering, the M.S. degree in Computer Science, and the Ph.D. degree in Computer Engineering from Chosun University. He is a senior researcher in the Department of Computer Intelligence Research at Korea Institute of Science and Technology Information (KISTI). His research focuses on Semantic Web Technologies, Semantic Information Processing and Retrieval, Word Sense Disambiguation and Knowledge Acquisition.

Do-Heon Jeong received the M.S. in Information Science from Yonsei University in 2003. He is in charge of analytics service development team in the Department of Computer Intelligence Research, Korea Institute of Science and Technology Information (KISTI). His research focuses on Informetrics, Semantic Web and Text Mining.

Jinhyung Kim received the M.S. degree and the Ph.D. degree in Computer Science from Korea University respectively. He is a postdoctoral researcher in the Department of Computer Intelligence Research at the Korea Institute of Science and Technology Information (KISTI). His research focuses on Semantic Web Technologies, Semantic Information Processing and Retrieval, Word Sense Disambiguation and Knowledge Acquisition.

Sa-kwang Song received his B.S. degree in Statistics in 1997 and his M.S. degree in Computer Science in 1999 at Chungnam National University, Korea. He received his Ph.D. degree in Computer Science at Korea Advanced Institute of Science and Technology (KAIST). He is currently a senior researcher in the Department of Computer Intelligence Research, Korea Institute of Science and Technology Information (KISTI). His research interest includes Text Mining, Natural Language Processing, Information Retrieval, Semantic Web, and Health-IT.

Hanmin Jung works as the head of the Department of Computer Intelligence Research and chief researcher at Korea Institute of Science and Technology Information (KISTI), Korea since 2004. He received his M.S. and Ph.D. degrees in Computer Science and Engineering from POSTECH, Korea in 1994 and 2003. His current research interests include technology intelligence based in the Semantic Web and text mining technologies, human-computer interaction (HCI), and natural language processing (NLP).

Received: September 01, 2012; Accepted: March 25, 2013

DOI: 10.2298/CSIS121110034C

Cognitive RBAC in Mobile Heterogeneous Networks

Hsing-Chung Chen*,[1], Marsha Anjanette Violetta[2], Chien-Erh Weng[3], and Tzu-Liang Kung[4]
[1]Department of Computer Science and Information Engineering, Asia University, Taichung City 41354, Taiwan Members, IEEE & IET
cdma2000@asia.edu.tw also shin8409@ms6.hinet.net
[2]Institute of Computer Science and Information Engineering, Asia University, Taichung City 41354, Taiwan
marsha.violette@gmail.com
[3]Department of Electronic Communication Engineering, National Kaohsiung Marine University, Kaohsiung City 81157, Taiwan
ceweng@mail.nkmu.edu.tw
[4]Department of Computer Science and Information Engineering, Asia University, Taichung City 41354, Taiwan
tlkung@asia.edu.tw

Abstract. In communication networks, a cognitive network (CN) is a new type of data network which is used to solve some of the problems that face current networks. Cognitive radio (CR) is part of a cognitive network and a smart wireless communication system. CR is conscious of its surrounding environment, and learns from the environment. It adapts its internal states by making corresponding real-time changes in certain operating parameters. In this paper, we propose a novel Cognitive RBAC (Role-Based Access Control) scheme which can be applied to Mobile Heterogeneous Networks (MHNs). The MHNs consist of mobile communication systems and Wi-Fi systems. The required new definitions for the RBAC model are proposed in this paper. They can improve the ability of conventional RBAC model to meet new challenges. In our scheme, we assume that a Cognitive Server (CS) provides and manages the permissions of services, and Network Providers support and manage a variety CRs and CNs, individually. For more efficiently managing CR and CN and meeting the large scale heterogeneous networks, we let mobile user can perceive network candidate actively to access services, in which the permissions are depending to the contract made by CS with each Network Provider. In this paper, the new generalized cognitive RBAC model and their definitions are proposed, and could be applied to new applications in a MHNs environment.

Keywords: cognitive radio, cognitive networks, role based access control, heterogeneous networks, mobile communication system, contract RBAC.

1. Introduction

A cognitive radio (CR) is a transceiver that automatically changes its transmission or reception parameters in such a way so as to allow opportunistic selection of available wireless channels. The main process is also known as dynamic spectrum management. A cognitive radio, as defined by the researchers at Virginia Polytechnic Institute and State University, is a software defined radio with a cognitive engine brain. The Cognitive Server (CS) [20] and the Network Provider(s) (Called as Operator(s), hereafter) are becoming a more and more popular means for a mobile user's local machine to access content from the databases, e.g. some databases of agents [22, 24], even those belonging to heterogeneous networks [1, 2, 5, 6]. The two cognitive processes derived from the CS and Operator(s) are both associated with perceiving the current access network conditions that are learned from consequences of end user actions. One process is the Cognitive Network (CN), and the other is the CR. One of the core techniques in a CS and Operator(s) are Access Control (AC), which is the means by which the availability of data and resources accessible by users in a system is restricted and which both defends against illegal access by malicious attackers and prevents honest users from gaining inappropriate access and possibly causing administrative errors. A new AC technique, Role-based Access Control (RBAC) [1, 5] has established itself as a generalized approach for handling access control in large organizations. It differs from conventional identity-based access control models in that it takes advantage of simplifying access control policies by using the concept of role relations [1, 5]. Based on the aforementioned [1, 5], the RBAC model's manners of constraining users' access to computer systems and the maturity of its models have been widely investigated [20].

The future trend is to access multiple large scale heterogeneous networks (LHNs), e.g., GPRS/3G-GPRS, WiMax, Wi-Fi [7, 23], on the same time. The new challenges will often arise when users' smart devices, such as cellular phones and personal digital assistants (PDAs) are suddenly handover to another access network in a Mobile Heterogeneous Networks (MHNs), which is a kind of LHNs, on while accessing the same resources. It is thus important to decide the best network that can fulfill user requirements based on QoS and individual consideration. Wireless cognitive technologies are used to adaptively select a better access network and the related wireless access radio channel for visiting various resources. Therefore a better quality of

*Correspondence: Hsing-Chung Chen (Jack Chen), Department of Computer Science and Information Engineering, Asia University, Taichung Country, Taiwan 41354, Fax, +886-4-23305737, E-Mails: shin8409@ms6.hinet.net also cdma2000@asia.edu.tw.

accessing services will be obtained by adaptively selecting wireless access radio channels, and also improving access capability.

However, conventional RBAC model [1, 5] does not address the model of CN and CR via the cognitive processes, in which current access network situations are perceived and learned from the consequences of end users' actions. Therefore, in order to cope with the CN and CR on a conventional RBAC model, we propose a novel Cognitive RBAC model to meet the new management requirements in MHNs. In this paper, we assume that the services of a CS also support the RBAC mechanism, which can assign and manage the access privileges depending on the user's corresponding permissions [6]. In this model, the CS and Operator(s) are assumed to have the contract that will enable their system to detect to which authenticated user the device belongs to. The Operator(s) will then allocate the specified access network as well as available access radio channel to role(s) corresponding to pre-registered user(s). Subsequently, dynamic cognitive access network environments will be provided that are necessary for the required Quality of Service (QoS).

The rest of this paper is organized as follows: the related works of CR, CN and RBAC are addressed in Section 2. In Section 3 we review the Cognitive RBAC model in a Small Heterogeneous Networks (SHNs) proposed by Hsing-Chung Chen and Marsha Anjanette Violetta. We then propose a new generalized Cognitive RBAC for MHNs. The new definitions for CS and Operator(s) are described in Section 4. There are three application scenarios in Section 5. The comparisons are provided in Section 6. Finally, we make our conclusions in Section 7.

2. Related Works

2.1 Cognitive Radio

With the growing number of wireless devices and increased spectrum occupancy, the unlicensed spectrum is getting increasingly. Additionally, large portions of the licensed spectrum, even in urban areas, are underutilized. To address the potential spectrum exhaustion problem, new wireless communication paradigms have been proposed for future wireless communication devices. The CR [3, 11, 13, 20, 21] concept is a new wireless communication approach that improves spectrum usage efficiency by exploiting the existence of spectrum holes [20].

The concept of cognitive radio was first proposed by Mitola *et al.* in a seminar at the Royal Institute of Technology in Stockholm, in 1998 [11]. The research article, which was published by Mitola *et al.* in 1999 [11], describes a novel approach to wireless communications in which: "*The point in which wireless smart devices and the related networks are sufficiently computationally intelligent about radio resources and related computer-to-*

computer communications to detect user communications needs as a function of use context, and to provide radio resources and wireless services most appropriate to those needs" [13]. CR is considered to be an ideal goal towards which a software-defined radio platform should evolve: *i.e.* a fully reconfigurable wireless transceiver that automatically adapts its communication parameters to network and user demands [11, 13, 20].

CR is the wireless communication technology having cognitive and reconfigurable properties and the capability to detect unoccupied spectrum holes and change frequency, thus enhancing computer-to-computer communications. In most of the existing proposals there are three steps for the basic functionality of CR. Observing and sensing is the first step of the cognitive process. The next step is to identify and analyze the spectrum. The last step is that of sharing the spectrum information and executing spectrum assignments [12, 20].

2.2 Cognitive Network

In communication networks, a CN is a new type of data network that makes use of cutting edge technology from several research areas; *i.e.* machine learning, knowledge representation, computer network, and network management, to solve some problems current networks are faced with. CN is different from CR as it covers all the layers of the OSI model, not only layers 1 and 2, as with CR [14, 20].

One of the attempts to define the concept of a cognitive network was made in 2005 by Thomas, Da Silva and Mac Kenzie [15] and is based on the older idea of the Knowledge Plane described by Clark *et al* [16]. Since then, several research activities in the area have emerged. A survey [17] and an edited book [18] reveal some of these efforts.

The Knowledge Plane is "a pervasive system within the network that builds and maintains high level models of what the network is supposed to do, in order to provide services and advice to other elements of the network" [16]. The concept of a large scale cognitive network was made in 2008 by Song [19], where such a Knowledge Plane was clearly defined for large scale wireless networks as the information about the availability of radio spectrum and wireless stations [20].

In [15], the authors define the CN as a network with a cognitive process that can perceive current network conditions, plan, decide, act on those conditions, learn from the consequences of its actions, all while following end-to-end goals. This loop, the cognition loop, senses the environment, plans actions according to input from sensors and network policies, decides which scenario fits best its end-to-end purpose using a reasoning engine, and finally acts on the chosen scenario as discussed in the previous section. The system learns from the past (situations, plans, decisions, actions) and uses this knowledge to improve decisions in the future [12].

CRs have been mentioned in previous papers [3, 11, 13, 20]. Mitola *et al.* [11] makes brief mention of how his cognitive radios could interact within the

system-level scope of a cognitive network. CNs are the future of information technology, in which the movement of network intelligence from controlling resources to understanding users' needs will help to manage the networks by facing out towards further network intelligences. CNs are with respect to future, one of the core technologies of mobile IP networks, and it is probable that the context sensitivity of these networks could have an interesting application in the field as cognitive radios. A CN should provide, over an extended period of time, better end-to-end performance than a non-cognitive network. Cognition could be used to improve resource management, Quality of Service (QoS), security, access control, and many other network goals [10, 20].

2.3 Role Based Access Control

RBAC is a well-known method for easily managing users to access resources via her/his authorized role. The main function of RBAC is to prevent unauthorized user from gaining information to which they are not entitled. Access rights are grouped by role name, and the use of resources is restricted to individuals authorized to assume the associated role. The use of roles to control access can be an effective means for developing and enforcing enterprise-specific security policies, and for streamlining the security management process.

Basic RBAC [1, 5] model is defined in terms of four model components: Core RBAC, Hierarchical RBAC, Static Separation of Duty Relations, and Dynamic Separation of Duty Relations. This basic model includes user-role assignment and permission-role assignment relations. Core RBAC includes sets of five basic data elements such as Users (U), Roles (R), Objects (OBJ), Operations (OPT) and Permissions ($PRMS$). Users are considered as human being, machines, networks, or intelligent agents that can perform some activities. Roles are described as a set of permissions necessary to access the resources. Permissions are approvals to execute operations on one or more objects. Operations are some executions for a program or a specific function which is invocated by a user. Objects are the entities that contain or receive information, or have exhaustible system resources. Furthermore, Core RBAC introduces the concept of role activation as part of a user's session within a computer system [1, 5].

3. Cognitive RBAC in SHNs

The Cognitive RBAC in SHNs was first proposed by Hsing-Chung Chen and Marsha Anjanette Violetta [20]. In their proposed scheme, it is assumed that a smart mobile device can take on cognitive characteristics, and be able to integrate various types of access networks in the SHNs. The intelligent capabilities [4] for the SHNs are located in the CS [20]. Assigned networks

and the available channels are two important aspects. For the SHNs, assigned networks and available channels are integrated as access resources. There were three phases proposed in [20] for a user using her/his device(s) to register and access a CS: the user registration is illustrated in Fig. 1; the device registration is illustrated in Fig. 2; the access phase is illustrated in Fig. 3.

The proposed scheme addressed the new basic concept for a Cognitive RBAC model [20], which consists of the following component sets: Users (U), Roles (R), Permissions ($PRMS$), Sessions (S), Devices ($D\upsilon$), Channels (CH), and Networks ($N\varepsilon$), representing the set of users, roles, permissions, sessions, devices, channels, and networks set. Users are a set U considered as authenticated users who can establish (wireless) communication with the resources of the CS to perform some activities. Roles are described as a set R of permissions to access the resources of the CS. Permissions are a set $PRMS$ of approvals to execute operations on one or more objects of the CS. Sessions are a set S of the mappings between networks $N\varepsilon$ and an activated subset of the set of roles R. Devices are a set $D\upsilon$ considered as the mobile units used by assigned users to activate the roles and access permissions. Channels are a set CH considered as conveyers of the information signals from senders to receivers for the operation. Networks are a set $N\varepsilon$ considered as computers interconnected by communication channels that allow sharing of resources and information [12].

It is assumed that the CS can identify the device whether it is registered or not, and determine who is the device owner. The CS also assigns and manages all of the access networks and available channels for each role, and dynamically adapts the access networks and available radio channels, depending on their environment, as needed for application performance. The generalized model of cognitive RBAC is described in *Definition 1* [20].

Definition 1: *The generalized model of Cognitive RBAC;*

- *Users (U), Roles (R), Permissions ($PRMS$), Sessions (S), Devices ($D\upsilon$), Channels (CH), and Networks ($N\varepsilon$), representing the set of users, roles, permissions, sessions, devices, channels, and networks set which are assigned by the CS, respectively;*
- $UA \subseteq U \times D\upsilon \times R$ *, the user assignment relation that associates users with their devices will be assigned the available roles after successful user and device authentication;*
- $r\_au(r \in R) \to 2^{U \times D\upsilon}$ *, the mapping of a role* r *onto a power set of authenticated users with their devices where function* $r\_au(\bullet)$ *is defined as* $r\_au(r \in R) = \{(u, d\upsilon) \in U \times D\upsilon \,|\, (u, d\upsilon, r) \in UA\}$*;*
- $N\varepsilon_{\psi'}$*, a set of networks, where* ψ' *is a CS system;*
- $CH_{\eta\varepsilon}$ *is a the set of channels and corresponding to a network* $\eta\varepsilon$*;*
- $ch \in CH_{\eta\varepsilon}$*,* $CH_{\eta\varepsilon} \in N\varepsilon_{\psi'}$*, is an available channels in a channel set which belongs to a set of networks;*

- $d\upsilon \subseteq D\upsilon$, $CHA_{\eta\varepsilon} \subseteq CH_{\eta\varepsilon} \times D\upsilon$ *the channel assignment relation that the available channels,* $CH_{\eta\varepsilon}$*, assigned to a smart device,* $d\upsilon$*, via a network* $\eta\varepsilon$ *which is managed by a CS;*
- $PA \subseteq R \times N\varepsilon_{\psi} \times CH_{\eta\varepsilon} \times PRMS$ *, the role assignment relation that assigns permission to an available role, network and channel;*
- $r\_p\left(r \in R, \eta\varepsilon \in N\varepsilon_{\psi}, ch \in CH_{\eta\varepsilon}\right) \to 2^{PRMS}$ *, the mapping of a role* r*, a network* $\eta\varepsilon$ *and a channel* ch *onto a power set of permissions where the function* $r\_p(\bullet)$ *is defined as* $r\_p(r, \eta\varepsilon, ch) = \{p \in PRMS | (r, \eta\varepsilon, ch, p) \in PA\}$*;*
- $r\_n(r \in R) \to 2^{N\varepsilon_{\psi}}$ *is the mapping of a role onto a power set of networks;*
- $ch\_r\left(ch \in CH_{\eta\varepsilon}\right) \to 2^{R}$ *is an assigning function for available channels in a network onto a power set of roles;*
- $u\_s((u, d\upsilon) \in U \times D\upsilon) \to 2^{S}$ *is the mapping of a user with his a device,* $(u, d\upsilon)$*, onto a power set of sessions;*
- $s\_r(\varsigma \in S) \to 2^{R}$ *is the mapping of a session to a power set of roles;*
- $avail\_s\_p\left(\varsigma \in S, , \eta\varepsilon \in N\varepsilon_{\psi}, ch \in CH_{\eta\varepsilon}\right) \to 2^{PRMS}$ *is the mapping of a power set of available permissions from a network* $\eta\varepsilon$ *and a channel* ch *in a session* ς *; the user finally gets the permissions,* $\bigcup_{r \in s\_r(\varsigma)} r\_p(r, \eta\varepsilon, ch)$.

□

Hierarchies in the Cognitive RBAC model are defined as an inheritance relationship between two roles managed by the CS, such that a role, $r_i \in R$, inherits the permissions from role, $r_j \in R$, if all permissions of r_j are also the permissions of r_i. We present a hierarchical Cognitive RBAC model for the CS in *Definition 2* [20]. In this model, permissions are assigned to a role.

Definition 2: *Role hierarchies in a Cognitive RBAC model;*

- $RH \subseteq R \times R \times N\varepsilon_{\psi} \times CH_{\eta\varepsilon}$ *is a partial order of roles, called the ascendancy relation combined with networks and channels , written as '*$\succeq$*' , where* $r_i \succeq r_j$*, is such that role,* $r_i \in R$*, inherits all permission,* $r_j \in R$ *, are assigned to all the users of* r_i *which are also assigned to all the users of* r_j*;*
- $r\_p\left(r_i \in R, \eta\varepsilon_m \in N\varepsilon_{\psi}, ch_c \in CH_{\eta\varepsilon}\right) \to 2^{PRMS}$ *is the mapping of a role,* r_i *, a network* $\eta\varepsilon_n$ *and a channel* ch_c *onto a power set of permissions. The permission set is assigned directly together with the permissions which are assigned to its successive roles, specifically:* $r\_p(r_i, \eta\varepsilon_m, ch_c) = r\_p(r_i) \cup \left\{ \cup_{\forall r_j \,:\, r_i \succeq r_j} r\_p\left(r_j, \eta\varepsilon_n, ch_e\right) \right\}$*;*

- $r\_au(r_i \in R) \to 2^{U \times D\upsilon}$ *is the mapping of a role ,* r_i *, onto a power set of authenticated users with their devices in the presence of a role hierarchy, specifically:* $r\_au(r_i) = \{(u, d\upsilon) \in U \times D\upsilon \,|\, r_i \succeq r_j, (u, d\upsilon, r_i) \in UA\}$;
- $r\_n(r_i \in R) \to 2^{N\varepsilon_{\psi'} \times Ch}$ *is the mapping of a role ,* r_i*, onto a power set of network together with channel. The set of network together with channel assigned directly to its successive roles, specifically:*
- *From the above sub-definitions, it follows that if* $r\_p(r_j, \eta\varepsilon_n, ch_e) \subseteq r\_p(r_i, \eta\varepsilon_m, ch_c)$ *and* $r\_au(r_j) \subseteq r\_au(r_i)$.

□

Separations of Duties are defined in *Definition 3* [20] and *Definition 4* [20] as those are to be enforced on a set of roles that may not be executed simultaneously by a user. Their model would be similar to the well-known RBAC model.

1) Static Separation of Duty (SSD) relations place constraints on the assignments of users to roles. Membership of one role may prevent the user from being a member of one or more other roles, depending on the SSD enforced rules.

Definition 3: *SSD relation in the Cognitive RBAC model;*

- SSD , $SSD \subseteq 2^R \times 2^{N\varepsilon_\psi} \times 2^{CH_{\eta\varepsilon}} \times N$ *is a collection of four* $(\alpha, \beta, \chi, n) \in \left(2^R, 2^{N\varepsilon_\psi}, 2^{CH_{\eta\varepsilon}}, N\right)$ *for CS where each* $\alpha \in 2^R$ *is a role set,* $\beta \in 2^{N\varepsilon_\psi}$ *is a network set,* $\chi \in 2^{CH_{\eta\varepsilon}}$ *is a channel set and* $n \in N$ *is a natural number,* $n \geq 2$ *with the property that no user can be assigned to* n *or more roles from the set* α *in any network* β *or channel* χ *. Specifically:* $\forall(\alpha, \beta, \chi, n) \in SSD, \forall \eta \in \alpha, \beta, \chi : |\eta| \geq n \Rightarrow \bigcap_{r \in \rho} r\_au(r) = \varnothing$.

2) Dynamic Separation of Duty (DSD) relations differ from SSD relations by the context in which these limitations are imposed. DSD requirements limit the availability of the permissions by placing constraints on the roles that can be activated within or across a user's sessions.

Definition 4: *DSD relation in the Cognitive RBAC model;*

- DSD , $DSD \subseteq 2^R \times 2^{N\varepsilon_\psi} \times 2^{CH_{\eta\varepsilon}} \times N$ *is a collection of four* $(\alpha, \beta, \chi, n) \in \left(2^R, 2^{N\varepsilon_\psi}, 2^{CH_{\eta\varepsilon}}, N\right)$ *for CS where each* $\alpha \in 2^R$ *is a role set,* $\beta \in 2^{N\varepsilon_\psi}$ *is a network set,* $\chi \in 2^{CH_{\eta\varepsilon}}$ *is a channel set and* $n \in N$ *is a natural number,* $n \geq 2$ *with the property that no user may activate* n *or more roles from the set* α *in any network* β *or channel* χ *. Specifically:*

$$\forall(\alpha,\beta,\chi,n)\in DSD,\forall\varsigma\in S,\forall\rho\subseteq s\_r(\varsigma)\cap\alpha,\beta,\chi:|\rho|\geq n\Rightarrow\bigcap_{r\in\rho}r\_au(r)=\varnothing$$

.

□

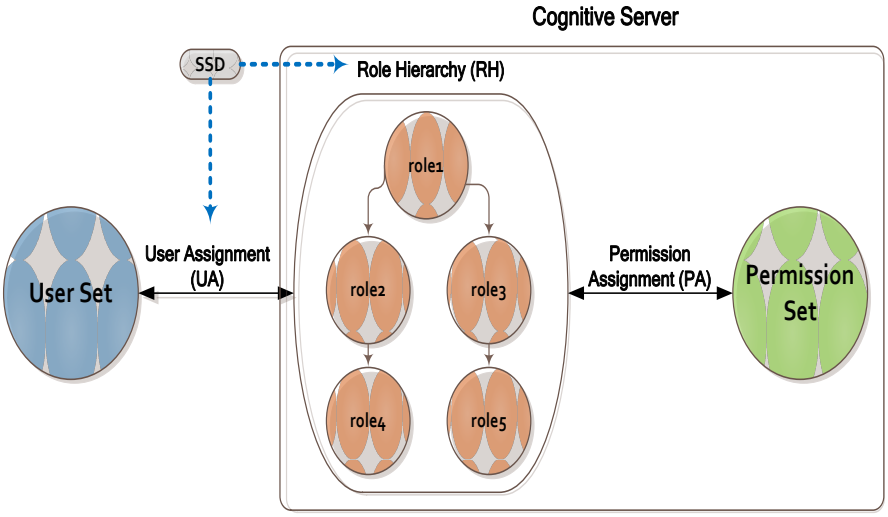


Fig. 1. User registration phase

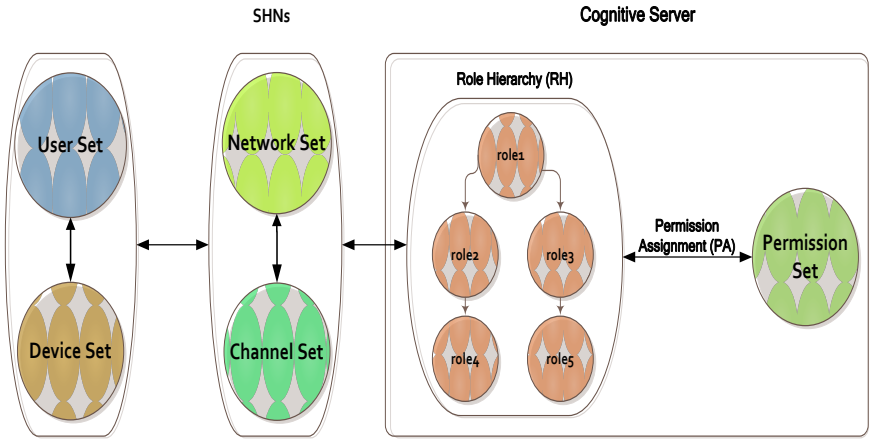


Fig. 2. Device registration phase

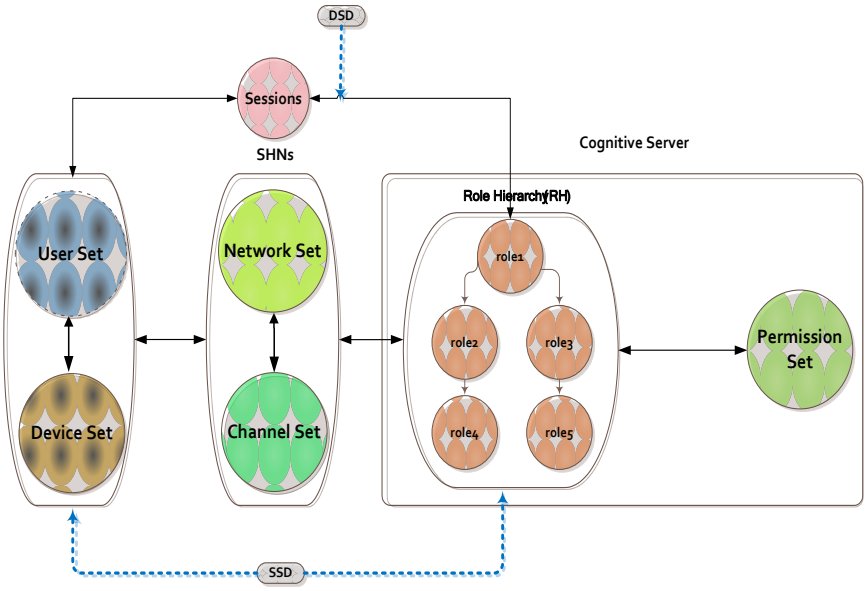


Fig. 3. Access phase

4. Generalized Cognitive RBAC model in MHNs

In order to deal with the CR and CN on a LHNs environment, we propose a novel Generalized Cognitive RBAC model to meet the new management requirements in MHNs. The MHNs will take on cognitive radio characteristic, and are able to integrate various types of access networks. This model can provide cognitive network characteristics such as QoS, better end-to-end performance, security, and access control. Registered devices, assigned networks, available channels and contract roles are four important aspects. In the MHNs, assigned networks and available channels are integrated and managed by Operator(s); permissions are assigned are managed by the CS for getting resources and services. We first propose a new role is 'contract role' which can achieve to flexibly manage the role mapping in MHNs. The contract role is depending on the contract made by the CS together with each Operator.

At first, we give the basic definition for the generalized Cognitive RBAC model combined with the contract concept as below.

4.1 The basic definition of the generalized Cognitive RBAC model in MHNs

The basic concept of the generalized Cognitive RBAC model consists of the following component sets: Users (U), Contract Roles (CR), Roles of Operator (Ro), Roles of CS (Rs), Permissions ($PRMS$), Sessions (S), Devices ($D\upsilon$), Channels (CH), Networks ($N\varepsilon$), and Contract ($CT^{(k)}$) representing the sets of users, contract roles, roles of operator, roles of CS, permissions, sessions, devices, channels, networks and contract. Users are a set of U considered as the authenticated users who can through the system of Operator, *e.g.* mobile communication system or Wi-Fi system, to access the resources of the CS. Roles of Operator are described as a set of roles, Ro, with the privileges to obtain the network(s) and channel(s) from the mobile communication or Wi-Fi system from an Operator. Roles of CS are described as a set of roles, Rs, with the permissions to access the resources of the CS.

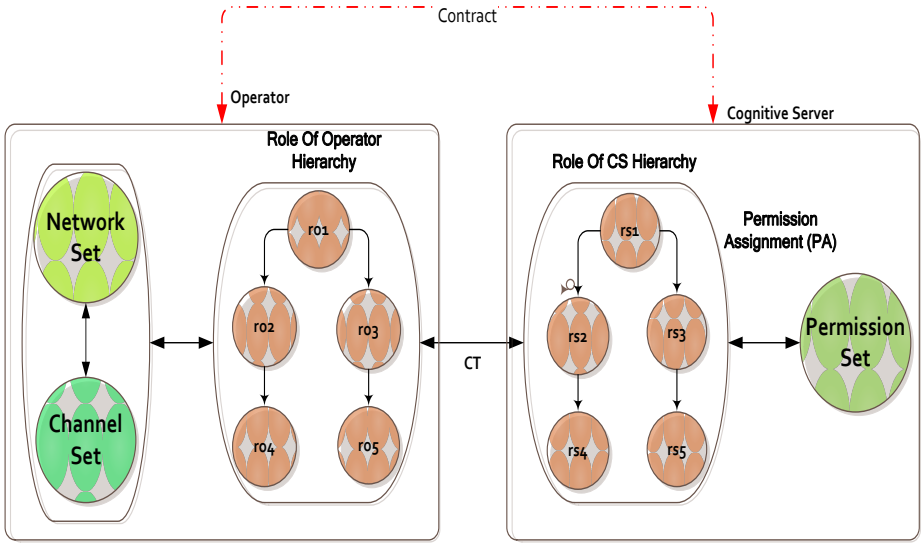


Fig. 4. Cognitive RBAC Model with the contract made by CS and one Operator in MHNs.

Contract Roles are combined the mapping roles from both Roles of CS and Roles of Operator, representing as a set of Contract Roles, CR, depending on the contract made by each Operator and the CS (See Fig. 4 and Fig. 5), individually. To access resources of CS through the assigned network(s) and the corresponding to radio channel(s) is managed by the Operator(s). Permissions, based on the contract roles set, CR, are a set of approvals to execute operations on one or more services of the CS. Sessions, based on the contract roles set, CR, are a set S of the mappings between networks, $N\varepsilon$, and an activated subset of the set of the contract roles set, CR. Devices

are a set of $D\upsilon$ which are considered as the mobile units used by assigning users to activate the contract roles and access the permissions constrained by the contract. Channels, based on the contract roles set, CR, is a set of CH considered as conveyers of the information signals from senders to receivers for the operation.

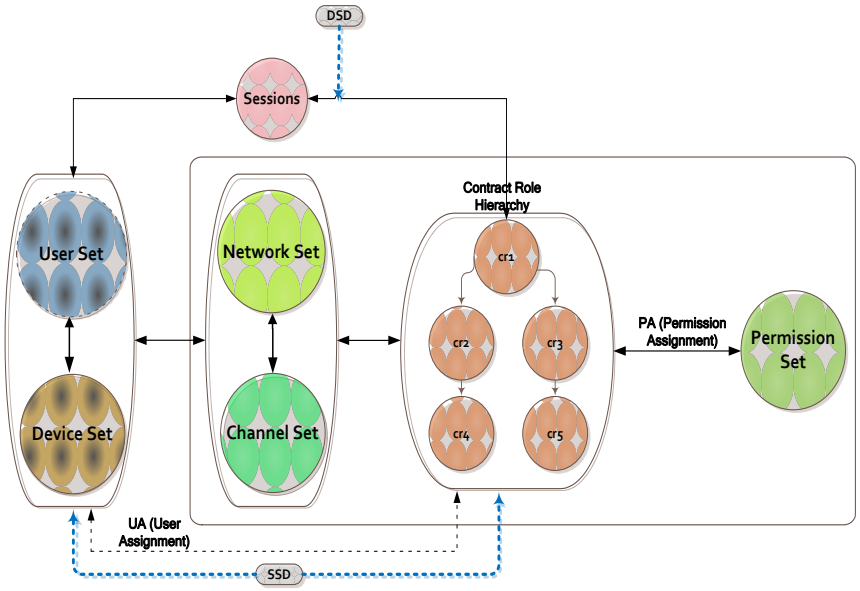


Fig. 5. Generalized Cognitive RBAC Model with combined roles which corresponding to the contract made by CS and one Operator in MHNs.

Networks, based on the contract roles set, CR, is a set of $N\varepsilon$ considered computers interconnected by communication channels that allow sharing of resources and information [12]. Contract is the agreement between CS and Operator(s) (See *Table 1*, *Table 2*, and *Table 3*). It is assumed that the CS and the Operator(s) can identify the device whether it is registered or not, and determine who is the device owner. Each Operator assigns and manages all of the access networks and available channels for each contract role, and dynamically adapts the access networks and available radio channels, depending on their environment and contract for getting better QoS. The CS manages the permissions which let mobile user can access resources from its server. The generalized model of Cognitive RBAC in MHNs is described in *Definition 5*.

Definition 5: *The generalized model of Cognitive RBAC in MHNs;*

- *Users* (U), *Contract Roles* (CR), *Roles of Operator* (Ro), *Roles of CS* (Rs), *Permissions* ($PRMS$), *Sessions* (S), *Devices* ($D\upsilon$), *Channels* (CH), *Networks* ($N\varepsilon$),*and Contract* ($CT^{(k)}$) *representing the set of users, contract roles, Roles of Operator, Roles of CS, permissions, sessions, devices, channels, networks and contract set which are corresponding the contract*

made by CS and one Operator k, respectively (See Fig. 4 and Fig. 5);

- $UA \subseteq U^{(k)} \times D\upsilon^{(k)} \times CR^{(k)}$ *the user assignment relation that associates users with registered devices to contract roles ;* $CR^{(k)} = Ro^{(k)} \times Rs^{(k)}$*;* k *is the name the name of serving Operator;*
- $r\_au^{(k)}\left(d\upsilon^{(k)} \in D\upsilon^{(k)}, cr^{(k)} \in CR^{(k)}\right) \to 2^{U^{(k)}}$*, the mapping of a device,* $d\upsilon^{(k)}$*, and a contract role,* $cr^{(k)}$*, onto a power set of authenticated users where function* $r\_au^{(k)}(\bullet)$ *is defined as* $r\_au^{(k)}\left(d\upsilon^{(k)}, cr^{(k)}\right) = \left\{u^{(k)} \in U^{(k)} \middle| \left(u^{(k)}, d\upsilon^{(k)}, cr^{(k)}\right) \in UA\right\}$*;*
- $PA \subseteq Rs^{(k)} \times PRMS^{(k)}$ *, the permission assignment relation that assigns permissions to available contract roles;*
- $r\_p^{(k)}\left(rs^{(k)} \in Rs^{(k)}\right) \to 2^{PRMS^{(k)}}$ *,the mapping of a role of CS,* $rs^{(k)}$ *, onto a power set of permissions where the function* $r\_p^{(k)}(\bullet)$ *is defined as* $r\_p^{(k)}\left(rs^{(k)}\right) = \left\{p^{(k)} \in PRMS^{(k)} \middle| \left(rs^{(k)}, p^{(k)}\right) \in PA\right\}$*;*
- $CRPA^{(k)} \subseteq CR^{(k)} \times PRMS^{(k)}$ *, the contract role permission assignment relation that assigns permissions to available contract roles;*
- $CRPRMS^{(k)} \subseteq 2^{PRMS^{(k)}}$ *is the permission set of contract role;*
- $CR^{(k)} \subseteq Ro^{(k)} \times Rs^{(k)}$ *, the contract role relation that associates Roles of Operator with Roles of CS to Contract Roles;*
- $r\_r^{(k)}\left(ro^{(k)} \in Ro^{(k)}, rs^{(k)} \in Rs^{(k)}\right) \to 2^{CR^{(k)}}$ *is the mapping of a role of operator and a role of CS onto a power set of contract roles, where function* $r\_r^{(k)}(\bullet)$ *is defined as* $r\_r^{(k)}\left(ro^{(k)}, rs^{(k)}\right) = \left\{cr^{(k)} \in CR^{(k)} \middle| \left(cr^{(k)}, ro^{(k)}, rs^{(k)}\right) \in CR^{(k)}\right\}$*;*
- $D\upsilon^{(k)}$*, the set of devices;*
- $\left\{d\upsilon^{(k)}\right\} \subseteq D\upsilon^{(k)}$*,* $CH_{\eta\varepsilon}{}^{(k)} \in d\upsilon^{(k)}$ *is the available channels which are belonging to a networks set for registered device;*
- $N\varepsilon_{\psi}{}^{(k)}$*, the set of networks;* ψ *is a system of Operator;* $\eta\varepsilon^{(k)} \in N\varepsilon_{\psi}{}^{(k)}$ *;* $m, w \in \psi$ *; m represented as a mobile communication system; and w represented as a Wi-Fi system;*
- $r\_n^{(k)}\left(ro^{(k)} \in Ro^{(k)}\right) \to 2^{N\varepsilon_{\psi}{}^{(k)}}$ *is the mapping of a role of operator onto a power set of networks;*
- $CH_{\eta\varepsilon}{}^{(k)}$ *, the set of channels;* $\eta\varepsilon$ *represented as a network;*
- $\left\{ch^{(k)}\right\} \subseteq CH_{\eta\varepsilon}{}^{(k)}$*,* $CH_{\eta\varepsilon}{}^{(k)} \in N\varepsilon_{\psi}{}^{(k)}$ *is the available channels in a channel set which belongs to a networks set;*
- $ch\_r^{(k)}\left(ch^{(k)} \in CH_{\eta\varepsilon}{}^{(k)}\right) \to 2^{Ro^{(k)}}$ *is an assigning function for available channels in a network onto a power set of roles of operator;*
- $u\_s^{(k)}\left(u^{(k)} \in U^{(k)}\right) \to 2^{S^{(k)}}$ *is the mapping of a user onto a power set of sessions;*
- $s\_r^{(k)}\left(\varsigma^{(k)} \in S^{(k)}\right) \to 2^{CR^{(k)}}$ *is the mapping of a session onto a power set of contract roles;*

- $avail\_s\_p^{(k)}\left(\varsigma^{(k)} \in S^{(k)}, \eta\varepsilon^{(k)} \in N\varepsilon_{\psi}{}^{(k)}, ch^{(k)} \in CH_{\eta\varepsilon}{}^{(k)}\right) \to 2^{SRPRMS^{(k)}}$ *is the mapping of a power set of available permissions from a network* $\eta\varepsilon^{(k)}$ *and a channel* $ch^{(k)}$*, in a session* $\varsigma^{(k)}$ *constrained by the assigned contract role* $\forall cr^{(k)} \in CR^{(k)}$ *; and the user gets her/his permissions as* $\bigcup_{cr^{(k)} \in s\_r^{(k)}(\varsigma^{(k)})} \left\{r\_p^{(k)}\left(cr^{(k)}, \eta\varepsilon^{(k)}, ch^{(k)}\right)\right\}$.

□

4.2 Hierarchical role in Generalized Cognitive RBAC in MHNs

Hierarchies in the Cognitive RBAC model are defined an inheritance relationship between the role of CS, $rs^{(k)}$, such that a role of CS, $rs_x{}^{(k)} \in Rs^{(k)}$, inherits the permissions from a role of CS, $rs_y{}^{(k)} \in Rs^{(k)}$, if all permissions of $rs_y{}^{(k)}$ are also the permissions of $rs_x{}^{(k)}$. We also define an inheritance relationship in Roles of Operator for the Operator(s), such that a role of operator, $ro_m{}^{(k)} \in Ro^{(k)}$, inherits all the networks and the channels from a role of operator $ro_n{}^{(k)} \in Ro^{(k)}$, if all networks and the channels of $ro_n{}^{(k)}$ are also the networks and the channels of $ro_m{}^{(k)}$. We present a hierarchical Cognitive RBAC model for the CS and the Operator(s) in *Definition 6*, such that a contract role, $cr_i^{(k)} \in CR^{(k)}$, inherits the permissions, access networks and available channels from a contract role, $cr_j{}^{(k)} \in CR^{(k)}$.

Definition 6: *Role hierarchies in a Cognitive RBAC model in MHNs;*

- $RH \subseteq CR^{(k)} \times CR^{(k)}$ *is a partial order of contract roles, called the ascendancy relation, written as "* $\succeq$ *" , where* $cr_i^{(k)} \succeq cr_j{}^{(k)}$*, means that a contract role,* $cr_i^{(k)} \in CR^{(k)}$*, inherits all the permissions from a contract role,* $cr_j{}^{(k)} \in CR^{(k)}$ *, and all the users who are assigned the role,* $cr_i^{(k)}$ *, also can access the permissions of* $cr_j{}^{(k)}$*;*
- $r\_p^{(k)}\left(rs^{(k)} \in Rs^{(k)}\right) \to 2^{PRMS^{(k)}}$ *is the mapping of a role of CS,* $rs_x{}^{(k)}$ *, onto a set of permissions. The permission set will be assigned directly to its successive Roles of CS,* $Rs^{(k)}$ *, specifically:* $r\_p^{(k)}\left(rs_x{}^{(k)}\right) = r\_p^{(k)}\left(rs_x{}^{(k)}\right) \cup \left\{ \bigcup_{\forall rs_y{}^{(k)} : rs_x{}^{(k)} \succeq rs_y{}^{(k)}} r\_p^{(k)}\left(rs_y{}^{(k)}\right) \right\}$*;*
- $r\_au^{(k)}\left(d\upsilon^{(k)} \in D\upsilon^{(k)}, cr^{(k)} \in CR^{(k)}\right) \to 2^{U^{(k)}}$ *the mapping of a device,* $d\upsilon^{(k)}$ *,*

and a contract role $cr^{(k)}$ *and onto a set of authenticated users in the presence of a contract role hierarchy, specifically:* $r\_au^{(k)}\left(d\upsilon^{(k)}, cr^{(k)}\right) = \left\{u^{(k)} \in U^{(k)} \middle| cr_i^{(k)} \succeq cr_j^{(k)}\left(u^{(k)}, d\upsilon^{(k)}, cr_i^{(k)}\right) \in UA\right\}$;

- $r\_n^{(k)}\left(ro^{(k)} \in Ro^{(k)}\right) \to 2^{N\varepsilon^{(k)}}$ *is the mapping of a role of operator,* $ro_m^{(k)}$*, onto a set of network. The networks set assigned directly together with the networks assigned to its successive Roles of Operator,* $Ro^{(k)}$*, specifically:* $r\_n^{(k)}\left(ro_m^{(k)}\right) = r\_n^{(k)}\left(ro_m^{(k)}\right) \cup \left\{ \bigcup_{\forall ro_n^{(k)} : ro_m^{(k)} \succeq ro_n^{(k)}} r\_n^{(k)}\left(ro_n^{(k)}\right) \right\}$;
- *From the above definitions, it follows that if* $r\_p^{(k)}\left(rs_y^{(k)}\right) \subseteq r\_p^{(k)}\left(rs_x^{(k)}\right)$, $r\_n^{(k)}\left(ro_n^{(k)}\right) \subseteq r\_n^{(k)}\left(rs_m^{(k)}\right)$ *and* $r\_au^{(k)}\left(d\upsilon^{(k)}, cr_j^{(k)}\right) \subseteq r\_au^{(k)}\left(d\upsilon^{(k)}, cr_i^{(k)}\right)$.

□4.3 Separation of duties constrained in the Cognitive RBAC in MHNs

Separations of Duties are defined in *Definition 7* and *Definition 8* as those are to be enforced on a set of contract roles that may not be executed simultaneously by a user. Our model would be similar to the well-known RBAC model, but we add a registered device in this concept.

1) SSD relations place constraints on the assignments of users and their devices to contract roles. Membership in one contract role may prevent the user with her/his devices from being a member of one or more other contract roles, depending on the SSD enforced rules for all networks.

Definition 7: *SSD relation in the Cognitive RBAC model in MHNs;*

- $SSD^{(k)}$, $SSD^{(k)} \subseteq 2^{CR^{(k)}} \times 2^{D\upsilon^{(k)}} \times N^{(k)}$ *is a collection of three* $(\alpha, \upsilon, n) \in \left(2^{CR^{(k)}}, 2^{D\upsilon^{(k)}}, N^{(k)}\right)$ *for CS and Operator(s) where each* $\alpha \in 2^{CR^{(k)}}$ *is a contract role set,* $\upsilon \in 2^{D\upsilon^{(k)}}$ *is a device set and* $n \in N^{(k)}$ *is a natural number,* $n \geq 2$ *with the property that no user can be assigned to* n *or more contract roles from the set* α *using any device from the set* υ*.*

  Specifically:

  $\forall (\alpha, \upsilon, n) \in SSD^{(k)}, \forall d\upsilon^{(k)} \in \upsilon, \forall \eta \subseteq \alpha : |\eta| \geq n \Rightarrow \bigcap_{cr^{(k)} \in \eta} r\_au^{(k)}\left(cr^{(k)}, d\upsilon^{(k)}\right) = \varnothing$.

2) DSD relations differ from SSD relations by the context in which these limitations are imposed. DSD requirements limit the availability of the permissions by placing constraints on the contract roles that can be activated within or across a user's sessions when contract roles can be activated by the user's devices through all networks and channels.

Definition 8: *DSD relation in the Cognitive RBAC model in MHNs;*

- $DSD^{(k)}$, $DSD^{(k)} \subseteq 2^{CR^{(k)}} \times 2^{D\upsilon^{(k)}} \times N^{(k)}$ *is a collection of three*

$(\alpha,\upsilon,n)\in\left(2^{CR^{(k)}},2^{D\upsilon^{(k)}},N^{(k)}\right)$ *for CS and Operator(s) where each* $\alpha\in 2^{CR^{(k)}}$ *is a contract role set,* $\upsilon\in 2^{D\upsilon^{(k)}}$ *is a device set and* $n\in N^{(k)}$ *is a natural number,* $n\geq 2$ *with the property that no user may activate* n *or more contract roles from the set* α *using any device from the set* υ *. Specifically:*

$$\forall(\alpha,\upsilon,n)\in DSD^{(k)},\forall d\upsilon^{(k)}\in\upsilon,\forall\varsigma^{(k)}\in S,\forall\rho\subseteq s\_r(\varsigma)\cap\alpha:|\rho|\geq n$$
$$\Rightarrow\bigcap_{cr^{(k)}\in\rho}r\_au^{(k)}\left(cr^{(k)},d\upsilon^{(k)}\right)=\varnothing.$$

□4.4. Contract and Registration Phase, and Access Phase

In this subsection, the phase for CS and the Operator(s) can be divided into two sub-phases: the first one is the *Contract and Registration Phase*, and the second phase is the *Access Phase*. In this subsection, we introduce the two phases for the generalized Cognitive RBAC model in MHNs.

4.4.1. Contract and Registration Phase

A contract is an agreement entered into voluntarily by two or more parties with the intention of creating a legal obligation, which may have some contract items in writing, though contracts can be made formally. Registration is having formally submitted a document to, and received approval for a specific activity from, the appropriate official or authority. In this phase, we will describe the contract and registration which is suitable for this application scenario.

1) **CS makes contract with Operator(s)**

In our assumption, CS is a kind of permissions provider for some specific applications, and Operator(s) should be an access network(s) provider which provides either mobile communication network or Wi-Fi network. The CS and Operator(s) will make the contracts that provide permissions to access the CS through the system of Operator. They will share information of registered users and registered devices. For examples, there are a CS server and three Operators: A, B, C, shown in *Table 1*, *Table 2*, and *Table 3*. The permissions are mapping from CS to networks and channels which belong to the distinct Operator.

Table 1. An example to illustrate the contract between CS and Operator(s) by $CT^{(A)}$

Role $CR^{(A)}$	Privileges				
	Operator A			Cognitive Server	
	Roles of Operator $Ro^{(A)}$	Network Set $N\varepsilon_{\psi}{}^{(A)}$	Channel Set $CH_{\eta\varepsilon}{}^{(A)}$	Roles of CS $Rs^{(A)}$	Permission $PRMS^{(A)}$
$cr_1^{(A)}$	$ro_1^{(A)}$	$m_1^{(A)}$ $m_2^{(A)}$ $w_3^{(A)}$	$\{ch_1, ch_2\} \subseteq CH_{m_1}{}^{(A)}$ $\{ch_1, ch_2\} \subseteq CH_{m_2}{}^{(A)}$ $\{ch_1, ch_3\} \subseteq CH_{w_3}{}^{(A)}$	$rs_1^{(A)}$	$prms_1^{(A)}$ $prms_2^{(A)}$ $prms_3^{(A)}$
$cr_2^{(A)}$	$ro_2^{(A)}$	$m_4^{(A)}$ $w_5^{(A)}$	$\{ch_2, ch_3\} \subseteq CH_{m_4}{}^{(A)}$ $\{ch_3, ch_5\} \subseteq CH_{w_5}{}^{(A)}$	$rs_2^{(A)}$	$prms_4^{(A)}$ $prms_5^{(A)}$
$cr_3^{(A)}$	$ro_3^{(A)}$	$m_6^{(A)}$ $w_7^{(A)}$	$\{ch_4\} \subseteq CH_{m_6}{}^{(A)}$ $\{ch_5, ch_6\} \subseteq CH_{w_7}{}^{(A)}$	$rs_3^{(A)}$	$prms_6^{(A)}$ $prms_7^{(A)}$
$cr_4^{(A)}$	$ro_4^{(A)}$	$w_8^{(A)}$ $m_9^{(A)}$	$\{ch_7, ch_8\} \subseteq CH_{w_8}{}^{(A)}$ $\{ch_8, ch_9\} \subseteq CH_{m_9}{}^{(A)}$	$rs_4^{(A)}$	$prms_8^{(A)}$ $prms_9^{(A)}$
$cr_5^{(A)}$	$ro_5^{(A)}$	$m_{10}^{(A)}$	$\{ch_9, ch_{10}\} \subseteq CH_{m_{10}}{}^{(A)}$	$rs_5^{(A)}$	$prms_{10}^{(A)}$

Table 2. An example to illustrate the contract between CS and Operator(s) by $CT^{(B)}$

Role $CR^{(B)}$	Privileges				
	Operator B			Cognitive Server	
	Roles of Operator $Ro^{(B)}$	Network Set $N\varepsilon_{\psi}{}^{(B)}$	Channel Set $CH_{\eta\varepsilon_{\psi}}{}^{(B)}$	Roles of CS $Rs^{(B)}$	Permission $PRMS^{(B)}$
$cr_1^{(B)}$	$ro_1^{(B)}$	$w_1^{(B)}$ $m_2^{(B)}$ $m_7^{(B)}$	$\{ch_1, ch_2\} \subseteq CH_{w_1}{}^{(B)}$ $\{ch_1, ch_2\} \subseteq CH_{m_2}{}^{(B)}$ $\{ch_1, ch_2\} \subseteq CH_{m_7}{}^{(B)}$	$rs_1^{(B)}$	$prms_1^{(B)}$ $prms_2^{(B)}$ $prms_3^{(B)}$
$cr_2^{(B)}$	$ro_2^{(B)}$	$w_3^{(B)}$ $m_4^{(B)}$	$\{ch_2, ch_3\} \subseteq CH_{w_3}{}^{(B)}$ $\{ch_3, ch_5\} \subseteq CH_{m_4}{}^{(B)}$	$rs_2^{(B)}$	$prms_4^{(B)}$ $prms_5^{(B)}$
$cr_3^{(B)}$	$ro_3^{(B)}$	$m_6^{(B)}$ $m_{10}^{(B)}$	$\{ch_1, ch_2\} \subseteq CH_{m_6}{}^{(B)}$ $\{ch_9, ch_{10}\} \subseteq CH_{m_{10}}{}^{(B)}$	$rs_3^{(B)}$	$prms_6^{(B)}$ $prms_7^{(B)}$
$cr_4^{(B)}$	$ro_4^{(B)}$	$m_8^{(B)}$ $m_9^{(B)}$	$\{ch_7, ch_8\} \subseteq CH_{m_8}{}^{(B)}$ $\{ch_8, ch_9\} \subseteq CH_{m_9}{}^{(B)}$	$rs_4^{(B)}$	$prms_8^{(B)}$ $prms_9^{(B)}$
$cr_5^{(B)}$	$ro_5^{(B)}$	$m_6^{(B)}$ $m_{10}^{(B)}$	$\{ch_1, ch_2\} \subseteq CH_{n_6}{}^{(B)}$ $\{ch_9, ch_{10}\} \subseteq CH_{m_{10}}{}^{(B)}$	$rs_5^{(B)}$	$prms_{10}^{(B)}$

Table 3. An example to illustrate the contract between CS and Operator(s) by $CT^{(C)}$

Role	Privileges				
	Operator C			Cognitive Server	
$CR^{(C)}$	Roles of Operator $Ro^{(C)}$	Network Set $N\varepsilon_{\psi}{}^{(C)}$	Channel Set $CH_{\eta\varepsilon_{\psi}}{}^{(C)}$	Roles of CS $Rs^{(C)}$	Permission $PRMS^{(C)}$
$cr_1^{(C)}$	$ro_1^{(C)}$	$m_1^{(C)}$ $m_2^{(C)}$ $w_7^{(C)}$	$\{ch_1, ch_2\} \subseteq CH_{m_1}{}^{(C)}$ $\{ch_1, ch_2\} \subseteq CH_{m_2}{}^{(C)}$ $\{ch_1, ch_2\} \subseteq CH_{w_7}{}^{(C)}$	$rs_1^{(C)}$	$prms_1^{(C)}$ $prms_2^{(C)}$ $prms_3^{(C)}$
$cr_2^{(C)}$	$ro_2^{(C)}$	$m_3^{(C)}$ $w_4^{(C)}$	$\{ch_2, ch_3\} \subseteq CH_{m_3}{}^{(C)}$ $\{ch_3, ch_5\} \subseteq CH_{w_4}{}^{(C)}$	$rs_3^{(C)}$	$prms_4^{(C)}$ $prms_5^{(C)}$
$cr_3^{(C)}$	$ro_3^{(C)}$	$m_6^{(C)}$ $w_{10}^{(C)}$	$\{ch_1, ch_2\} \subseteq CH_{m_6}{}^{(C)}$ $\{ch_9, ch_{10}\} \subseteq CH_{w_{10}}{}^{(C)}$	$rs_2^{(C)}$	$prms_6^{(C)}$ $prms_7^{(C)}$
$cr_4^{(C)}$	$ro_5^{(C)}$	$w_8^{(C)}$ $m_9^{(C)}$	$\{ch_7, ch_8\} \subseteq CH_{w_8}{}^{(C)}$ $\{ch_8, ch_9\} \subseteq CH_{m_9}{}^{(C)}$	$rs_4^{(C)}$	$prms_8^{(C)}$ $prms_9^{(C)}$
$cr_5^{(C)}$	$ro_4^{(C)}$	$m_6^{(C)}$ $w_{10}^{(C)}$	$\{ch_1, ch_2\} \subseteq CH_{m_6}{}^{(C)}$ $\{ch_9, ch_{10}\} \subseteq CH_{w_{10}}{}^{(C)}$	$rs_5^{(C)}$	$prms_{10}^{(C)}$

2) User and Device(s) Registration from the System of Operator

At first time, each user will be asked to go to the office centre of Operator, e.g. the Mobile Communication Company. Then, the contract will be signed by both the user and the Operator in order to finish the user registration via face to face. After the user gets his/her user ID, $uID_u^{(k)}$,she/he further enters the information of device(s), in an online system, in order to do the device registration(s) through the system of Operator, e.g. the mobile communication system or Wi-Fi system, which is depended on the signed contract. In the online system, the system of Operator(s) will check whether the user and her/his device(s) registration have been finished the registration or not according to *Algorithm 1*. If all the requirements of the user's contract are satisfied, the system of Operator will assign a contract role $cr_i^{(k)} \in CR^{(k)}$ to the user and her/his device(s), where the role will be mapped to (1) the role of operator consisting of the network(s) and available channel(s), (2) the role of CS consisting of permissions. We show an example to illustrate how users and their devices are mapped to the contract roles, for examples in *Table 4*.

Algorithm 1: User and Device(s) Registration Phase

Input: *user ID,* $uID_u^{(k)}$*; device information;*

Output: *contract role* $cr_i^{(k)} \in CR^{(k)}$ *; a default network* $N\varepsilon_\psi^{(k)}$ *; a default channel* $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$*;*

Begin

The user enters the information of device(s), in an online system, in order to do the device registration(s) through the system of Operator;

If (the system of Operator received $uID_u^{(k)}$ *and device(s) information) then*

If (check $uID_u^{(k)} \notin U^{(k)}$*) then*

a. *the system of Operator notifies the user need to do user registration;*
b. *the User and Device(s) Registration Phase should be ended;*

Else

the system of Operator will find out the signed contract;

End If;

While (all device(s) are registered in the system of Operator)

Do {

a. *the user inputs the information of device one by one;*
b. *the system of Operator issues the device ID,* $uID_u^{(k)}$ *to user according to the contract;*
c. *the system of Operator assigns the contract role* $cr_i^{(k)} \in CR^{(k)}$ *to the user, where contract role with the role mapping:*
   i. *the role of operator* $ro_m^{(k)}$ *mapping to network,* $N\varepsilon_\psi^{(k)}$ *,also available channels,* $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$*;*
   ii. *the role of CS* $rs_x^{(k)}$ *mapping to* $p_p^{(k)} \in PRMS^{(k)}$*;*
d. *the system of Operator assigns a default network* $N\varepsilon_\psi^{(k)}$ *and a default channel* $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$ *to each device according to the assigned contract role;*

};

End While;

Return*;*

□

Table 4. An example to illustrate how users and their devices are mapped to the contract roles

$U^{(k)}$	$D\upsilon^{(k)}$	$CT^{(A)}$	$CT^{(B)}$	$CT^{(C)}$
$u_1^{(k)}$	$d\upsilon_1^{(k)}, d\upsilon_2^{(k)}, d\upsilon_3^{(k)}$	$cr_1^{(A)}$	$cr_3^{(B)}$	$cr_2^{(C)}$, $cr_4^{(C)}$
$u_2^{(k)}$	$d\upsilon_4^{(k)}, d\upsilon_5^{(k)}, d\upsilon_6^{(k)}$	$cr_3^{(A)}$	*None*	*None*
$u_3^{(k)}$	$d\upsilon_7^{(k)}, d\upsilon_8^{(k)}$	*None*	$cr_2^{(B)}$, $cr_3^{(B)}$	$cr_4^{(C)}$
$u_4^{(k)}$	$d\upsilon_9^{(k)}$	$cr_5^{(A)}$	$cr_5^{(B)}$	$cr_3^{(C)}$
$u_5^{(k)}$	$d\upsilon_{10}^{(k)}$	$cr_5^{(A)}$	*None*	*None*

4.4.2 Access Phase

The user uses her/his registered device through a default channel $CH_{\eta\varepsilon}^{(k)}$ of the network $N\varepsilon_{\psi}^{(k)}$ which is constrained by the assigned contract role $cr_i^{(k)} \in CR^{(k)}$ activates an access request to the CS. In *Algorithm 2*, the contract role and the device authenticated by the system of Operator are legal or not. After passing the authentication, the user can activate multiple contract roles via an available channel of assigned network in order to access the resources from CS. For supporting cognitive network and cognitive radio (channel) to get better QoS, the device will sense received strength and quality of alternative channels of networks, periodically. After analysing the measured report, when the best strength and quality are better than the threshold values compared to the serving channel of network, the device will activate a new access request via the new channel of network which is having the best QoS according to the contract role. The CS will continue the services via new channel of network with the same permissions to the device.

Algorithm 2: Access Phase

Input: *contract role* $cr_i^{(k)} \in CR^{(k)}$***;*** *default network,* $N\varepsilon_{\psi}^{(k)}$*; default channel* $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$*;*

Output: *activate the contract role* $cr_i^{(k)} \in CR^{(k)}$ *to access the resources from CS ;*

Begin

User activates an access request to the CS by using his/her contract role $cr_i^{(k)} \in CR^{(k)}$ *through the default channel* $CH_{\eta\varepsilon}^{(k)}$ *of the default*

network $N\varepsilon_{\psi}^{(k)}$*;*

If (the contract role and the device authenticated by the system of Operator are illegal)
reject the access request for system of Operator and CS;
Else
activate the contract role $cr_i^{(k)} \in CR^{(k)}$ *to access the resources from CS via the default channel* $CH_{\eta\varepsilon}^{(k)}$ *of the default network* $N\varepsilon_{\psi}^{(k)}$;
End If;
While ((the device sense best $RSS_{c_i} + RSS_{TH}$ (dBm) *and* $RQ_{c_i} + RQ_{TH}$ *of alternative channel of network) > (*$RSS_{c_i} + RSS_{TH}$ (dBm) *and* $RQ_{c_i} + RQ_{TH}$ *));*
DO {
If (the alternative channel of network belongs to the contract role $cr_i^{(k)} \in CR^{(k)}$*)*
a. the device activate a new access request via the new channel of network $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$ *to CS;*
b. the CS will continue the services via the new channel of network $ch^{(k)} \in CH_{\eta\varepsilon}^{(k)}$ *with the same permissions* $p^{(k)} \in PRMS^{(k)}$ *to the device;*
End If;
};
End While;
Return*;*

---

where
RSS = receiving signal strength;
RQ = receiving signal quality;
c_i = channel i, $i = 1,2,3,\ldots$;
RSS_{TH} = treshold value of RSS;
RQ_{TH} = treshold value of RQ.

5. Application Scenarios

Assume that there are a CS and three distinct Operators managed by distinct Providers: $CT^{(A)}$, $CT^{(B)}$, and $CT^{(C)}$ (the notations are defined for the simple), respectively. We then assume that the users u_1, u_2, u_3, u_4, u_5 and $dv_1, dv_2, dv_3, dv_4, dv_5, dv_6, dv_7, dv_8, dv_9, dv_{10}$, after registration procedure in the

CS and the Operators according to *Algorithm 1*, have satisfied the UA relations $UA^{(A)}$, $UA^{(B)}$, and $UA^{(C)}$where the relations are defined according to *Definition 5* as follows: $u_1\left(d\upsilon_1{}^{(A)}, d\upsilon_2{}^{(A)}, d\upsilon_3{}^{(A)}\right) \in UA^{(A)}, u_2\left(d\upsilon_4{}^{(A)}, d\upsilon_5{}^{(A)}, d\upsilon_6{}^{(A)}\right) \in UA^{(A)}$, $u_4\left(d\upsilon_9{}^{(A)}\right) \in UA^{(A)}$,and $u_5\left(d\upsilon_{10}{}^{(A)}\right) \in UA^{(A)}$ where $UA^{(A)} \subseteq U^{(A)} \times D\upsilon^{(A)} \times CR^{(A)}$; Also $u_1\left(d\upsilon_1{}^{(B)}, d\upsilon_2{}^{(B)}\right) \in UA^{(B)}$, $u_2\left(d\upsilon_5{}^{(B)}, d\upsilon_6{}^{(B)}\right) \in UA^{(B)}$, and $u_4\left(d\upsilon_9{}^{(B)}\right) \in UA^{(B)}$ where $UA^{(B)} \subseteq U^{(B)} \times D\upsilon^{(B)} \times CR^{(B)}$; Finally, $u_1\left(d\upsilon_3{}^{(C)}\right) \in UA^{(C)}$, $u_3\left(d\upsilon_7{}^{(C)}, d\upsilon_8{}^{(C)}\right) \in UA^{(C)}$ and $u_4\left(d\upsilon_9{}^{(C)}\right) \in UA^{(C)}$ where $UA^{(C)} \subseteq U^{(C)} \times D\upsilon^{(C)} \times CR^{(C)}$. We illustrate cases with the following.

***Case 1*:** The users: u_2 using $d\upsilon_4, d\upsilon_5, \text{or } d\upsilon_6$, and u_4 via $d\upsilon_9$ shown in *Table 3*, are assigned the contract roles as $cr_3{}^{(A)}$ and $cr_5{}^{(A)}$, respectively, by the $CT^{(A)}$, where the two contract roles satisfy the ascendancy relation as $cr_3{}^{(A)} \succeq cr_5{}^{(A)}$. In other words, according to *Definition 6*, the user u_2 using $d\upsilon_4, d\upsilon_5, \text{or } d\upsilon_6$ can access not only the resources with the permissions $prms_6{}^{(A)}$ and $prms_7{}^{(A)}$ via the available channels $\{ch_4\} \subseteq CH_{m_6}{}^{(A)}$ and $\{ch_5, ch_6\} \subseteq CH_{w_7}{}^{(A)}$and the assigned networks $m_6{}^{(A)}$ and $w_7{}^{(A)}$ set of the contract role, $cr_3{}^{(A)}$, but also the resources with the permissions via the available channels and the assigned networks set of the contract role, $cr_5{}^{(A)}$. On the contrary, however, the user u_4 using $d\upsilon_9$ cannot access the resources with the permissions $prms_{10}{}^{(A)}$ via the available channels $\{ch_9, ch_{10}\} \subseteq CH_{m_{10}}{}^{(A)}$ and the assigned networks $m_{10}{}^{(A)}$ of the contract role, $cr_3{}^{(A)}$. This implies that the user u_4 is assigned to the contract role $cr_5{}^{(A)}$, that is, the user u_4 is not assigned to another contract role $cr_3{}^{(A)}$. In *Table 2*, if the $CT^{(B)}$ server enforces these two contract roles $cr_2{}^{(B)}$ and $cr_3{}^{(B)}$, such that these two contract roles share a SSD relation according to *Definition 7*, and if these two contract roles are conflicting, then the user u_2 using $d\upsilon_5$ or $d\upsilon_5$ may never assign to these two contract roles $cr_2{}^{(B)}$ and $cr_3{}^{(B)}$, i.e. $\left(\left\{cr_2{}^{(B)}, cr_3{}^{(B)}\right\} \mid \left\{d\upsilon_5{}^{(B)}, d\upsilon_6{}^{(B)}\right\}, 2\right) \in SSD^{(B)}$. In *Table 3*, according to *Definition 8*, no users using any devices are allowed to activate both contract roles $cr_2{}^{(C)}$ and $cr_4{}^{(C)}$ in a single session, i.e. $\left(\left\{cr_2{}^{(C)}, cr_4{}^{(C)}\right\} \mid \left\{d\upsilon_3{}^{(C)}\right\}, 2\right) \in DSD^{(C)}$. The fact is that no DSD constraint on $cr_2{}^{(C)}$ and $cr_4{}^{(C)}$ is specified for the other contract roles. Only the $CT^{(C)}$ server enforces the DSD constraint that the user u_1 via $d\upsilon_3$ may never activate these two contract roles for a single user's session. Finally, after successfully performing a user's authentication by way of $CT^{(k)}$, a user can be allowed to

access this CS via registered devices through available channels and assigned networks according to the contract roles.

***Case 2*:** A user, u_5, who uses her/his device, dv_{10}, is accessing the services from the network of $CT^{(A)}$ via current channel $ch_{10} \in CH_{m_{10}}{}^{(A)}$ by activating the assigned contract role, $cr_5^{(A)}$. On the meantime, the device performs *Algorithm 2.* Once the device detects an alternative channel of network, which having the better receiving strength and quality. The device will reselect the new channel, e.g. $ch_9 \in CH_{m_{10}}{}^{(A)}$, and requests it to the network of $CT^{(A)}$. Therefore, the network of $CT^{(A)}$ can reallocate the new channel to the device dv_{10}. After receiving response from the network of $CT^{(A)}$, the device, dv_{10}, will access the same services with the same permissions $prms_{10}{}^{(A)}$ via new channel $ch_9 \in CH_{m_{10}}{}^{(A)}$ by using according the contract role $cr_5^{(A)}$.

6. Discussions

The important characteristics in our generalized Cognitive RBAC model are analysed in the following. The CS may define Roles of CS (Rs) and Permissions ($PRMS$), the Operator(s) may define Roles of Operator (Ro), allocate Channels (CH), and assign Networks ($N\varepsilon$), and the contract between them may concern about Users (U), Contract Roles (CR), Sessions (S), Devices (Dv), and Contract ($CT^{(k)}$)representing the set of users, contract roles, Roles of Operator, Roles of CS, permissions, sessions, devices, channels, networks and contracts set respectively; and furthermore, the CS and Operator(s) may also define the relationship UA, PA, RH, SSD and DSD.

6.1. Dynamic Changing Contract Role(s)

We assume that the CS and the system of Operator(s) make the contract(s), so that when system of Operator wants to delete and update a contract, in which the roles-mapping table should be reconstructed depending on the new contract, again. Therefore, the system of Operator is able to make sure that no user can activate the old contract role in any session to access permissions in the CS using any device through the channel of network. We propose two functions: $AddRole(\bullet)$ and $DeleteRole(\bullet)$ shown as *Function 1* and *Function 2* below, to support dynamic changing the contract role(s) mapping tables, e.g. *Table 1*, *Table 2* and *Table 3*.

Function 1. $AddRole(\bullet)$

$AddRole(cr^{(k)}, ro^{(k)}, rs^{(k)} : Name)$
$cr^{(k)} \notin CR^{(k)}, ro^{(k)} \notin Ro^{(k)}, rs^{(k)} \notin Rs^{(k)}$
$CR'^{(k)} = CR^{(k)} \cup \{cr^{(k)}\}$
$Ro'^{(k)} = Ro^{(k)} \cup \{ro^{(k)}\}$
$Rs'^{(k)} = Rs^{(k)} \cup \{rs^{(k)}\}$
$assigned\_user' = assigned\_user \cup \{cr^{(k)} \mapsto \varnothing\}$
$assigned\_network' = assigned\_network \cup \{ro^{(k)} \mapsto \varnothing\}$
$assigned\_channel' = assigned\_channel \cup \{ro^{(k)} \mapsto \varnothing\}$
$assigned\_permission' = assigned\_permission \cup \{rs^{(k)} \mapsto \varnothing\}$

Function 2. $DeleteRole(\bullet)$

$DeleteRole(cr^{(k)}, ro^{(k)}, rs^{(k)} : Name)$
$cr^{(k)} \notin CR^{(k)}, ro^{(k)} \notin Ro^{(k)}, rs^{(k)} \notin Rs^{(k)}$
$\varsigma^{(k)} \in S^{(k)} \bullet r^{(k)} \in s\_r(\varsigma^{(k)}) \Rightarrow DeleteSession$
$UA' = UA\{u : U, d\upsilon : D\upsilon \bullet u, d\upsilon \mapsto cr\}$
$assigned\_user' = assigned\_user\{cr \mapsto assigned\_user(cr)\}$
$PA' = PA\{PRMS \bullet prms \mapsto rs\}$
$assigned\_permission' = assigned\_permission\{r \mapsto assigned\_permission(rs^{(k)})\}$
$assigned\_network' = assigned\_network\{r \mapsto assigned\_network(ro^{(k)})\}$
$assigned\_channel' = assigned\_channel\{r \mapsto assigned\_channel(ro^{(k)})\}$
$CR'^{(k)} = CR^{(k)}$
$Ro'^{(k)} = Ro^{(k)}$
$Rs'^{(k)} = Rs^{(k)}$

6.2. Comparisons

Finally, our Cognitive RBAC model is compared to influential research paper [1, 20] mentioned above. R. S. Sandhu *et al.*'s [1] proposed RBAC Models and Hsing-Chung Chen and Marsha Anjanette Violetta [20] proposed Cognitive RBAC in SHNs. To distinguish them with our model, we make table to compare some influential research papers with our scheme. The comparisons among Conventional RBAC, Cognitive RBAC in SHNs and MHNs are listed in *Table 5*.

Table 5. Comparison among Conventional RBAC, Cognitive RBAC in SHNs and MHNs

Items	Schemes		
	RBAC	Cognitive RBAC in SHNs [20]	Cognitive RBAC in MHNs
Multiple Assigned Networks	Not Addressed	Addressed	Addressed
Available Channels	Not Addressed	Addressed	Addressed
QoS	Not Addressed	Addressed	Addressed
Multiple Registered Devices	Not Addressed	Not Addressed	Addressed
SSD and DSD Constraint	User	User and Device	User and Device
Role Contract	Not Addressed	Not Addressed	Addressed between CS and Operator(s)
Cognitive Area Networks	Not Addressed	Home Area Network Local Area Network	Wi-Fi, Mobile Communication Network

7. Conclusions

In this work, we propose the generalized cognitive RBAC model in MHNs. In our scheme, a registered device constrained by its contract role(s) can intelligently sense the environment of networks, and choose a better access network as well as radio channel depending on its needed for the required performance of applications. A CS can support the services with cognitive RBAC management in LHNs the environment of networks. The contract should be made by the CS and each Operator, individually. The Operator can be as a mobile communication system provider or Wi-Fi provider. However, the MHNs consist of more than one mobile communication networks of operators and Wi-Fi of operators. In the model, we divide into two phases; there are the Contract and Registration Phase and the Access Phase. We also propose the SSD and DSD constraint based on not only a user but also on registered device(s). The cognitive RBAC model that we proposed has satisfied the requirements, in which the device can sense a better channel or network via a role, then the serving CS and the Operator can adaptively assign the better channel or the better network as well as its corresponding available radio channel to the role depended on the contract. Finally, the device can get the services from the CS using the role via accessing the new network as well as the new channel. Therefore, we then propose a very convenient RBAC model which can achieve more efficient management for

LHNs, and let mobile user can active perceive current network situations to get permissions and access corresponding services under the contract have made by CS with Network Providers.

Acknowledgements. This work was supported in part by Asia University, Taiwan, under Grant 101-asia-28, also by the National Science Council, Taiwan, Republic of China, under Grant NSC99-2221-E-468-011.

References

1. R. S. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman, “Role-Based Access Control Models,” *IEEE Computer*, vol. 29, no. 2, pp. 38 – 47, Feb. 1996.
2. Y. Cai, K. A. Hua, G. Cao., T. Xu, “Real-time Processing of Range-Monitoring Queries in Heterogeneous Mobile Databases,” *IEEE Trans. on Mobile Computing*, vol. 5, no. 7, pp. 931 – 942, Jul. 2006.
3. G. Baldini, S. Braghin, I. Nai Fovino, A. Trombetta, “Adaptive and Distributed Access Control in Cognitive Radio Networks,” *Proc. of The 24th IEEE International Conference on Advanced Information Networking and Applications (AINA 2010)*, pp. 988 - 995, 2010.
4. G. Chen, Z. Yong, M. Song, X. Wang, “Cognitive Access Control in Cognitive Heterogeneous Networks,” *Proc. of IEEE International Conference on Communications Technology and Applications (ICCTA -2009),* pp. 707 - 711, 2009.
5. D. F. Ferraiolo, R. Sandhu, S. Gavrila, D.R. Kuhn, and R. Chandramouli, “Proposed NIST Standard for Role-Based Access Control,” *ACM Trans. on Information and System Security*, vol. 4, no. 3, pp. 224 – 274, Aug. 2001.
6. H.-C. Chen, S.-J. Wang, J.-H. Wen, and C.-W. Chen, “Temporal and Location-Based RBAC model,” *Proc. of the Fifth International Joint Conference on INC, IMS and IDC (MCM 2010)*, pp. 2111 – 2116, Seoul, Korea, Aug. 25 – 27, 2009.
7. X. Cai, F. Liu, “Network Selection for Group Handover in Multi-Access Networks,” *Proc. of IEEE International Conference on Communications,* pp. 2164-2168, 2008.
8. H.-Y. Cui, Q.-J. Yan, “Heterogeneous Network Selection using a Novel Multi-Attribute Decision Method,” *Proc. of the Third International Conference on Communications and Networking in China*, pp.153-157, 2008.
9. X. Dong, J. Wang, Y. Zhang, M. Song, R. Feng, “End to End QoS Provisioning in Future Cognitive Heterogeneous Networks,” *Proc. of IEEE International Conference on Communications Technology and Applications (ICCTA -2009),* pp. 425 – 429, 2009.
10. R. W. Thomas, D. H. Friend, L. A. DaSilva, A. B. MacKenzie, “Cognitive Networks: Adaptation and Learning to Achieve End-to-End Performance Objectives,” *IEEE Communications Magazine*, vol. 44, no. 12, pp. 51 – 57, 2006.
11. J. Mitola, III Maguire, G.Q., Jr. R., “Cognitive Radio: Making Software Radios More Personal,” *IEEE Personal Communications*, vol. 6, no 4, pp. 13 – 18, Aug. 1999.
12. Wikipedia, “Cognitive Network,” http://en.wikipedia.org/wiki/Cognitive_network, 2012.
13. H. Nan, Z.-G. Cao, “Wireless Cognitive Networks: Concept and Instance,” *Computer Engineering and Applications*, vol. 45, no.2, pp. 1-6 , 2009.
14. J. Mitola, "Cognitive Radio – An Integrated Agent Architecture for Software

Defined Radio," Ph.D. Dissertation, Royal Institute of Technology, Kista, Sweden, May 8, 2000.
15. R. W. Thomas, L. A. Da Silva, A. B. Mac Kenzie, "Cognitive Networks," *Proc. of the First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Network*s, Baltimore, MD, USA, Nov. 8–11, 2005.
16. D. D. Clark, C. Partrige, J. C. Ramming, J. T. Wroclawski, "A Knowledge Plane for the Internet," *Proc. of the SIGCOMM*, Karlsruhe, Germany, Aug. 25–29, 2003.
17. C. Fortuna, M. Mohorcic, "Trends in the Development of Communication Networks: Cognitive Networks," *Computer Networks*, 2009.
18. Q. Mahmoud, "Cognitive Networks: Towards Self-Aware Networks," *John Wiley and Sons*, 2007.
19. S. Liang, "Cognitive Networks: Standardizing the Large Scale Wireless Systems". *Proc. of the 5th IEEE Consumer Communications and Networking Conference (CCNC 2008)*, pp. 988-992, Jan. 2008.
20. H.-C. Chen and M.-A. Violetta, "Cognitive RBAC in Small Heterogeneous Networks," *Proc. of the Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2012)*, pp. 494 – 499, Palermo, Italy, Jul. 4- 6, 2012.
21. R. Yu, Y. Zhang, S. Gjessing, C. Yuen, S. Xie, M. Guizani, "Cognitive-Radio-Based Hierarchical Communications Infrastructure for Smart Grid," *IEEE Network*, Sept. 2011.
22. Li, L., Yan, H., Li, H., Zhang, C., "Velocity Adaptation for Synchronizing a Mobile Agent Network," *Computer Science and Information Systems*, vol. 8, no. 4, pp.1303-1315, Oct. 2011.
23. D. Mitrović, M. Ivanović, Z. Budimac, M.Vidaković, "Supporting Heterogeneous Agent Mobility with ALAS," *Computer Science and Information Systems*, vol. 9, no. 3, pp. 1203-1230, Sept. 2012.
24. M. Ganzha, A. Omelczuk, M. Paprzycki, M. Wypysiak, "Information Resource Management in an Agent-based Virtual Organization—Initial Implementation," *Computer Science and Information Systems*, vol. 9, no. 3, pp. 1307-1330, Sept. 2012.

Dr. Hsing-Chung Chen received the B.S. degree in Electronic Engineering from National Taiwan University of Science and Technology, Taiwan, in 1994, and the M.S. degree in Industrial Education from National Normal University, Taiwan, in 1996, respectively. He received the Ph.D. degree in Electronic Engineering from National Chung Cheng University, Taiwan, in 2007. During the years 1991-2007, he had served as a Mobile Communication System Engineer at the Department of Mobile Business Group, Chunghwa Telecom Co., Ltd. During February 2008–present, he has served as the Assistant Professor of the Department of Computer Science and Information Engineering at Asia University, Taiwan. Currently, he is interested in researching Information Security, Cryptography, Role-based Access Control, Wireless Heterogeneous Networks, Mobile Communications, and Medical Image Processing. He was Workshop Co-Chair of NCM-2009, NCM-2010, ITAUC-2011, ITAUC-2012, ITAUC-2013, SMEUCE-2012, SMEUCE-2013, ATIMCN-2012, ATSME-2012, NGWMN-2010, NGWMN-2011, and NGWMN-

2012, Track Co-Chair of NBiS-2012 and NBiS-2013, Web & Local Arrangement Co-Chair of ISIC-2012. He was Program Committee Co-Chair of EMC-2012. Presently, he is also the Program Committee Co-Chair of CISIS-2013 and IMIS-2013. He is a member of IEEE, IET, CCISA, and ICCIT. He was as the Leader Guest Editor of SI: "Recent Advances in Networked Computing" of *Journal of Networks*. From July 2012 – present, Dr. Chen is also the Editor-in-Chief of *Newsletter of TWCERT/CC (From February 2013, it has been re-named as 'Taiwan Information Security Newsletter').*

Marsha Anjanette Violetta received the BS degree in Information System from Amikom College, Indonesia, in 2010, and the MS degree in Computer Science and Information Engineering from Asia University, Taiwan, in 2013. She is current a PhD Student studying in Asia university, Taiwan, from September 2013. She is also a lecturer in the Department of Informatics Engineering, Islamic University of Indonesia and also the Department of Information System, Amikom College, Indonesia. Her research interests include Role-Based Access Control, E-mail Phishing, Database Security, and Cloud Computing.

Dr. Chien-Erh Weng received the M.S. degree in Electrical Engineering from the National Yunlin University of Science & Technology, Yunlin, Taiwan, and the Ph.D. degree in electrical engineering from the National Chung Cheng University, Chiayi, R.O.C., in 2000 and 2007, respectively. Since Sep. 2010, he joined the Department of Electronic Communication Engineering at National Kaohsiung Marine University, Kaohsiung, Taiwan, R.O.C., as an Assistant Professor. His research interest is in the fields of performance study of UWB communication systems, wireless sensor networks and cooperative radio networks.

Dr. Tzu-Liang Kung received the BS degree in industrial administration from the National Taiwan University in 1997, the MS degree in statistics from the National Chiao Tung University, Taiwan, in 2001, and the PhD degree in computer science from the National Chiao Tung University in 2009. From 2001 to 2004, he served as a Senior Engineer at the Behavior Design Corporation, Taiwan. He is currently an assistant professor in the Department of Computer Science and Information Engineering, Asia University. His research interests include multivariate data analysis, machine translation, natural language processing, interconnected systems, fault-tolerant computing, algorithm design, and wireless networks.

Received: November 10, 2012; Accepted: March 25, 2013

DOI: 10.2298/CSIS120716035L

Content-based Image Retrieval using Spatial-color and Gabor Texture on a Mobile Device

Yong-Hwan Lee[1], Bonam Kim[2] and Sang-Burm Rhee[3]

[1] Department of Applied Computer Engineering, Dankook University
152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 448-701, Korea
hwany1458@empas.com

[2] Division of Electrical and Computer Engineering, Chungnam National University
99, Daehak-ro, Yuseong-gu, Daejon, 305-764, Korea
kimbona@cnu.ac.kr**

[3] Department of Applied Computer Engineering, Dankook University
152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 448-701, Korea
sbrhee@dankook.ac.kr

Abstract. Mobile image retrieval is one of the most exciting and fastest growing research fields in the area of multimedia technology. As the amount of digital contents continues to grow users are experiencing increasing difficulty in finding specific images in their image libraries. This paper proposes a new efficient and effective mobile image retrieval method that applies a weighted combination of color and texture utilizing spatial-color and second order statistics. The system for mobile image searches runs in real-time on an iPhone and can easily be used to find a specific image. To evaluate the performance of the new method, we assessed the Xcode simulations performance in terms of average precision and F-score using several image databases and compare the results with those obtained using existing methods such as MPEG-7. Experimental trials revealed that the proposed descriptor exhibited a significant improvement of over 13% in retrieval effectiveness, compared to the best of the other descriptors.

Keywords: Mobile Content-Based Image Retrieval, Image Representation and Recognition, Image Descriptor, Spatial-Color, Gabor Texture

1. Introduction

The mobile phones that we use in our everyday life have become popular multimedia devices, and it is not uncommon to observe users capturing photos on their mobile phones, instead of using dedicated digital cameras or video cameras. This ease of use means that the number of digital images will only continue to rise with the rapid development of mobile devices, and individual users rapidly accumulate very large image repositories including both their personal photo archives and image from wider-access digital libraries. At present, users generally browse their personal multimedia repositories on mobile devices by

** Corresponding author : Bonam Kim

scrolling through image libraries or by manually creating a series of folders to fit their needs, and browsing through these folders. However, as the amount of digital content continues to increase end-users are beginning to suffer difficulties in locating specific images. As a result, research into more effective image retrieval techniques is currently receiving a great deal of attention, and image retrieval is now one of the most exciting and fastest growing areas in the field of multimedia technology [11].

There are two main approaches to image retrieval: text-based retrieval and content-based retrieval [22]. The popular text-based method requires images to be annotated with one or more keywords that can then be easily searched. However, this method involves a vast amount of labor and tends to be colored by personal subjectivity; the resulting lack of clarity often leads to mismatches in the retrieval process. In particular, this approach runs into critical problem concern the possibility of mismatches in a personal photo database. The alternative content-based method indexes images in a database by identifying similarities between them based on lower-level visual features such as color, texture, shape and spatial information [22] [23]. Although some systems are designed for a specific domain such as medical image retrieval or personal identification [19], a CBIR (Content-based Image Retrieval) system typically requires the construction of an *image descriptor*, which is characterized by two primary functions [23]. One is an extraction process that encodes the image into feature vectors, and another is a similarity measure that compares two images. The image descriptor D is formulated into 2-tuples as (F_D, S_D), where $F_D : \{I\} \rightarrow R^n$ is an extraction function that extracts a feature vector $\boldsymbol{f}$ from image $\boldsymbol{I}$, and $S_D : R^n \times R^n \rightarrow R$ is a distance measure function that computes the similarity between two feature vectors corresponding to images.

This paper proposes a new and more efficient mobile image descriptor that utilizes a weighted combination of color and texture features based on spatial-color and second-order statistical texture. This paper is an extension of work first presented in [9], here we provide a more thorough experimental comparison, and demonstrate much improved performance[4].

This paper is organized as follows: Section 2 reviews the related research on image retrieval and mobile image searching. In Section 3, we provide full details of our proposed descriptor. Section 4 presents some experimental results obtained using the proposed approach and assesses its performance compared to the methods currently used. Lastly, Section 5 concludes the paper by summarizing the study and discussing possible directions for future research.

[4] These improvements are due to three changes: First, a better algorithm is applied in the pre-processing stage to enhance computational time and memory efficiency. These involve checking the SubjectArea tag of EXIF metadata and reducing the size of the query image, either with main area or down sampling. Second, we utilize the Haar wavelet filter to analyze the image. Third, image databases were prepared more carefully and reasonably, focusing on the type of natural images that would be commonly found in personal photo libraries.

2. Related Work

This section summaries recently published research on content-based image retrieval including a consideration of the features commonly used in image searches and the issues involved in feature extraction from a color image.

2.1. Content-based Approaches

Many general-purpose image retrieval systems have been developed and proposed by both industrial and academic research laboratories, and it is not practicable to attempt to survey all of these in the limited space available. Hence, we focus on those works that deal specifically with standardization, especially those that feature in SC29/WG1 JPSearch. The standard approach involves two steps [10]:

Algorithm 1: RETRIEVAL finds out the relevant images from database

Input: A query image
Output: Relevant images in the database
1 $Extracting\ image\ features\ to\ a\ distinguishable\ extent$
2 $Matching\ these\ features\ to\ yield\ a\ result\ that\ is\ visually\ similar$
3 **return** $relevant\ images$

Fig. 1 shows the approaches traditionally used to search digital images. The type of image search and retrieval systems shown schematically on the left hand side of the figure require each image to be associated with one or more keywords entered by a human operator, while those on the use an image as a query and then attempts to retrieve other images which are similar. This represents the current state-of-the-art in CBIR systems described in SC29/WG1 JPSearch, International Standardization [10].

There already exist many research systems which apply image retrieval techniques to mobile search. IDeixis is an image-based mobile search system which combines image retrieval method and text-based search techniques. It uses CBIR methods to search the Web repository and/or other databases for matching images, and their result pages are based on current users location to find relevant images [26]. But, they still have the possibility of mismatching with text-based method. Reference [1] presents a content-based multimedia retrieval system designed for mobile platforms running Symbian-based OS. It is built on client-server architecture, and the system basically focusing on server-side application, while the client-side consists of the user interface and controllers. MOSIR is also a CBIR system for mobile phones, which facilitates instantaneous search on mobile phones for images similar to photos taken via phone cameras [18]. To find similar images, edge-based and color-layout features are used. It also enables region-based queries by detecting salient regions and extracting their features. They utilized image segmentation techniques to image

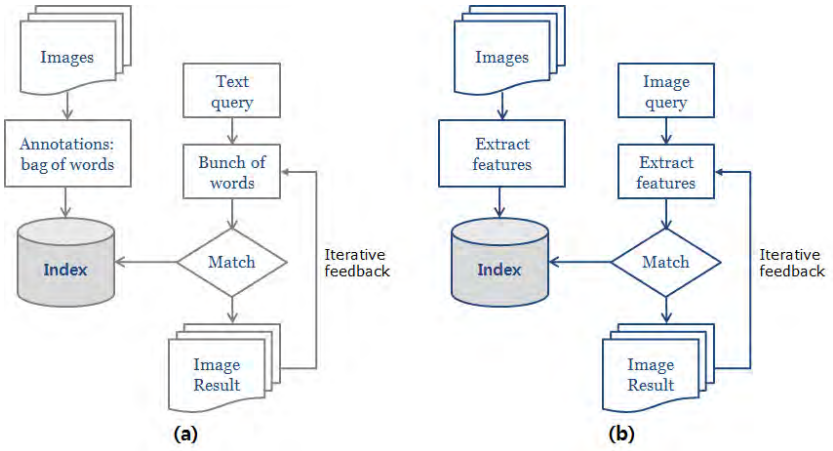


Fig. 1. Naive system view of image search and CBIR [10].

filtering processes prior to image analysis, while we use EXIF data, which gives a better way to find approximate position within image. In [24], they present mobile search systems which support image queries and audio queries, covering typical design for mobile visual and audio search. But, their work flow of matching core logic is running on the server, and client-side module is only working as interface for the user.

2.2. Conventional Features Relevant to the Current Research

This subsection explains the conventional features used in the proposed retrieval method, namely *auto-correlograms* for color features and *gray-level co-occurrence metrics* as texture features.

A histogram is a graphical display of frequencies that represents the total distribution in a digital image [4]. The histogram for color c_i of image $\boldsymbol{I}$ is formulated as follows:

$$H_{c_i}(I) \;=\; probability_{p\in I}[p \in I_{c_i}] \tag{1}$$

Since the histogram simply corresponds to the probability of there being any pixels of color c_i in an image, this feature does not take into account the spatial distribution of color across different areas of the image. A correlogram characterizes not only the color distributions of pixels, but also the spatial correlations of pairs of colors (c_i, c_k) [20]. This feature describes the probability of finding a pixel p_2 of special color c_k at a distance d for a pixel p_1 of given color c_i. The correlogram for a color pair (c_i, c_k) is formulated as follows:

$$C^{d}_{c_i,c_k}(I) \;=\; probability_{p_1\in I_{c_i},p_2\in I}[p_2 \in I_{c_k} || p_1 - p_2| = d] \tag{2}$$

With all possible combinations of color pairs, the size of the correlogram is inevitably very large so a formulation, known as an auto-correlogram, is generally used. An auto-correlogram gives the probability of capturing the spatial correlation between identical colors only, and this is effectively a simplified subset of the correlogram, signified by $\Gamma_c^d(I) = C_{c_i,c_k}^d(I)$. Thus, the auto-correlogram is formulated as follows:

$$\Gamma_c^d(I) \;=\; probability_{p_1 \in I_c, p_2 \in I}[p_2 \in I_c || p_1 - p_2| = d] \tag{3}$$

Gray Level Co-occurrence Matrices (GLCM) contain information about the positions of pixels having similar gray level values [6]. The GLCM is defined by calculating how often a pixel with the intensity value i occurs in a specific spatial relationship to a pixel with the value j [12]. That is, the GLCM is created by first specifying a displacement vector $d = (d_x, d_y)$ and then counting all the pairs of pixels separated by d having gray levels i and j. The GLCM G is therefore computed over an $n \times m$ image, parameterized by an offset $(\Delta x, \Delta y)$ as follows:

$$G_d[i,j] = \sum_{p=1}^{n} \sum_{q=1}^{m} \begin{cases} 1, \text{ if } I(p,q) = i \text{ and } I(p+\Delta x, q+\Delta y) = j \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

These spatial relationships can be specified in a number of different ways, but the default is that between a pixel and its immediate neighbor to its right. However, it is possible to specify this relationship with different offsets and angles, as described in Section 3.

3. Proposed Mobile Image Descriptor

Fig. 2 depicts a block diagram of the proposed retrieval approach. A single feature may lack sufficient discriminatory information to permit the retrieval of relevant images [21], so for this study we opted to use multiple features utilizing a combination of color and texture features that have been extracted separately. In addition, local features have been proved to be effective in image analysis [7]. When a query image I_Q is entered into the retrieval system, it must first be preprocessed by reducing the size of the querying image and by performing color space conversion and color channel separation. Each of the channels is then wavelet decomposed into a wavelet image, after which color features $\boldsymbol{f}_c$ and texture features $\boldsymbol{f}_t$ are extracted from the transformed image. Next, the system combines two features with appropriate weighting to generate the query vector $\boldsymbol{F}$, which is applied to the extraction process used to encode images into feature vectors. In order to respond to the user's query, the system then computes the similarity between the query vector $\boldsymbol{F}_Q$ and each of the target vectors $\boldsymbol{F}_T$ in

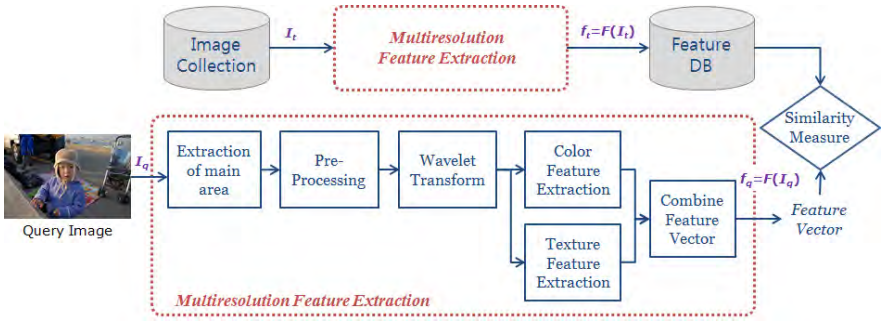


Fig. 2. Block Diagram of the Proposed Mobile Image Retrieval Method.

the database. Finally, it returns similar target images from the image database according to their similarity rankings.

Details of how the proposed algorithm extracts the color features using a wavelet spatial-color correlogram and the texture features using second-order statistical data for the texture based on Gabor wavelet transforms are provided in the following subsection.

3.1. Extraction of Spatial-color Based on Main Focus Region

The procedure for extracting the color features from an input image is shown in Fig. 3. Given a RGB query image, the SubjectArea tag of EXIF is first checked for resizing the entered image. The tag indicates the location and area of the main subject in the overall scene [5] [3], enabling the main area of the image to be extracted from the original if the tag is set as the region of interest (ROI). The main area is decided by computing the value of the tag, and by choosing the largest areas of intersection, as shown in Fig. 4. Alternatively the whole image can be reduced using bi-directional down-sampling algorithm such as YCbCr $4:2:0$ Co-sited of the JPEG standard [5]. Then, the RGB color space of the reduced image is converted into HSV space, presented as I^c, where $c \in \{H, S, V\}$. When extracting a color feature with a correlogram, HSV color space is known to provide better correspondence with human perceptions of color similarities than other color spaces [20].

Next, each of the three channels is wavelet transformed into two consecutive levels using a Haar filter, which is a good compromise between computational time and performance [13], denoted as $W_{s,l}^C$ where s indicates the four orientations of the sub-bands $s \in \{LL, LH, HL, HH\}$, l is the level of wavelet decomposition, and C represents the color channels. Thereafter, wavelet coefficients are quantized into $Q_{s,l}^C$ with different levels for each scale and sub-band. The number of quantization levels for each sub-band is weighted, according to the ratio $LL : LH : HL : HH = 2 : 1 : 1 : 0$. The correlogram for the HH sub-band cannot be computed because wavelet coefficients corresponding to HH have no significant spatial correlation. In order to reduce the computational

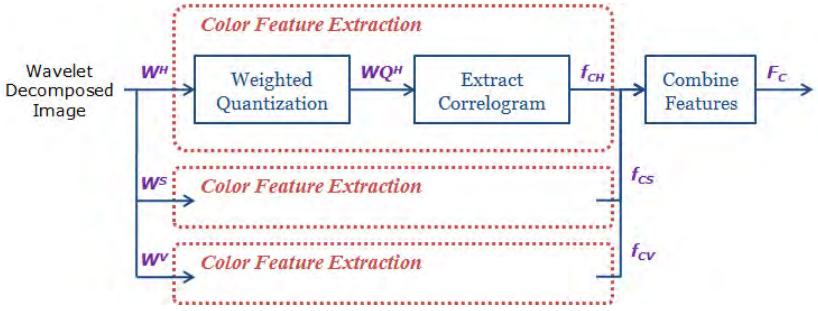


Fig. 3. Procedure for Weighted Wavelet Spatial Color Extraction.

Fig. 4. Five Types of Main Area of Query Image with SubjectArea Tag (Red point and circle).

time required for the extraction of the feature vectors a quantized color *codebook* was created for the proposed algorithm that functions as a color lookup table. Next, the horizontal, vertical and both directional correlograms for the quantized coefficients are calculated for the LH, HL and LL sub-bands in each scale. The correlogram of image I, which comprises the pixels $p(x, y)$ is then re-formulated from the definition of equation (X):

$$\Gamma_c^d(I) \;=\; \frac{|\{p(x,y) \mid I(x,y) = c_i; I(x \pm d, y \pm d) = c_i\}|}{|\{p(x,y) \mid I(x,y) = c_i\}|} \tag{5}$$

where c_i is the distinct value of each color and d is the fixed distance of the correlation.

Thus, the correlogram of wavelet coefficients for LL is computed as follows:

$$\alpha_{c_i}^d(W_{LL}) \;=\; \frac{|\{(x,y) \mid W_{LL}(x,y) = c_i; W_{LL}(x \pm d, y \pm d) = c_i\}|}{8 \times d \times |\{(x,y) \mid W_{LL}(x,y) = c_i\}|} \tag{6}$$

where W_{LL} is the wavelet decomposed image of LL sub-band, c_i is the quantized color, and d is the correlation distance.

Wavelet coefficients for LH correspond to the low pass filter and the high pass filter in the horizontal and vertical directions, respectively. Correlogram calculations on the LH sub-band can logically proceed only in a horizontal direction (low pass filtering), so the horizontal correlogram of the LH coefficients

is computed as follows:

$$\alpha_{c_i}^{d}(W_{LH}) = \frac{|\{(x,y) \mid W_{LH}(x,y) = c_i; W_{LH}(x, y \pm d) = c_i\}|}{2 \times d \times |\{(x,y) \mid W_{LH}(x,y) = c_i\}|} \quad (7)$$

Similarly, the vertical correlogram of the HL coefficients is computed using only the vertical direction, as described in equation (8):

$$\alpha_{c_i}^{d}(W_{HL}) = \frac{|\{(x,y) \mid W_{HL}(x,y) = c_i; W_{HL}(x \pm d, y) = c_i\}|}{2 \times d \times |\{(x,y) \mid W_{HL}(x,y) = c_i\}|} \quad (8)$$

Fig. 5 represents the neighboring pixels of point p with distance $d = 1$, used in the proposed approach.

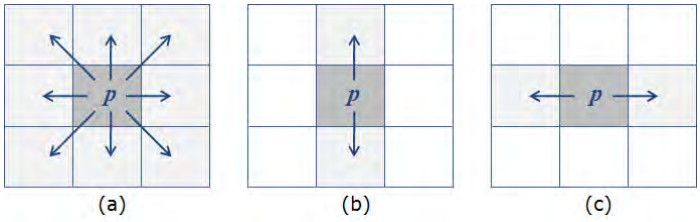


Fig. 5. Neighboring Pixels of Point p with Distance $d = 1$ (a) 8-directions for LL, (b) 2-directions for LH, and (c) 2-directions for HL.

Next, the wavelet color-spatial feature is combined with different weights for the sub-bands in the wavelet transform for each color channel, as follows:

$$f_c(C) = \lfloor \omega_{LL} \times \alpha_{LL}^{d},\ \omega_{LH} \times \alpha_{LH}^{d},\ \omega_{HL} \times \alpha_{HL}^{d} \rfloor \quad (9)$$

where the large C indicates the channel of the color image that satisfies the condition $C \in \{H, S\}$, and ω is the weighted value for LL, LH and HL sub-bands.

In this extraction process the results of the feature vectors inherit the multi-scale and multi-resolution properties from the wavelet and the translation invariant property from the correlogram.

3.2. Extraction of Texture Feature Using GLCM

The second part of the proposed descriptor consists of the texture feature extraction, which is shown in Fig. 6.

In order to extract a texture feature from the transformed domain, a Gabor wavelet filter is commonly used, as this is known to outperform tree-structured wavelet transforms, pyramid-structured wavelet transforms and multi-resolution simultaneous auto-regressive models [16] [27]. In the first step, the image is

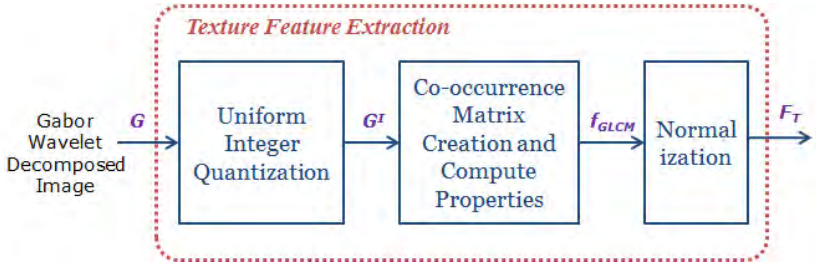


Fig. 6. Texture Feature Extraction Procedure in the Gabor Wavelet Domain.

converted to a gray-scale image I_G and a Gabor filter with two scales and four orientations is constructed. Gabor wavelet decomposition of the converted image is then performed, after which a GLCM is generated with five displacements ($0, 45, 90, 135$ and 315, as shown in Fig. 7) after performing integer quantization.

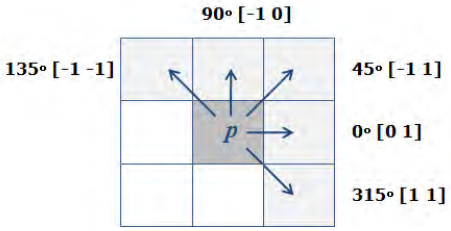


Fig. 7. Spatial Relationship used for Computing GLCM.

Five common texture features, namely contrast, correlation, energy, entropy and homogeneity are then calculated with the GLCM, as shown below:

$$\begin{aligned} Contrast &= \Sigma\Sigma(i-j)^2 \cdot P(i,j) \\ Correlation &= \Sigma\Sigma\frac{(i-\mu_i)(j-\mu_j)\cdot P(i,j)}{\delta_i\delta_j} \\ Energy &= \Sigma\Sigma P(i,j)^2 \\ Entropy &= \Sigma\Sigma P(i,j)\cdot\log P(i,j) \\ Homogeneity &= \Sigma\Sigma\frac{P(i,j)}{1+|i-j|} \end{aligned} \tag{10}$$

where $P(i,j)$ is the $(i,j)_{th}$ entry in the co-occurrence matrix P, and $\delta_i\delta_j$ and $\mu_i\mu_j$ are the mean and standard deviation of P, respectively.

Since this similarity measure does not consider rotation invariance, relevant texture images with different orientations may be missed by the searching process, as they would be assigned a low rank. Many papers using the Gabor texture feature either fails to consider rotation invariance or consider shifting feature elements in every direction to find the best match between the query

image and images in the database [25]. However, both these approached require expensive calculations.

In this paper, we propose implementing a simple circular shift on the feature map to solve the rotation variance problem associated with Gabor texture features. The total energy for each orientation is calculated and then the orientation with the highest total energy is deemed to be the dominant orientation. The feature elements in the dominant element are then shifted to become the first element in the feature vector ft and the other elements are circularly shifted accordingly. For example, if the original feature vector is $\boldsymbol{f} = [\epsilon_1, \epsilon_2, \cdots, \epsilon_n] = [1, 3, 2, 5, 2, 3]$ and "5" is the dominant orientation, the circularly shifted feature will be $\boldsymbol{f}_{CSF} = [5, 2, 3, 1, 3, 2]$.

3.3. Combination With Both Visual Features

The final step is to combine the two feature vectors. The color and texture features must first be normalized to reduce the effect of different feature dimensions and variances of the feature components. The normalized multiple feature F_D is computed as follows:

$$F_D \;=\; \lfloor \omega_c \times \frac{f_c}{N_c \times \delta_c \mu_c},\; \omega_t \times \frac{f_t}{N_t \times \delta_t \mu_t} \rfloor \tag{11}$$

where N_c and N_t are the dimensions of the color and texture feature vectors, $\delta_c \delta_t$ and $\mu_c \mu_t$ are the mean and standard deviations for color and texture, respectively, and ω_c and ω_t indicate the weights of color and texture, over the ranges $0 \leq \omega_c, \omega_t \leq 1$ and $\omega_c + \omega_t = 1$.

3.4. Similarity Measure

Once the features of the image have been extracted the retrieval results are obtained by measuring the similarity between the features of the query image and the pre-extracted features of the images in the database.

One of the most important parts of the matching process is the similarity function, because this decides how similar two features are. There are two methods commonly used to perform this function: the Minkowski-form metric and the Quadratic-form metric [4]. While the former compares only the corresponding bins between the histograms, the latter also considers the cross-relationships between the bins.

For the similarity measure, we can compute a distance that consists of the sum of the normalized distances for the visual feature:

$$S_D(F^Q, F^T) \;=\; \sum_{i=0}^{n-1} \frac{|F_i^Q - F_i^T|}{1 + F_i^Q + F_i^T} \tag{12}$$

where n is the number of feature dimensions, and F^Q and F^T are the query feature vector and target feature vector, respectively.

4. Experiments and Results

4.1. Datasets

To evaluate the performance of the proposed descriptor, we selected three datasets of images. Two of the datasets were the Corel photo gallery and the MPEG-7 common color dataset (CCD), both of which are widely used in the field of image retrieval. The third dataset included natural photos obtained from the website www.freeimages.co.uk. Each collection has images with a range of resolutions (e.g., $320 \times 240, 384 \times 256, 640 \times 420, 768 \times 512$ and $1,600 \times 1,200$) formatted as JPEGs and various kinds of images, including humans, flowers, vehicles, structures, fruits, materials, and so on. We used a subset of three datasets consisting of 2,200 images belonging to 85 classes of different kinds of images, chosen to estimate the effectiveness of image retrieval. To identify the SubjectArea of the EXIF tag, each of the images was manually generated to highlight a ROI. Fig. 8 shows just sample images from the MPEG-7 CCD, representing 50 image classes.

The ground truth sets (GTS) for evaluation were provided with the class in the experiments, but this was only used to calculate the effectiveness of the proposed approach. Retrieved images were considered to be relevant if they belonged to the same class as the query image.

Fig. 8. Sample Images from MPEG-7 CCD, from 50 Image Classes.

4.2. Experimental Results

The most common evaluation measures used in information retrieval (IR) are precision and recall, usually presented as a precision-recall curve [4]. Precision

denotes the ratio of retrieving an image that is relevant to the query, and recall indicates the ratio of the relevant images being retrieved, calculated as follows:

$$\begin{array}{l} precision \ = \ \frac{no.of relevant image retrieved}{no.of relevant images in collection} = \frac{a}{a+b} \\ recall \ = \ \frac{no.of relevant image retrieved}{no.of images retrieved} = \frac{a}{a+c} \end{array} \tag{13}$$

where a is the number of relevant images retrieved, b is the number of irrelevant images retrieved, and c is the number of relevant images that were not retrieved.

Since precision and recall are not always the most appropriate measures for evaluating IR, precision and recall scores are often combined into a single measure of performance, known as the *F-score* [2]. Higher values of the F-score are obtained when both precision and recall are higher. The formula for calculating the F-score is:

$$Fscore \ = \ \frac{2 \times precision \times recall}{precision + recall} \tag{14}$$

The following experimental approach was adopted to evaluate the search results and quantify any improvement in the retrieval performance. Leaveone-out cross validation (LOO-CV) performance was applied to obtain more reliable estimates compared to previous experiments whose results were based on a small number of queries, for example MPEG-7 [25]. Thus, each image in the database was selected in turn as the query image, and queried against the remaining images.

Fig. 9 shows the results for the comparison of retrieval effectiveness over the entire query. The source codes of C++ for other image descriptors [14], [6], [25], [20], [16], [17] are called within Objective-C++, to compare the effectiveness of the retrieval results. The values shown are computed in terms of recall and precision after the top 50 images have been retrieved, denoted as P(50) and R(50), respectively. The other methods included for comparison included a higher proportion of irrelevant images during the search, as indicated by their low precision and high recall. Based on the average from all queries for the F-score, 57.7% of all relevant images were retrieved by the new algorithm, which compares favorably to the best of the comparable methods, which retrieved just 44.7% of the relevant images in the image collection and was much better than the worst case, which retrieved only 13.3%. SCD, which is one of representative descriptor of MPEG-7, achieved 44.3% of relevant images, as shown in experimental results.

Experimentally the proposed method achieved an overall retrieval result score of 57.7%, markedly better than the 47.3% achieved by reducing query image through down sampling without checking main area of the image tag.

These results clearly indicated that the retrieval results achieved by the proposed approach achieved a higher ranking than any of the other methods tested based on its ability to cope with different scales and resolutions in the dataset of relevant images. Thus, this test shows that the proposed descriptor offers a more efficient way to conduct multi-resolution and multi-scale image retrieval.

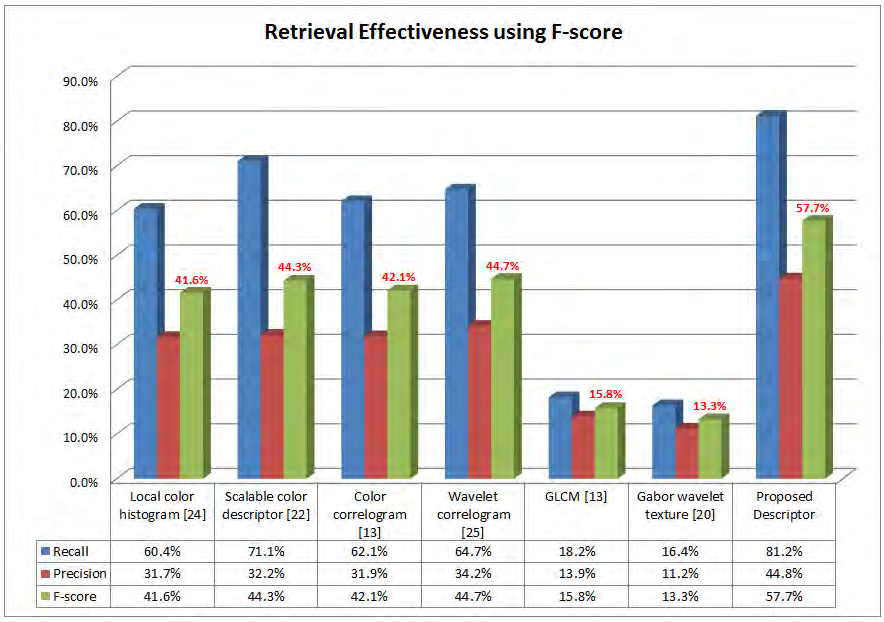


Fig. 9. Comparisons of Retrieval Effectiveness using F-Score.

Dimensionality of the feature vector is one of the most important factors affecting not only the amount of storage space needed, but also the retrieval accuracy and computational time [15]. Although the retrieval accuracy generally tends to improve as the dimension increases the amount of stage space and the computation time also increase. Thus, it is very important to choose the optimum dimension of the feature vector that will result in acceptable retrieval accuracy without incurring an excessive amount of storage space and computation time in CBIR. Since the ways of composing feature vectors in the search methods are quite different, it is necessary to fix all their feature vector dimensions to the same value for a fair comparison. Here, we chose the vector dimensions of image color features to be around 100, close to the dimension of the extracted vector for the proposed descriptor.

We implemented all approaches using Objective-C and C++ (source code for other methods) with Xcode 4.4.1 on a MacBook Pro running OS X 10.7 (Lion). Table 1 shows the computational characteristics of each method in order to compare the dimension of the feature vector and computational time required for each. Each of the computational times was calculated by the averages through batch processing. The retrieval time for exhaustive search is the sum of two times: T_{sim} and T_{sort} [8]. T_{sim} is the time to calculate the similarity between the query and every image in the database, and T_{sort} is the time to rank all the images in the data according to their similarity to the query. However, the retrieval time is highly depends only on the measuring the simi-

larity, especially on the time of extracting features. At this point, the proposed algorithm requires more computing time than others, enhancement of time is necessary for computing power and this is remained in future work.

Table 1. Computational time and dimensions of feature vectors for descriptors. Column (A) is the average time for extraction of visual features [sec/image], and (B) is the number of dimensions used for the feature vectors.

Method	(A)Extraction time	(B)Feature vector
Local color histogram	0.081	96
Scalable color descriptor	0.094	64
Color correlogram	0.127	96
Wavelet correlogram	0.131	96
GLCM	0.067	16
Gabor wavelet texture	0.098	48
Proposed Method	0.176	72

Fig. 10 depicts the user interfaces for our prototype system, which were used solely in extraction and retrieval mode for this study. Thus, users of the new system would only use the retrieval interface presented in Fig. 10(a) and 10(b). Fig. 10(c) shows how users can scroll through the retrieved results. However, the system does not yet support the re-query procedure, which remains for future work.

5. Conclusions

This paper proposes a new, more efficient mobile image descriptor that utilizes a combination of color features based on color-spatial information and texture features that make use of the Gabor texture of an image. In the preprocessing stage, the query image is resized, either by extracting the main area of the image or by down sampling to avoid a memory leak, taking into account the EXIF metadata. When using a correlogram for the color features, more computational time is required than for a histogram-based approach. For this reason, we incorporated a wavelet transform, whose coefficients provide information that is independent of the original image resolution, and appropriately weight the LL, LH and HL sub-bands. Also, the use of a color codebook helps to reduce the computational time needed.

The results of extensive experimental trials revealed that the proposed method produced a significant improvement of around 13% in retrieval effectiveness compared to the best of the other descriptors tested. However, the memory efficiency still needs further improvement, since limited memory resources remain a critical problem for mobile devices.

The main contribution of this paper lies in its weighted combination of color and texture for the use of mobile image retrieval based on spatial-color and

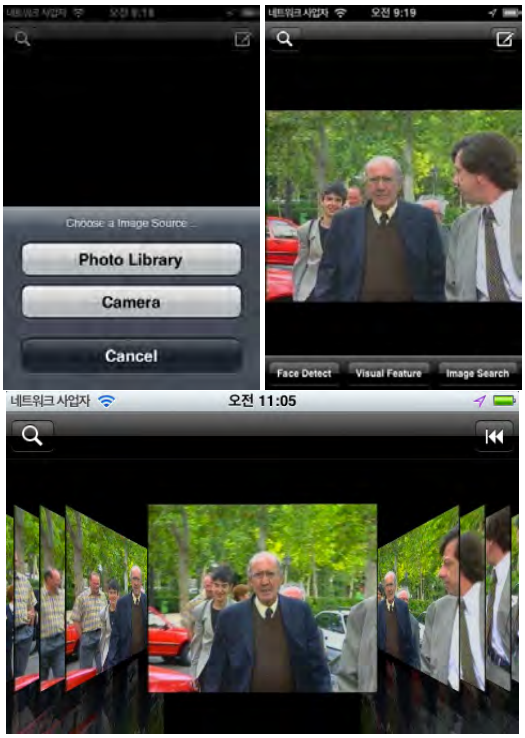


Fig. 10. Screenshots of the Prototype System: (a) User Interface for Selecting Query Image, (b) Query Image Chosen for Mobile search, and (c) Retrieved Images shown as Cover-flow on a iPhone Simulator, which is the Target Device used in Our Work.

second order statistics. As for future work, there are two main avenues for further development to enable the system to operate on smart phones such as the iPhone and Android. The first is the addition of an automatic procedure to identify the main area of an image, which had to be performed manually for these experiments. For example, automatic face detection and recognition would be particularly helpful. The second is the addition of textual or semantically related information such as geo-location and user events to the existing algorithm to enable users to search for photographs associated with specific features.

Acknowledgments. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program.

References

1. Ahmad, I., Gabbouj, M.: A generic content-based image retrieval framework for mobile devices. Multimedia Tools Applications (2010)

2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology behind Search. 2nd Editions. ACM Press Books, USA (2011)
3. Camera, Association, I.P.: Exchangeable image file format for digital still cameras: Exif version 2.3 (2010), [Online]. Available http://www.cipa.jp/english/hyoujunka/kikaku/pdf/DC-008-2010E.pdf (current March 2013)
4. Castelli, V., Bergman, L.D.: Image Databases - Search and Retrieval of Digital Imagery. Wiley Inter-Science (2002)
5. Group, M.W.: Guidelines for handling image metadata version 2.0. Motorola Labs, Paris (2010), [Online]. Available http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf (current March 2013)
6. Howarth, P., Ruger, S.: Evaluation of texture features for content-based image retrieval. In: Lecture Notes of Computer Science. pp. 326–334 (2004)
7. Jiang, X., Sun, T., Fu, G.: Multi-scale image semantic recognition with hierarchical visual vocabulary. Computer Science and Information Systems (2011)
8. Krishnamachari, S., Abdel-Mottaleg, M.: Hierarchical clustering algorithm for fast image retrieval. In: Part of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII. pp. 427–435 (1999)
9. Lee, Y.H., Kim, B., Rhee, S.B.: Content-based image retrieval using wavelet spatial-color and gabor normalized texture in multi-resolution database. In: International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. pp. 1–1. IMIS, Australia (2012)
10. Leong, M.K., Chang, W.: Framework and system components. ISO/IEC JTC1/SC29/WG1N3684 pp. 1–1 (2006)
11. Leong, M.K., Chang, W.: Iso/iec pdtr 24800-1: Jpsearch - part 1: Framework and system components. ISO/IEC JTC1/SC29/WG1N4203 pp. 1–1 (2007)
12. M.A., A., Maldague, Z., W.B., L.: A new color-texture approach for industrial products inspection. Journal of Multimedia pp. 44–50 (2006)
13. Makris, C.: Wavelet tress: a survey. Computer Science and Information Systems 9(2) (2012)
14. Malinga, B., Raicu, D., Frust, J.: Local vs. global histogram-based color image clustering. Technical Report TR06-010, School of Computer Science, De Paul University (2006)
15. Man, W.K.: Content-based Image Retrieval using MPEG-7 Dominant Color Descriptor. Master Thesis, Dept. of Electronic Engineering, City University of HongKong (2004)
16. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of large image data. IEEE Transactions on Pattern Analysis and Machine Intelligence (Special Issue on Digital Libraries) 18(8), 837–842 (1996)
17. Moghaddam, H.A., Khajoie, T.T., Rouhi, A.: A new algorithm for image indexing and retrieval using wavelet correlogram. In: Proceedings of International Conference on Image Processing. pp. 497–500 (2003)
18. Nakagawa, A., Kutics, A., Phyu, K.H.: Mosir: Mobile-based segment and image retrieval. In: International Conference on Wireless Communications, Networking and Mobile Computing. pp. 1–4. WiCOM (2011)
19. Ogiela, M.R., Ogiela, L.: Personal identification based on cognitive analysis of selected medical visualization. Journal of Internet Services and Information Security 2(3/4), 148–153 (2012)

20. Ojala, T., Rautiainen, M., Matinmikko, E., Aittola, M.: Sementic image retrieval with hsv correlograms. In: Proceeding 12th Scandinavian Conference in Image Analysis. pp. 621–627 (2001)
21. Skulsujirapa, P., Aramvith, S., Siddhichai, S.: Development of digital image retrieval technique using autocorrelogram and wavelet based texture. In: IEEE International Midwest Symposium on Circuits and Systems. pp. 273–276 (2004)
22. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
23. Torres, R.S., Falcao, A.Z.: Content-based image retrieval: Theory and applications. In: Brazilian Symposium on Computer Graphics and Image Processing. pp. 161–185. SIBGRAPO (2006)
24. Xing Xie, Lie Lu, M.J.H.L.G.S.W.Y.M.: Mobile search with multimodal queries. In: Proceedings of the IEEE. pp. 589–601. WiCOM (2008)
25. Yamada, A., Ocallaghan, R., Kim, S.K.: Mpeg-7 visual part of experimentation model (xm) version 27.0. ISO/IEC JTC1/SC29/WG11/N7808 (2006)
26. Yeh, T., Tollmar, K., Grauman, K., Darrell, T.: A picture is worth a thousand keywords: Image-based object search on a mobile platform. In: Proceeding Conference Hyman Factors in Computing System (2005)
27. Zhang, D., Wong, A., Indrawan, M., Lu, G.: Content-based image retrieval using gabor texture features. IEEE Transactions PAMI pp. 13–15 (2000)

Yong-Hwan Lee is received his Ph.D. degree in Electronics and Computer Engineering and M.S.degree in Computer Science from Dankook University, Korea, in 2007 and 1995, respectively. He is a research professor in Department of applied computer engineering at Dankook University. His research interests are the area of Image/Video Representation and Retrieval, Face Recognition, Augmented Reality, Mobile Programming and Multimedia Communication.

Bonam Kim is received the Ph.D. degree in Computer Science and Software Engineering from the Auburn University, Alabama, USA in 2006. She is a research professor in division of electrical and computer engineering at Chungnam National University since 2010. She joined the School of Electrical and Computer Engineering, CNU in March 2007. Her current research interests are in the areas of wireless ad hoc and sensor networks, network security and MIPv6.

Sang-Burm Rhee is received the Ph.D. degrees in Electronics Engineering from Yonsei Univ. in 1986. Now he is a professor at Dankook Univ. since 1979. His research interests are the area of Microprocessor, SoC(System-On-Chip), Pattern Recognition, Multimedia Processing. They include topics such as Object-oriented Methods for Audio/Video Watermarking, Pattern Recognition and HDL for SoC..

Received: July 16, 2012; Accepted: January 18, 2013.

DOI: 10.2298/CSIS121114036H

Design and Implementation of an Efficient and Programmable Future Internet Testbed in Taiwan

Jen-Wei Hu[1, 2], Chu-Sing Yang[1], and Te-Lung Liu[2]

[1] Institute of Computer and Communication Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C
{hujw, csyang}@ mail.ee.ncku.edu.tw
[2] National Center for High-Performance Computing, Tainan, Taiwan, R.O.C
{hujw, tlliu}@nchc.narl.org.tw

Abstract. Internet has played an important part in the success of information technologies. With the growing and changing demands, there are many limitations faced by current Internet. A number of network testbeds are created for solving a set of specific problems in Internet. Traditionally, these testbeds are lacking of large scale network and flexibility. Therefore, it is necessary to design and implement a testbed which can support wide range of experiments and has the ability of programmable network. Besides, there has been a big change enabled by cloud computing in recent years. Although networking technologies have lagged behind the advances in server virtualization, the networking is still an importance component to interconnect among virtual machines. There are also measurement issues with growing number of virtual machines in the same host. Therefore, we also propose integrating management functions of virtual network in our testbed. In this paper, we design and create a Future Internet testbed in Taiwan over TWAREN Research Network. This testbed evolves into an environment for programmable network and cloud computing. This paper also presents several finished and ongoing experiments on the testbed for multiple aspects including topology discovery, multimedia streaming, and virtual network integration. We will continue to extend our testbed and propose innovative applications for the next generation Internet.

Keywords: Future Internet, OpenFlow, Testbed, TWAREN.

1. Introduction

Internet has become the most important information exchange infrastructure that provides business transaction, personal communication, information sharing, etc. With wide range of applications and services applied to the Internet, some challenges are issued beyond its original design including scalability, security, mobility, flexibility, and so on [2], [10].

For resolving the increasing issues in current Internet, the U.S., E.U., Japan, and Korea have launched research projects for the Future Internet [5], [7], [13], [14], [20]. There were many issues discussed in these projects, especially on how to rethink and redesign decisions underlying current network architecture. Each project has its different aspects for Future Internet, but comes to the same conclusion, that is to provide an environment for performing research. Therefore, an experimental infrastructure on real networks is desirable to apply new protocols or develop new technologies. However, running experiments on the production network may be risky [4], and control-plane functions in most of network equipments are untouchable. There are some research projects focusing on eliminating the barriers of innovation, such as FEDERICA [8] and GENI [9]. The main goal is to develop a programmable network and enable multiple researchers to obtain a slice of resources by using network virtualization.

TaiWan Advanced Research & Education Network (TWAREN) [22] was established and managed by NCHC, which has been operating since Jan, 2004. It was developed using the latest network technologies and can offer users a variety of new services including IPv6, Multicast, and Light Path. The goals of TWAREN network design are:

- Hybrid technology: IP (routing) over optical Light Path (dark fiber, SDH, or Wavelength).
- Dual networks: production and research networks.
- Hierarchical topology: 3 tiers (cores, POPs, and end nodes).
- Multiple services.
- As shown in Fig. 1, TWAREN owns an island-wide network infrastructure in Taiwan. It plays an important role like Internet2 in the U.S. and GEANT in Europe. One mission of TWAREN is to continue developing and providing new technologies and environment for researchers. To meet this goal, we plan to deploy the Future Internet testbed in TWAREN and further extend into universities or research institutes.
- Besides, cloud computing has become a common word in IT industry. One key technology of cloud computing is hardware virtualization. A well-known of hardware virtualization techniques is the hypervisor (e.g., VMware, Xen, and KVM etc.) which allows multiple operating systems, called virtual machines (VMs), running concurrently on a same host machine. However, virtual networking technologies have lagged behind the advances in hardware virtualization [17]. The main reason is that cloud computing considers the service interaction more than network infrastructure. Each virtual machine shares same physical resources including network connection. Currently, most hypervisors use the existed network bridge to provide virtual machines connectivity [18]. As everything is virtualized in cloud environment, it gets even harder to manage. There still remains many research topics and open problems (e.g., traffic visibility, isolation, and security among VMs) in current cloud networking. In addition to supporting programmable network, we also expect our architecture to

provide a small cloud environment in which virtual switching services are enabled.

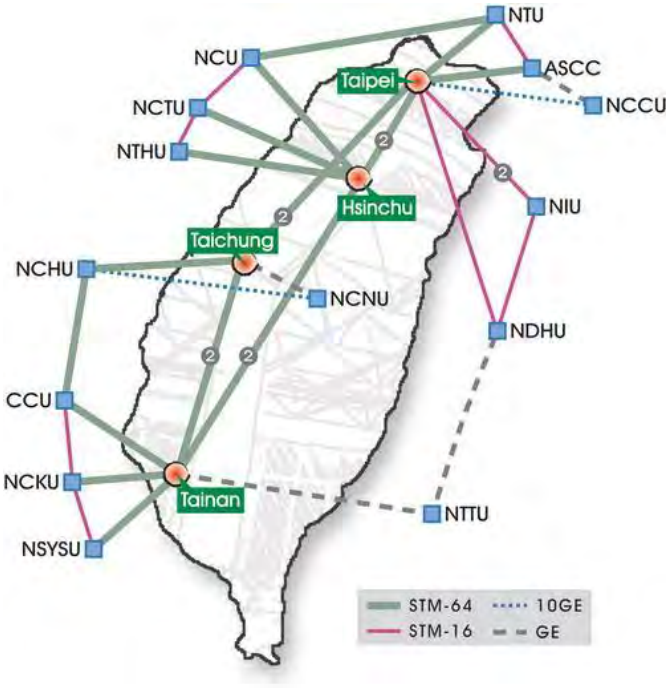


Fig. 1. TWAREN network structure [22]

The rest of this paper is organized as follows, In Section 2, we discuss related work in existing Future Internet testbed. In Section 3, we outline the implementation of our testbed and present current deployment status. Also, we briefly describe useful management modules that have been run on the testbed in Section 4. We present some performance results of our testbed in Section 5. Finally, we conclude this paper with a summary of this work in Section 6.

2. Background and Related Work

In this section, we present background information relevant to our work. We also survey related work and point out their relationship to our work. To design a future-proof testbed, there are some conditions that need to be considered. The network on this testbed should be programmable and isolable. Therefore, we first discuss Software defined network (SDN) and OpenFlow [15]. Then we introduce two famous testbeds based on the SDN architecture and discuss some similarities and differences between these existing testbeds and ours.

2.1. SDN and OpenFlow

The current Internet architecture is not sufficient to support the emerging applications in the future. One of the main reasons why new ideas cannot be tested on production networks is the closed support from the vendors. Legacy network devices, such as IP routers or Ethernet switches, run both data planes and control planes. All control functions are implemented by vendors and cannot be modified or touchable. To overcome these obstacles to testing innovative ideas and redesigning the Internet architecture, SDN approach was proposed. SDN separates data and control planes with well-defined protocol. The control functionalities are taken out of the equipment and given to a centralized or distributed system, while retaining only data plane functionality on the equipment.

OpenFlow is one of SDN implementations, which is an initiative by a group of people at Stanford University as part of their clean-slate program to redefine the Internet architecture. Processing packets decisions are moved to the OpenFlow controller. That means the network is programmable in OpenFlow. Each OpenFlow-enabled switch performs packets forwarding based on the flow table. The flow table contains a set of entries with packet header fields, an action, and flow statistics. Each flow entry is associated with actions that dictate how switch handles matching packets. Thus, OpenFlow uses distinct entries of flow tables to achieve isolation among experiments.

2.2. Future Internet Testbeds

Global Environment for Network Innovation (GENI) [3], [8] is a US program funded by the National Science Foundation (NSF). It is an experimental facility designed to form a federated environment to allow networking researchers to experiment on a wide variety of problems in communications, networking, distributed systems, cyber-security, and networked services and applications with emphasis on new ideas. GENI will provide an environment for evaluating new architectures and protocols, over fiber-optic networks equipped with optical switches, novel high-speed routers, radio networks and computational clusters [3].

The GENI architecture can be divided into three levels, Physical substrate, User services, and GENI Management Core (GMC). Physical substrate represents the set of physical resources, such as routers, switches; User services represent the set of services that are available for the users in order to fulfill their research goals; GMC defines a framework in order to bind user services with underlying physical substrate. In order to implement this, it includes a set of abstractions, interfaces and name spaces and provides an underlying messaging and remote operation invocation framework.

For constructing a topology of multiple substrates, GENI proposed the Aggregate Manager to control its own domain. Each Aggregate Manager has a unique RSpec which defines its Substrate resources. These RSpecs are

represented as a topology description of the individual substrate. However, how to automatically discover a global perspective of substrate topology is not mentioned.

The OpenFlow in Europe: Linking Infrastructure and Applications (OFELIA) is another famous testbed, which is funded by the European Union as part of its FP7 ICT work program. The OFELIA project consortium is made up several academic partners, commercial organizations, and telecom operators. Its infrastructure facility consists of five different islands spread across the Europe. Each island will host different capabilities to offer different functionalities to the researchers.

OFELIA architecture is still under development. However the architecture will be based on OpenFlow technology [3]. Currently, OpenFlow switches topology can be discovered when these reside in the single controller. With the growing OpenFlow domains, the environment of multiple controllers is needed for load balance. However, there does not have any mechanism which automatically retrieves the topology among OpenFlow switches controlled by multiple controllers.

3. Design and Implement Future Internet Testbed on TWAREN

We explain how to design and implement the future-proof testbed with OpenFlow in this section. As mentioned in Section 1, we expect the proposed architecture not only supporting OpenFlow but also providing virtual switching services for cloud networking research. To accomplish these goals, we propose the architecture as shown in Fig. 2. There are three parts in our design: Services layer, Networking layer, and Resources manager.

A number of controllers comprise the controller pool in the Services layer [1]. We provide different types of controllers (e.g., standalone, virtual machines) for researchers to request. If researchers would like to use their own host machine as a controller, binding a public IP is the only constrain. We use FlowVisor [19], a network virtualization layer of OpenFlow, to support these external controllers. Because FlowVisor contains a mapping table, we can maintain the relation between controllers from external users and our OpenFlow switches. There are several servers in the Services layer, some of them are classified as Virtualized Servers for concurrently running multiple virtual machines and the other are categorized into Bare-metal Servers for performance-concerned experiments.

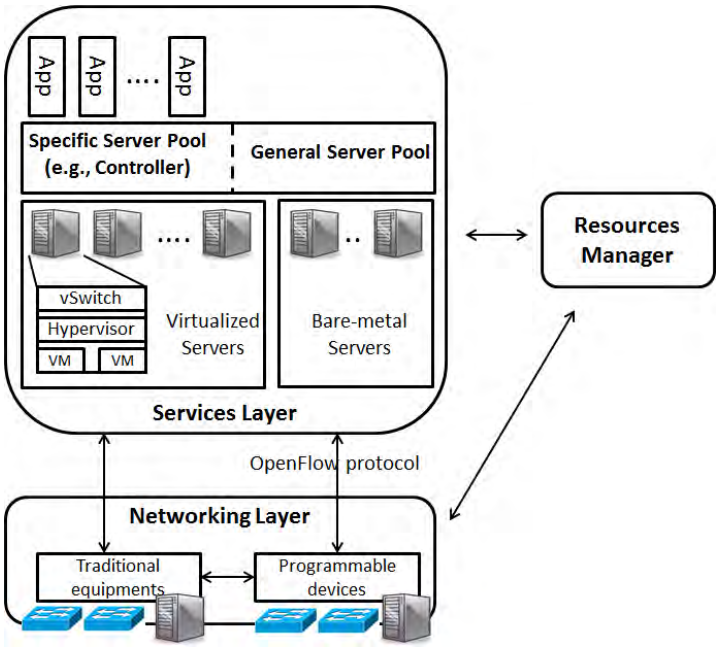


Fig. 2. Future Internet testbed architecture

For the Networking layer, we deploy legacy network equipments (e.g., switches, routers) and OpenFlow switches from different vendors including HP, Extreme, and PC with NetFPGA card. Since OpenFlow switches have to be operated at Layer 2 network, in this layer we provide hybrid solutions for extending our testbed smoothly. First, we use one of many TWAREN services, VPLS/VPN, which can connect multiple sites in the same local area network. This service is very useful for creating Layer 2 networks dynamically. However, there are some OpenFlow sites that cannot be applied directly to VPLS. For resolving this problem, we reserve several servers in the Service layer as tunneling servers in which software-based tunneling tools are installed (e.g., Capsulator [6]). About Resources manager, we use existing tools (e.g., OpenNebula, libvirt, virt-manager) to manage and control VMs in servers. It also maintains several services configurations, such as FlowVisor, tunneling, etc. We plan to develop a user interface for centralized management.

However, there are some OpenFlow sites that cannot be applied directly to VPLS. For resolving this problem, we reserve several servers in the Service layer as tunneling servers in which software-based tunneling tools are installed (e.g., Capsulator [6]). About Resources manager, we use existing tools (e.g., OpenNebula, libvirt, virt-manager) to manage and control VMs in servers. It also maintains several services configurations, such as FlowVisor, tunneling, etc. We plan to develop a user interface for centralized management.

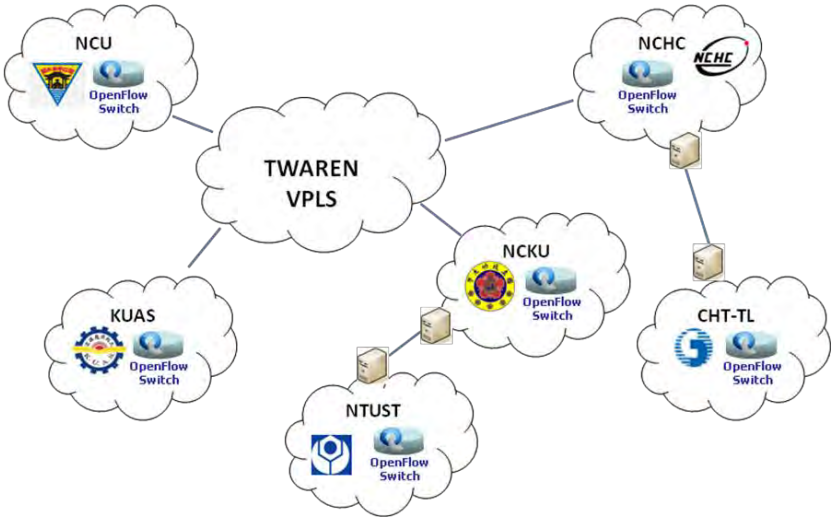


Fig. 3. Current OpenFlow connection in TWAREN research network

At the beginning of our project, two universities in Taiwan (e.g., NCKU, KUAS) participate in this Future Internet tesbed. Each site, including NCHC, is connected by Capsulator for operating at Layer 2 network. To deal with poor performance, we leverage VPLS service in TWAREN to provide a hardware-based tunneling. Many institutes that have interests in Future Internet research join our testbed, the current status is shown in Fig. 3.

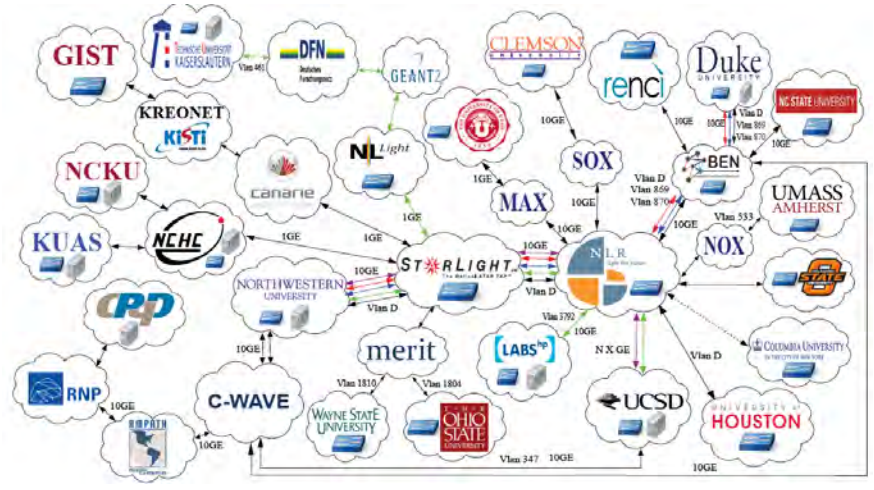


Fig. 4. Participating institutes of iGENI project [12].

In 2011, we joined iGENI [12] project by TWAREN international connection, as illustrated in Fig. 4. This will provide more real experiment network for our testbed.

4. Management Functionalities on TWAREN Testbed

In this section, we briefly describe some network management modules running on our testbed, which developed and resided in different aspects including inter-domain topology and virtual machines management.

4.1. Management of Inter-domain Connection

As mentioned previously, OpenFlow separates data and control plane. The only responsibility of OpenFlow switch is to forwarding received packets according to its flow table. Other complex works (e.g., routing decisions) are taken by controllers. Each OpenFlow switch has its own controller, so directly connected switches can be easily perceived by controllers. In addition, LLDP (Link Layer Discovery Protocol) packets are exchanged between any two OpenFlow switches to figure out neighbor switches. With these links information, controller can discover the topology in its controlled domain. As Fig. 5 shows, there are four OpenFlow switches (e.g., OF_A, OF_B, OF_C, and OF_D) residing in two different domains. $Controller_1$ for $Domain_1$ is responsible for OF_A and OF_B while OF_C, and OF_D are taken by $Controller_2$ in $Domain_2$. These two domains are directly connected by the link between OF_B and OF_C in Fig. 5. However, two controllers do not specify these links in their discovered object lists. That may cause the complexity of management and link provisioning.

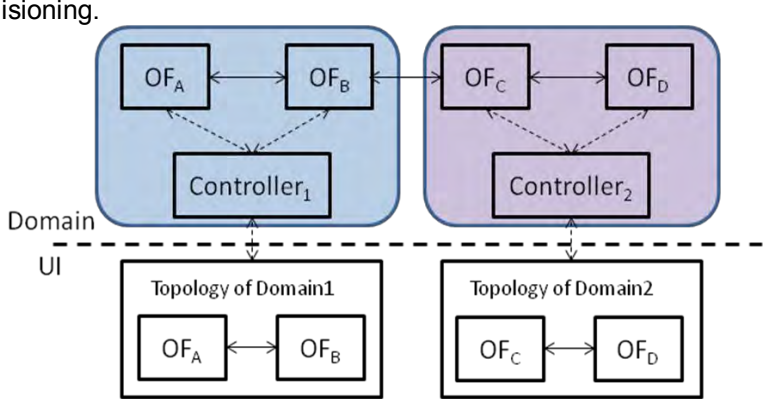


Fig. 5. Original inter-domain topology

For solving this problem, we proposed a mechanism to insert additional information into LLDP messages. In addition, we modify some applications in NOX for retrieving links among inter-domains. The full links information of our proposed solution is illustrated in Fig. 6.

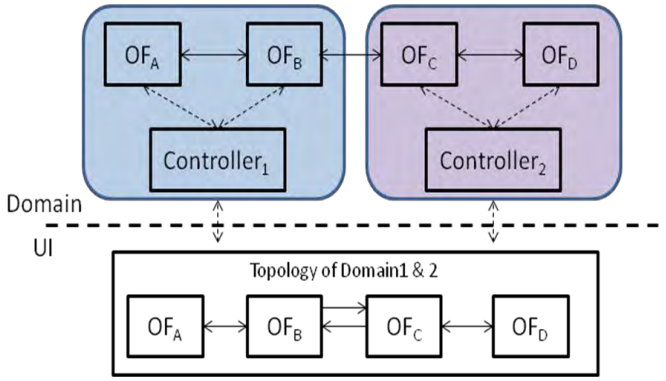


Fig. 6. Inter-domain information after applying proposed mechanism.

In general, LLDP information is sent by network devices from each of their interfaces periodically. A LLDP frame, as shown in Fig. 7, is composed by a series of LLDP Data Units (LLDPDUs). Each LLDPDU is a type-length-value (TLV) structure. There are four mandatory TLVs and zero or more optional TLVs in every LLDPDU.

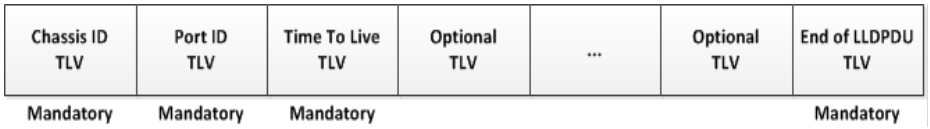


Fig. 7. LLDPDU format

As mentioned above, we can obtain topology information from devices but they must be resided in the same controller's domain. Hence, our main goal is to combine all topology information from different controllers. Through our experiments and observations, we found LLDP packets are also exchanged between any two directly connected devices. However, LLDP packets across different domains will be eventually dropped by receiving controller because they come from another domain. Since LLDP frame reserves optional TLVs to be extended by vendors or users, we add an optional TLV which contains controller information (e.g., IP and port) into generating application in NOX controller. Then, we modify the dropping policy and stored the received LLDP packets from different controller domains. Therefore, we aggregate all received information to build an overall topology. Fig. 8 shows the operations and relationship in modules of our mechanism.

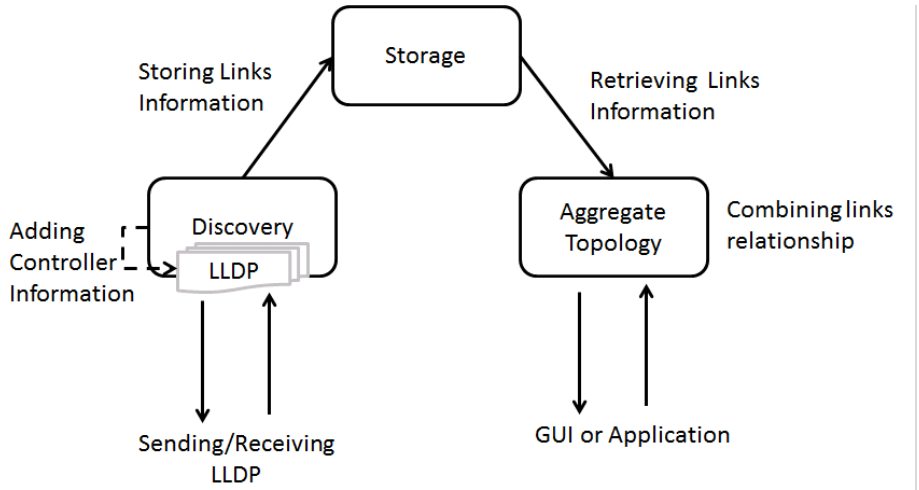


Fig. 8. Modules relationship in our mechanism

To verify the proposed mechanism in real OpenFlow network, we deploy it in three different domains including NCHC in Taiwan, NWU (Northwestern University) in the U.S., and CRC (Communications Research Centre) in Canada. Fig. 9 shows the links topologies of this experiment [11].

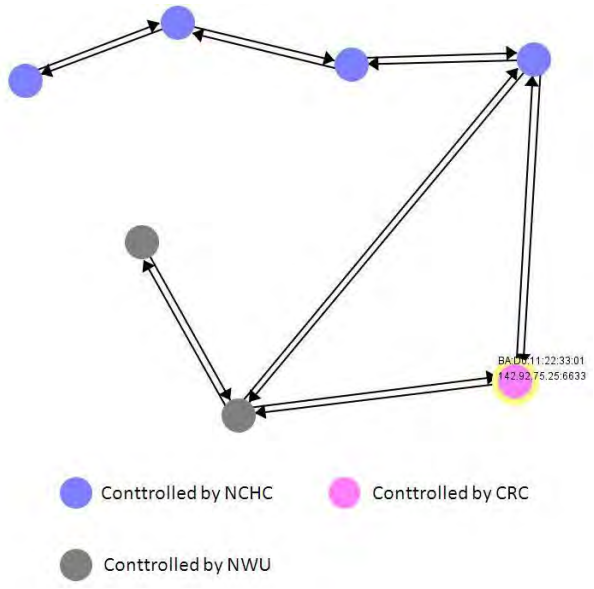


Fig. 9. Auto-discovery applies in a real multi-controller environment.

Since our proposed mechanism adds extra domain information in original LLDP packets, we quantified its processing overheads including CPU usage,

allocated memory, packet size, and processing time. We compare our proposal results against the original NOX controller. We setup two Linux hosts (Quad-code 2.53GHz, Xeon CPU, 4GB RAM, 1Gbps NIC), one to be an OpenFlow controller and the other uses Mininet to create the network topology which has 4 linear-connected OpenFlow switches.

Table 1. Comparison between original discovery application and our proposal

Mechanisms	CPU (%)	Memory (MBytes)	Packet size (Bytes)	Proc. Time (sec)
Original application	1%	23	60	1.5974
Our proposal (persistence version)	1%	23	60	7.5101
Our proposal (on the fly version)	1%	23	60	1.6102

Each application in OpenFlow controller is event-driven. When an OpenFlow switch receives packets, it will pass through all started applications and trigger their Packet_In event. For each mechanism, we generate 100 LLDP packets to measure the performance results shown in Table 1. There are no differences in CPU usage and allocated memory. The format of LLDP has only 14 bytes, but most network equipments will send it in 60-byte packet by padding the last few bytes. Although our mechanisms add extra information in original LLDP packet (e.g., Optional TLV), the size of modified LLDP packet is still less than 60 bytes. Therefore, the LLDP packet size is also no different from the original one. In order to discover multi-domain topology, we add a procedure to combine topology information received from each neighbor domain. For measuring overhead of our proposals, we define the processing time which represents a period starts from processing an incoming LLDP packet to storing its recognized information in controller. In our first proposal – persistence version, we had a poor performance than origin because it stored the topology information into persistent file for interoperating with multiple program languages application and recording current topology in our system. Furthermore, we developed another version (e.g., on the fly version) to solve this performance issue. It uses a compatible data structure instead of file and creates a thread to periodically write the topology information into file. There reduces much time when processing LLDP packets.

Considering scalability, in [21] they mentioned on an eight-core machine with 2GHz CPUs, NOX controller handles 1.6 million requests per second with an average response time of 2ms. We add additional topology information without affecting the original LLDP packet size and the time of processing LLDP packets is nearly same as the origin. Therefore, our proposals can have the same performance in real environment.

4.2. Management of Inter-domain Connection

In the past, standalone servers connect to physical switches directly. Many management functions, such as access control, port mirroring, and so on, are provided by network equipments. When moving to cloud, servers are replaced by VMs and reside in host machines. The network connections between servers and network devices have transferred to VMs and virtual switches. In this management module, we focus on integrating packet monitoring and network virtualization into our testbed, we call it VM manager module.

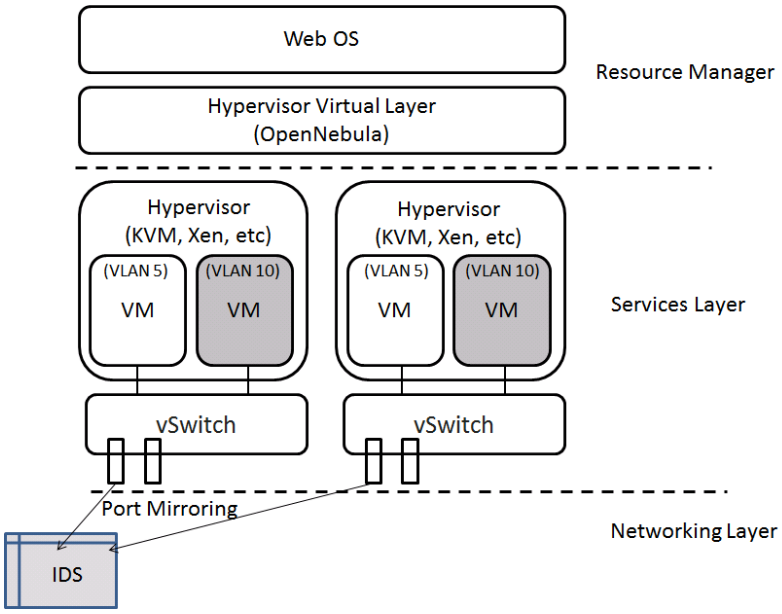


Fig. 10. VM Manager module

The common way for separating various users is to assign distinct ranges of private IP addresses. This mechanism can work properly in network connectivity, but all users will be resided in the same broadcast domain. That means users can access any virtual machines if they modify their own VM IP address to specific IP ranges. With increasing the number of VMs per host, this issue causes the difficulties of network management and security in cloud environment. Open vSwitch [16] is an open source tool fitting our requirements to resolve this problem. It implements 802.1q VLAN that can isolate different broadcast domain to keep inter-VM security. In addition to VLAN features, it also supports NetFlow, sFlow, and RSPAN for network visibility.

VM Manager module is shown in Fig. 10. It crosses the three layers of our testbed. In Resource Manager Layer, we use WebOS for our user interface and OpenNebula for hypervisor manager respectively. For Services Layer,

we set up several servers for deploying VMs. Each of them is installed Open vSwitch for virtual network and managed by OpenNebula. We implement some integrated programs to bind OpenNebula and Open vSwitch smoothly. Besides, we use Layer 2 technology, VLAN, to separate different VM users. But the valid VLAN range is from 1 to 4095, it is the limitation of our VM users in this status. We still develop and integrate other approaches to solve this limitation. Each of virtualized servers contains a management port which is connected to external analysis system for monitoring abnormal traffic among VMs. This integrated mechanism provides security capabilities in our VM users.

The VM resource allocation is an important issue for performance transmissions. In general, users require multiple VMs which are often arranged on the same host. Our VM manager module has different policies (e.g., Round-robin, Keep-in-one-host, and Random) to allocate multiple VMs requested from a single user. For suiting different types of VM services and allocating efficiency, we measure the performance by different packet sizes to provide allocating policy in our manager module. We setup two hosts (Quad-code 2.53GHz, Xeon CPU, 16GB RAM, 1Gbps NIC), each of them running 8 VMs and the measurement tool is “iperf”. Random choosing two VMs on each hosts (e.g., one is server and the other is client) to be the same host group. Then, we random choose one VM from other six VMs on each host respectively, and assign these two VMs as the different hosts group. Other VMs (e.g., five VMs in each host) run the same application which has ten megabytes in memory usage and one percent of CPU time. Our experiment results are shown in Fig. 11, larger packet size increases throughput because it generates less number of packets when transferring the same data frame. Each packet has fixed header, thus fewer packets will have less overhead (e.g., the source and destination addresses). We can also observe that the throughput for assigning two or more VMs on the different hosts is exceeds than arranging them on the same host as the packet size exceeds 32K. In this situation we find two hosts need using more memory to buffer and process packets when these two VMs on the same host. However, the memory usage will be shared and reduced if these two VMs are on different hosts. Therefore, we think packet size will be a factor in allocating VM resources. According to this experiment result, we extend the default VM allocation with a network-oriented policy which considers network factors (in currently, we just define default transferring packet size) to assign VM resources in our manager module.

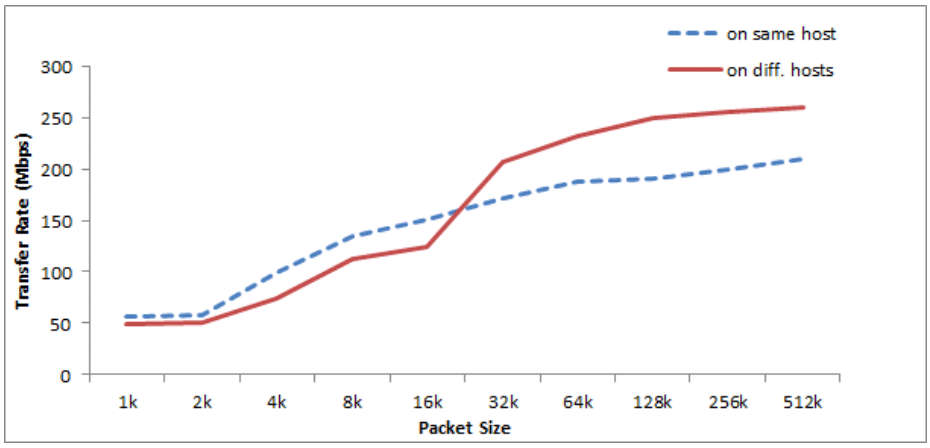


Fig. 11. Throughput of different VM assignment approaches.

5. Testbed Performance Result

As described the Networking layer of our testbed architecture in Section 3, we built two different mechanisms for network connection. In this section, we will do some performance experiments to measure the overhead of these two mechanisms in our testbed. In addition, VM Manager is also an important module in our testbed. We will compare its performance against original OpenNebula in this section.

In Table 2, NCHC-TN and NCHC-HC are the southern and the northern departments of our center respectively. The distance between NCHC-HC and NCHC-TN is around 230 km. Another site of our experiment, NCKU, is a university in southern Taiwan. The distance to NCKU from NCHC-TN is around 20 km while.

Our latency experiments used 100 64-byte packets. The first row in Table 2 shows the result. We also measured one-direction TCP throughput by different sizes of packets. For each case, we ran 20 30-second trials. The results show that VPLS technology is significantly faster and more efficient than the mechanism with tunneling software.

Table 2. Micro-benchmarks for TWAREN Future Internet testbed overheads

Cases	VPLS		Tunneling Software	
	NCHC-TN to NCHC-HC	NCHC-TN to NCKU	NCHC-TN to NCHC-HC	NCHC-TN to NCKU
RTT (ms)	3.512	0.895	5.822	2.873
Throughput (1M packet) (Mbps)	461	815	75.7	89.2
Throughput (10M packet) (Mbps)	473	831	76.5	89.3
Throughput (100M packet) (Mbps)	472	838	75.5	87.6

For comparing network throughput between VM Manager and original OpenNebula, we set up two Linux hosts which create four VMs in each host. Each VM has one-core 2.53GHz CPU, 512MB memory, and 1Gbps NIC. For the first trial, we compared the VM TCP throughput of VM Manager and original OpenNebula on the same host. We chose one host and divided its VMs into two groups. One VM of each group is running iperf server and the other is client. In this experiment, each group received a result and we chose minimum one of them to be TCP throughput. The Fig. 12 shows the first trial result. Clearly, VM Manager outperforms original OpenNebula at any sizes of transferring packets.

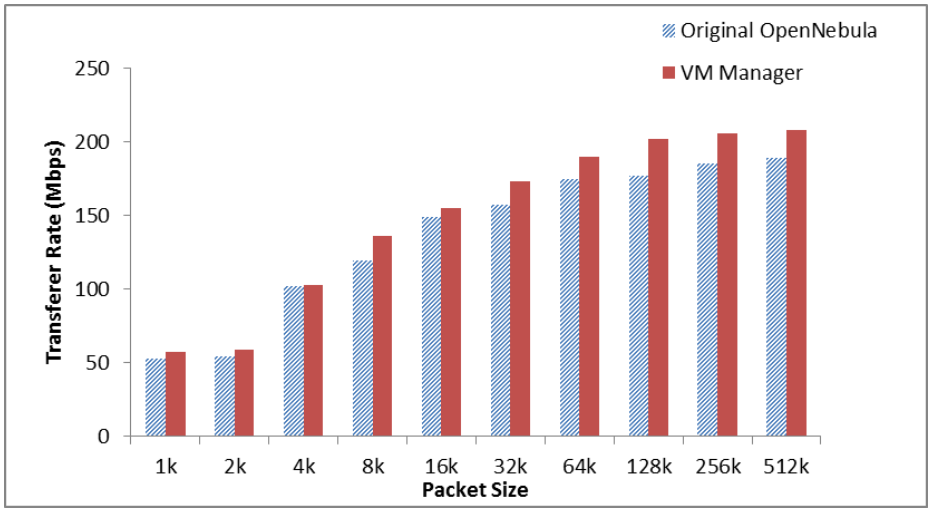


Fig. 12. Throughput with VMs on the same host

The second trial, we consider the network performance when VMs are arranged on different hosts. We classified two hosts into two groups, one host make its all VMs be iperf servers and all VMs of the other host are iperf clients. The experiment result is shown in Fig. 13, which appears VM Manager outperforms original OpenNebula by 19% in average.

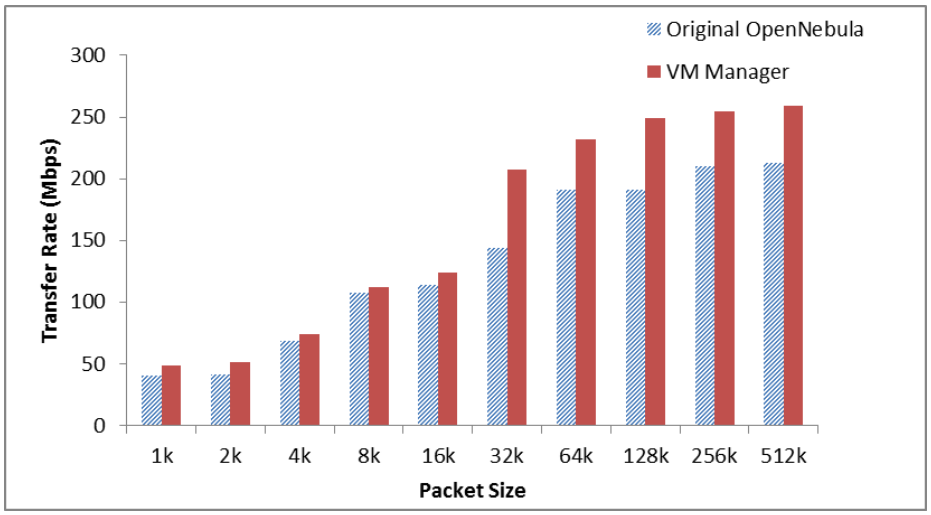


Fig. 13. Throughput with VMs on the different hosts

6. Conclusion

In this paper, we propose and create a Future Internet testbed which has the capabilities for programmable network and cloud. This testbed is deployed over TWAREN Research Network. We experiment and verify different research activities on this tesbed, including Future Internet and cloud. In our future work, we will keep developing more innovative functions for Future Internet. It will be useful to build and maintain a cross organization and a large scale multinational Future Internet platform. We believe the TWAREN Future Internet testbed opens up a new environment in Taiwan for network research. It enables us not only to design new thoughts, but also to solve and verify current issues in real network.

References

1. Bădică, C., Budimac, Z., Burkhard, H., Ivanović, M.: Software Agents: languages, tools, platforms. Computer Science and Information Systems, Vol. 8, No. 2, 255-296. (2011).

2. Bellovin, S. M., Clark, D. D., Perrig, A., and Song, D.: A Clean-Slate Design for the Next-Generation Secure Internet. GENI Design Document 05-05. (2005).
3. Belter, B., Campanella, M., Farina, F., Garcia-Espin, J., Jofre, J., Kaufman, P., Krzywania, R., Lechert, L., Loui, F., Nejabati, R., Reijs, V., Tziouvaras, C., Vlachogiannis, T., and Wilson, D.: Virtualisation Services and Framework – Study. Formal report from European Commission. (2012).
4. Bianco, A., Birke, R., Giraudo, L., and Palacin, M.: OpenFlow Switching: Data Plane Performance. Communications (ICC), 2010 IEEE International Conference, pp. 1-5. (2010).
5. Cameron, D.: Internet2: The Future of the Internet and Next-Generation Initiatives. Computer Technology Research Corp. (1999).
6. Capsulator, http://www.openflow.org/wk/index.php/Capsulator.
7. Fairhurst, G., Collini-Nocker, B., and Caviglione, L.: FIRST: Future Internet: A Role for Satellite Technology. IEEE International Workshop on Satellite and Space Communications (IWSSC). (2008).
8. FEDERICA: Federated E-infrastructure Dedicated to European Researchers Innovating in Computing network Architectures, http://www.fp7-federica.eu/.
9. GENI: Global Environment for Network Innovations, http://geni.net.
10. Greenberg, A., Hjalmtysson, G., Maltz, D. A., Myers, A., Rexford, J., Xie, G., Yan, H., Zhan, J., and Zhang, H.: A Clean Slate 4D Approach to Network Control and Management. ACM SIGCOMM Computer Communication Review, Vol. 35, Issue 5. (2005).
11. Huang, W. Y., Hu, J. W., Lin, S. C., Liu, T. L., Tsai, P. W., Yang, C. S., Yeh, F. I., Mambretti, J. J., and Chen, J. H.: The Implement of Automatic Network Topology Discovery System in Future Internet across Different Domains. 26th International Conference on Advanced Information Networking and Applications Workshops (WAINA). (2012).
12. iGENI: International Global Environment for. Network Innovations, http://groups.geni.net/geni/wiki/IGENI.
13. Kim, D. Y., Mathy, L., Campanella, M., Summerhill, R., Williams, J., Shimojo, S., Kitamura, Y., and Otsuki, H.: Future Internet: Challenges in Virtualization and Federation. Fifth Advanced International Conference on Telecommunications, (AICT), pp.1-8. (2009).
14. Lee, J., Kang, S., Lee, Y., and Lee, J.: A Study on the Future Internet Requirement and Strategy in Korea. 10th International Conference on Advanced Communication Technology (ICACT), Vol. 1, pp. 627-629. (2008).
15. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S., and Turner, J.: Openflow: enabling innovation in campus networks. SIGCOMM CCR, Vol. 38, no. 2, pp. 69-74. (2008).
16. Open vSwitch, http://www.openvswitch.org/.
17. Pettit, J., Gross, J., Pfaff, B., and Casado, M.: Virtual Switching in an Era of Advanced Edges. 2nd Workshop on Data Center – Converged and Virtual Ethernet Switching (DC-CAVES), ITC 22. (2010).
18. Pfaff, B., Pettit, J., Koponen, K.A.T., Casado, M., and Shenker, S.: Extending networking into the virtualization layer. Proceedings of the ACM SIGCOMM HotNets. (2009).
19. Sherwood, R., Gibb, G., Yap, K. K., Apenzeller, G., Casado, M., McKeown, N., and Parulkar, G.: FlowVisor: A Network Virtualization Layer. Tech. Rep. OPENFLOWTR- 2009-1, OpenFlowSwitch.org. (2009).
20. Stuckmann, P. and Zimmermann, R.: European research on future Internet design. IEEE Wireless Communications, Vol. 16, Issue 5, pp. 14-22. (2009).

21. Tootoonchian, A., Gorbunov, S., Ganjali, Y., Casado, and M., Sherwood, R.: On Controller Performance in Software-Defined Networks. 2nd USENIX Workshop on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services (Hot-ICE). (2012).
22. TWAREN Research Network, http://www.twaren.net/.

Jen-Wei Hu received the B.S. degree in Applied Mathematics from National Chung Hsing University, Taiwan, in 2001, and the M.S. degree in Computer Science and Engineering from National Sun Yat-sen University, Taiwan, in 2003. Currently, he works as an Assistant Engineer in the Network and Security Division of National Center for High-Performance Computing, Taiwan. His current research interests include Software-Defined Networking, Networking in data centers, and Multipath transmission.

Chu-Sing Yang is a Professor of Electrical Engineering in the Institute of Computer and Communication Engineering at National Cheng Kung University, Tainan, Taiwan. He received the B.Sc. degree in Engineering Science from National Cheng Kung University in 1976 and the M.Sc. and Ph.D. degrees in Electrical Engineering from National Cheng Kung University in 1984 and 1987, respectively. He joined the faculty of the Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan, as an Associate Professor in 1988. Since 1993, he has been a Professor in the Department of Computer Science and Engineering, National Sun Yat-sen University. He was the chair of the Department of Computer Science and Engineering, National Sun Yat-sen University from August 1995 to July 1999, and the director of the Computer Center, National Sun Yat-sen University from August 1998 to October 2002. He was the Program Chair of ICS-96 and Program Co-Chair of ICPP-2003 and MTPP-2010. He joined the faculty of the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, as a Professor in 2006. He participated in the design and deployment of Taiwan Advanced Research and Education Network and served as the deputy director of National Center for High-performance Computing, Taiwan from January 2007 to December 2008. His research interests include future classroom/meeting room, intelligent computing, network virtualization.

Te-Lung Liu received the B.S. and Ph.D. degrees in computer science from the National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1995 and 2002, respecively. He is currently a Research Scientist in National Center for High-Performance Computing, Tainan, Taiwan, R.O.C. He is also a Team Member of the Taiwan Advanced Research and Education Network (TWAREN) and now working on OpenFlow Testbed in Taiwan. His current research interests include Software-Defined Networking, Future Internet, optical networks, and network design.

Received: November 14, 2012; Accepted: April 05, 2013

DOI: 10.2298/CSIS120917037D

Key Management Approach for Secure Mobile Open IPTV Service

Inshil Doh[1], Jiyoung Lim[2]*, and Kijoon Chae[1]

[1]Ewha Womans University,
{isdoh1, kjchae}@ewha.ac.kr
[2]Korean Bible University
jylim@bible.ac.kr
*Correspondent Author

Abstract. In mobile open Internet Protocol TV (IPTV) which is one of the major attracting technologies recently, the security is a key issue for reliable service, because the mobility and the openness in IPTV could cause much more vulnerabilities to various attacks compared with traditional IPTV services. In this paper, we propose an energy-efficient and secure channel group key establishment and rekeying management scheme for mobile open IPTV services. Our scheme provides the data authentication between an Evolved Node B (eNB) or a Base Station and the mobile devices for the security enhancement and efficiently rekeys the group key when the membership changes. Additionally, it proposes a pairwise key establishment mechanism for open IPTV services through eNBs. Our proposal can cope with the security vulnerability in mobile open IPTV services and guarantee the secure group key rekeying in addition to decreasing the storage and communication overhead.

Keywords: group key; pairwise key; channel; security; rekeying; authentication; mobile open IPTV

1. Introduction

IPTV is a system through which television services are delivered using the Internet protocol suite over a packet-switched network. It has attracted a lot of interest as many intelligent devices appear and support IPTV related functions. Secure IP multicast may be used to support the secure transmission of IP packets to groups of receivers in IPTV services but neglects access control and network management. Key distribution solutions for secure group communications usually apply key refreshing techniques upon a group change (member join or leave) in order to impose both perfect forward and backward secrecy [1,2].

Recently, with the advance of mobile devices technology, users would want to receive their services through mobile devices anywhere, and mobility

is additionally required for IPTV service. However, the use of wireless environment has many risks and weaknesses when it is compared with the existing wired networks. There are two approaches for mobile IPTV security technologies as in Fig. 1. One is adding the mobility to IPTV, and the other is adding IP technologies to mobile TV such as DMB, DVB, and so on. In our work, we are focusing on adding the mobility to the fixed IPTV technology.

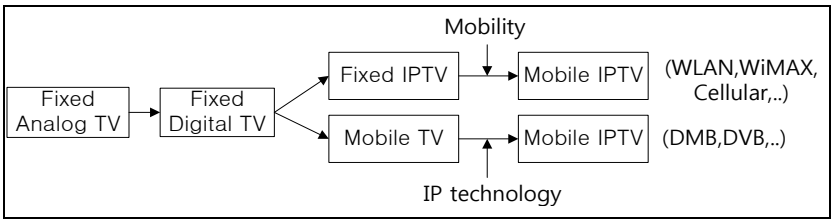


Fig. 1. Mobile IPTV Technology Approaches

In addition to mobile IPTV, in open IPTV, consumers refuse to be passive content users, but instead want to have influence as content providers and choose the contents they want [3]. Open IPTV is one of the application areas of Machine-to-Machine (M2M) communication services, which is a major paradigm shift. As the major standardization organization, the Open IPTV Forum [4] is developing an end-to-end solution to provide personalized IPTV services in a managed or non-managed network.

Because IP networks are open to anyone, IPTV based on IP network has the attacks such as unauthorized access and watching, illegal copy and circulation, and so on. To solve these problems, current broadcasting systems adopt encryption technologies such as the Condition Access System (CAS) [5] and Digital Rights Management (DRM) [6]. However, when mobility is considered, security technologies for traditional IPTV are not proper to adopt. For the security requirements for mobile IPTV, group management for users who subscribe the membership and watch the channel is essential. Fundamental of group management is the group keys.

In mobile IPTV, the users join in and leave the service often while moving. Every time the users join in, they need to be provided the group keys, but they are not supposed to know the previous contents, so the rekeying is required. When they leave the service, the keys need to be rekeyed for the leaving nodes not to get the service any more. This frequent rekeying makes the security vulnerable. Especially, in the open IPTV, where each user can be the service provider, it is much more complicated. If key management system becomes vulnerable due to its poor security, there is possibility that the security of the whole communication system becomes insecure. Other related works have not considered the frequent membership changes or the openness. Therefore, we propose a channel group key management mechanism based on Pre-distribution and local Collaboration-based Group Rekeying (PCGR) and an automatic group key rekeying mechanism considering membership changes and device mobility [7]. In addition, for the

users to communicate for open IPTV, pairwise keys between them need to be established. In this paper, we additionally propose a pairwise key management for efficient mobile open IPTV service. Our contributions are as follows.

- Our proposal basically supports data authentication functionality through eNBs by verifying the information received in the rekeying process.
- By considering the frequency of membership change, our mechanism increases the efficiency of channel group key rekeying with low communication, computation, and storage overhead.
- Device communication for each pair of users who participate in the open IPTV service is also described in our work.

The remainder of this paper is organized as follows. Section 2 describes the related works for IPTV and group key management schemes. We also briefly describe the PCGR which we partly adopt in our mechanism. Section 3 presents the previous proposed group key management mechanism which provides data authentication and automatic rekeying among IPTV users. Open IPTV service between devices is presented in section 4. Section 5 evaluates the effectiveness of advanced mechanism and analyzes the security issues. Finally, we conclude our paper in Section 6.

2. Related Works

Major researches on secure IPTV service are described in this section. In addition, in considering the mobility of devices and group communication security, group key management mechanisms including PCGR that we partially adopted in our work are presented in this section.

2.1. IPTV Security

As IPTV brings a lot changes in industrial and technological aspect, security becomes a key issue to solve for the service. To prevent the unauthorized watching of IPTV, user authentication and access control are required. CAS [5] and DRM [6], the major technologies for IPTV security, are frequently adopted [8]. They differ from each other in terms of how they are applied; however, they also complement each other at the same time. CAS is the core technology for securely transferring content encrypted with a the private key preloaded for each user, and it is used for content protection in traditional digital and satellite TV, as well as IPTV, etc. The structure of CAS is shown in Fig. 2. At the head-end, control word (CW) is used to initialize the generation of a pseudo random sequence number. The pseudo random sequence number is generated by a pseudo random sequence generator for scrambling and descrambling of video programs. The CW for each subscriber is encrypted with the authorization key (AK) of the corresponding channel and

the encrypted CW forms an entitlement control message (ECM). The AK is also encrypted using the private key (DK) and the encrypted AK forms an entitlement management message (EMM). The ECM, EMM and the scrambled program are re-multiplexed in a new transport stream which is broadcast in the form of a radio frequency signal. The subscriber management system is used to administer the issue of or update of the smart card for a subscriber, which contains the DK and other account information. At the receiver end, the receiver can descramble the program according to the reverse steps of the head-end with the cooperation of the smart card and Set-Top-Box (STB).

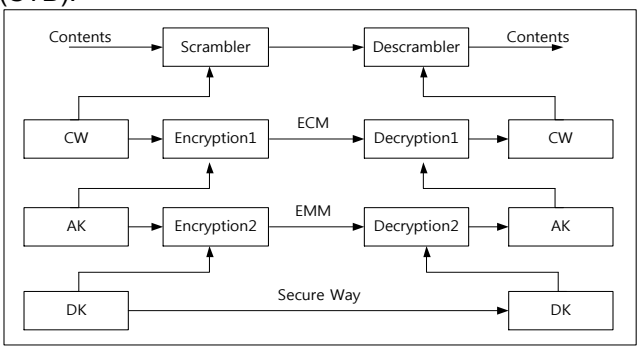


Fig. 2. Traditional CAS System

DRM [6] is a technology designed to prevent the unauthorized use and duplication of digital media. Technically, this technology is based on the encryption of the data. The key used for decryption is itself encrypted and bundled with the permissions. The encrypted data and corresponding licenses are typically associated with each other using unique content identifiers. In order for a single receiver to access and use the license, and thus the content key and content, the encrypted license should be known only to the sender and the intended receiver. This can conveniently be done by deploying a public-key infrastructure and surrounding trust system.

Several researchers [9], [10], [11] have considered the security problem for IPTV services. However, most of them deal with traditional IPTV which is static and cannot be applied for mobile IPTV services. For open IPTV network security, in our previous work, we proposed a secure user authentication and key distribution based on Kerberos for open IPTV security. We also proposed a contents sharing mechanism in home network [3].

2.2. Group Key Management Mechanisms

Group key management has been researched a lot, and the mechanisms can be classified into three categories.

In centralized key management schemes, a group manager generates group keys and distributes the key to authenticated group members and

manages key material and lists. Blundo, C.et al. proposed a mechanism in which a server chooses a t-degree polynomial randomly and distributes them to neighbor nodes and the member nodes substitute the polynomial with their IDs; hence, all the nodes share one group key [12]. Wang, Y. and Ramamurthy, B. proposed four safe group communication methods [13]. Information for group key rekeying is unicasted to each node. This creates a heavy overload when group size grows. Broadcasting is proposed to solve the overhead problem. The broadcasting mechanism requires heavier overhead when groups are generated; however, rekeying cost is relatively low. Overlapping is also proposed to prevent flooding attack. Finally, group information predistribution minimizes group generation time. Karuturi, N.N.et al. provide a generalized framework for centralized GKM along with a formal security model and definitions for the security properties that dynamic groups demand. A lot of researches have been done for centralized group key management. However, in mobile communication environment, parent-child relationship changes constantly because of devices movements. Even if centralized management is very stable and secure, it is not proper for adopting in mobile network.

In distributed key management, multiple key managers generate group keys and distribute them to authentic members. Zhang, W. and Cao, G. proposed a mechanism (PCGR) that predistributes key related information and generates group keys[14]. When group key rekeying is required, nodes cooperate and a new group key is computed. This scheme is applied in our proposal and will be more described in subsection 2.3. Huang, J. H.et al. proposed a level key infrastructure for multicast and group communication that uses level keys to provide an infrastructure that lowers the cost of nodes joining and leaving [15]. This scheme has a drawback in that process delay increases even when many nodes are changed. Zhu, S. et al. proposed a key management protocol for sensor network designed to support in-network processing, while at the same time restricting the security impact of a compromised node [16]. This mechanism is safer, because it uses four different kinds of keys. However, key update consumes much overhead. Adusumilli, P., Zou, X. and Ramamurthy, B. proposed a Distributed Group Key Distribution (DGKD) protocol which does not require existence of central trusted entities such as group controller or subgroup controllers [17]. Aparna, R. and Amberker, B.B. proposed a key management scheme for managing multiple groups. They uses a combination of key-based and secret share-based approach for managing the keys and showed that it is possible for members belonging to two or more groups to derive the group keys with less storage [18]. Kim, Y., Perrig, A, and Tsudik, G. investigated a novel group key agreement approach which blends key trees with Diffie-Hellman key exchange [19]. It yielded a secure protocol suite called Tree-based Group Diffie-Hellman (TGDH) that is both simple and fault-tolerant.

Contributed management mechanisms rekey the group keys through nodes' cooperation without specific key managers. Yu, Z. and Guan,Y. propose a group key management mechanism [20] in which basic matrix G and secret matrices A,B are assigned to each sensor node; each matrix is

used to generate group keys among nodes in the same groups and different groups, respectively. The advantage of this mechanism is that the probability of generating group keys is high. However, when the grid size is large, much energy is wasted and when the grid size is small, group keys may not be generated.

2.3. PCGR

This scheme was designed based on the idea that future group keys are generated by neighbors that can collaborate to protect the communication and appropriately use the preloaded keys [14].A detailed description is provided, since our proposal partly adopts this scheme.

Setup server constructs a unique univariate t-degree g-polynomial g(x), and g(0) is the initial group key (Fig.3(a)). After a device has been deployed and has discovered its neighbors, it randomly picks a bivariate e-polynomial and generates g'-polynomial (Fig.3(b)). The encryption polynomial is generated as follows.

$$e_u(c,y) = \sum_{j=0}^{\mu} B_j y^j \quad (1)$$

Encryption is conducted as,

$$g'(x) = g(x) + e_u(x,u). \quad (2)$$

After distributing the shares of $e_u(x, y)$ to its n neighbors as in Fig. 3(c), N_u removes $e_u(x,y)$ and g(x), but keeps g'(x). Fig. 3(d) illustrates the final distribution of g'(x) and $e_u(x,v_i)$.

Every device maintains a timer for rekeying. When the time expires, each innocent device N_u increases its c by one, and returns share $e_{vi}(c,u)$to each trusted neighbor, N_{vi}. Meanwhile, as shown in Fig.3 (e), N_u receives a share $e_u(c, v_i)$ from each trusted neighbor N_{vi}. Having received μ+1 shares, N_u can reconstruct a unique μ-degree polynomial as

$$\sum_{j=0}^{\mu} (v_i)^j B_j = e_u(c,v_i)(0 \le i \le \mu). \quad (3)$$

Finally, N_u computes the new group key g(c) = g'(c) - $e_u(c,u)$ as in Fig.3(f).This scheme has the advantage that even if some devices are attacked, the new group key is not revealed. However, major drawback of this scheme is that any node in the network can initiate the group key rekeying, causing heavy overhead.

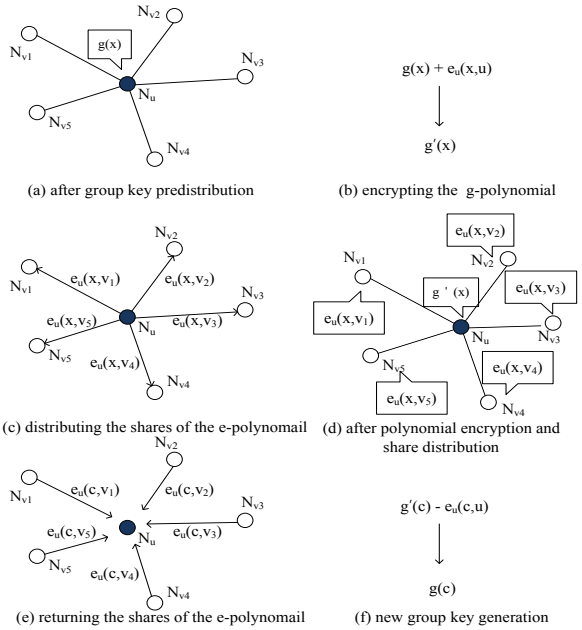


Fig. 3. PCGR: Polynomial Encryption, Share Distribution, and Key Updating

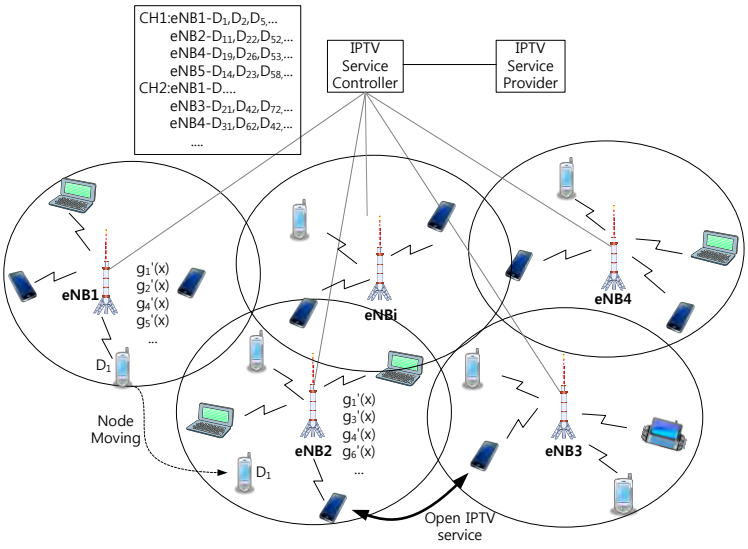


Fig. 4. System Environment including the Device Movement and Open IPTV Service of Our Proposal

3. Proposed Key Management for Secure Mobile IPTV Services

3.1. System Architecture and Basic Assumptions

Based on the cellular network where devices receive their data through eNBs, mobile devices are provided IPTV services through eNBs and ISC in our proposal. Devices watching the same channel share a group key which is used to encrypt the contents delivered through the eNB, which means individual key is assigned for each channel. Basic key materials (polynomial coefficient values) are assigned to eNBs and devices according to the rekeying cycle. Group keys are shared among devices which receive the IPTV service.

Before each device belongs to its own eNB, pairwise keys between the eNB and the device are preassigned. Routing is not considered in our work. As in fig. 4, in our proposal, devices move from one cell to another, and devices can communicate with each other when they subscribe in the open IPTV service.

3.2. Group Key Initialization

For IPTV service, there are many channels for the users to select, and the contents delivered through the channel need to be secured. We define the devices which subscribe and receive the contents from a channel as a group. For each group, group keys for encrypting the contents are required. The most important issue here is how to generate group keys and how to update them efficiently for secure IPTV service. Rekeying is required according not only to the rekeying cycle but also to the membership change. We also need to consider the members mobility. For these objectives, we partially adopt the PCGR for group key generation and rekeying for securing the contents.

Each channel requires individual encryption key for securing the IPTV contents. Because of device mobility, CAS is not proper for securing the contents because it is designed to be installed in STB for traditional IPTV service. We adopted a part of basic PCGR and modified it for channel key generation and rekeying when required. ISC generates the channel key polynomials, $g_i(x)$ for channel i and $e_i(x,y)$ for each $g_i(x)$, and distributes the information to each eNB under the channel service. eNBs receive as many g(x)s as the number of channels that the members of eNB belong to. For each channel i, ISC also generates $e_i(x,y)$ as

$$e_i(x,y) = a_i(x,y) \times d_i(x,y) + q_i(x,y). \quad (4)$$

With encryption polynomials as above, eNBs can verify the shares from devices to filter the false shares and decide which device is illegally receiving the IPTV service. Using the e-polynomial (i.e., $e_i(x,y)$), eNB encrypts the g-polynomial (i.e., $g_i(x)$) to get its g' polynomial (denoted as $g_i'(x)$). The encryption can be conducted as follows:

$$g_i'(x) = g_i(x) + e_i(x, i) \quad (5)$$

After receiving $g_i(x)$ and $e_i(x, i)$, eNB sends $g_i(0)$ to the member nodes. Next, eNB distributes the shares of $e_i(x, y)$ to its member devices D_{v_i} (i= 0, ⋯ , n−1). Specifically, each device D_{v_i} receives share $e_i(x, v_i)$. eNB unicasts this message to each device, including the individual encryption polynomial, $d_u(x)$, and $q_u(x)$, after encrypting the message with $g_i(0)$.

$eNB_i \Rightarrow D_v$: $E_{g_i(0)}\{e_i(x,ID_{D_v})\}$

(1≤v≤n, n is the number of devices in the group)

After transmission, eNB removes $e_i(x,y)$ and $a_i(x,y)$ that has been used to generate $e_i(x,y)$ for security, but keeps $g_i'(x)$. After group initialization, the following information is retained.

eNB : $g_i'(x)$, $d_i(x,y)$, $q_i(x,y)$

Device v: $e_i(x,ID_{D_v})$

Fig.5 shows the group key initialization processes.

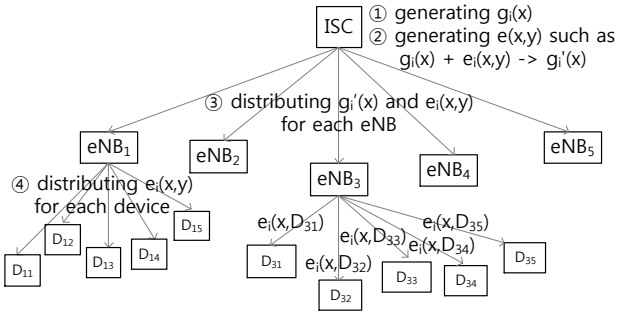


Fig. 5. Group Key Initialization Flow among ISC, eNB, and Devices

3.3. Group Key Update on Rekeying Cycle

Channel group keys are periodically renewed according to the following processes on rekeying cycle.

(1) On rekeying time, ISC sends the group key rekeying command to eNBs which have subscribers of the channel i.

(2) eNB sends out the request_share message with random number asking group key shares of $e_i(x,y)$ to its member devices.

(3) Devices receiving this message reply with the result value after computing the encryption polynomial. The value is encrypted with current group key $g_i(c)$ after being computed by substituting x with random number r, y with the ID of the device, ID_{Dv}.

$D_v \Rightarrow$eNB:$E_{g_i(c)}\{e_i(r,ID_{Dv})\}$(1≤*v*≤*n*, *n* is the number of devices in the cell)

(4) After receiving the key shares, eNB first verifies the values. Because the encryption polynomial was generated as $a_i(x,y) \times d_i(x,y) + q_i(x,y) = e_i(x,y)$,the return value is verified if $e_i(r,ID_{Dv}) \bmod d_i(r,ID_{Dv}) = q_i(r,ID_{Dv}) \bmod d_i(r,ID_{Dv})$.

If the result is true, eNB considers that the device is authenticated. After gathering μ+1 key shares from devices, $e_i(r, ID_{eNB})$ is computed and a new group key is generated, as follows.

$$g_i(r) = g_i'(r) - e_i(r, ID_{eNB}) \quad (6)$$

If $g_i(x)$ is a t-degree polynomial, at least t+1 key shares from neighbor devices are needed to compute the new group key, $g_i(r)$.

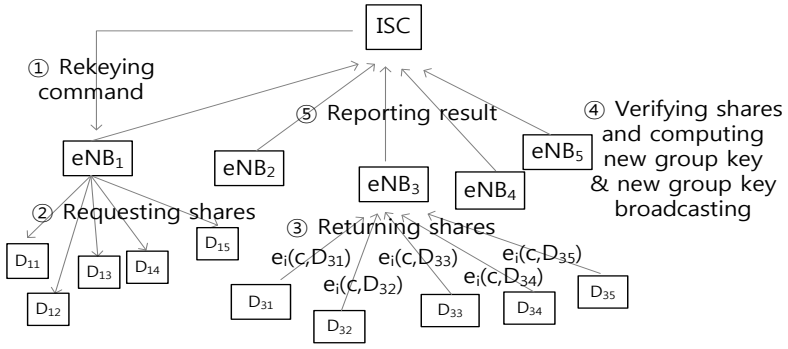


Fig. 6. Group Key Rekeying for All Devices

(5) eNB broadcasts new group key $g_i(r)$ after encrypting it with current group key $g_i(c)$ to its member devices in cell *i*.

eNB⇒ D_v: $E_{g_i(c)}\{g_i(r)\}$

(1≤*v*≤*n*, *n* is the number of devices under the channel service)

When a device fails to verify itself, it may be assumed to be an illegal watcher. The eNB notifies this to ISC to recheck the subscription. If it is illegal, another rekeying is induced. The result of rekeying whether it is successful or not is encrypted with the pairwise key between the eNB and the ISC and delivered to ISC. Fig.6 depicts the processes.

Terminologies for our proposal are described in Table 1.

Table 1.Terminologies for Our Proposal

Terminologies	Description
Di	Device node i
ID_{Dv}	ID of device D_v
$g_i(x)$	Group key polynomial for channel *i*
$e_i(x,y)$	Group key encryption polynomial of channel *i*
$g_i'(x)$	Encrypted group key polynomial of channel *i* using $e_i(x,y)$
$a_i(x,y),d_i(x,y),q_i(x,y)$	Polynomials of group *i* for generating $e_i(x,y)$
r	Random number *r*
$g_i(c)$	Group key of channel *i* in current session
$g_i(r)$	New group key of channel *i*
$w_i(x)$	Polynomial to exclude a leaving node
$f_i(r)$	Polynomial made with $w_i(x)$ and $f_i(r)$ for isolating the leaving device
TH	Threshold value for group key rekeying
THstd	Standard TH to start group key rekeying process

3.4. Group Key Rekeying Triggering

When membership changes occur, eNBs report this to ISC to check if group key rekeying is required or not.

Device Leaving from the Service Group When some devices don't want to receive the channel service anymore, group key rekeying is required for forward secrecy, which means leaving device should not get the future contents anymore. eNB notifies member leaving to ISC and then ISC checks if normal rekeying process is required. If normal group key rekeying is not required, temporary group key is adopted. For this, instead of encrypting the new group key with the old group key, ISC generates $f_i(x)$ as follows to isolate the device from the service group.

$$f_i(x) = g_i(x) \times w_i(x), \quad (7)$$

where, $w_i(x)=(x- x_1)(x- x_2)\ldots(x- x_{k-1})(x-ID_{D_x})$

(*k* is the number of devices in group)

D_x is the leaving node, which means that when a leaving device inputs its ID in the formula, $w_i(x)$ becomes zero and the node cannot compute the new group key. Other devices divide $F_i(r)$ by $w_i(ID_{Di})$ and get $f_i(r)$. They can take part in the new group session having obtained this new group key as follows.

$$eNB \Rightarrow D_v: Eg_i(c)\{f_i(x)||w_i(x)\} \quad (8)$$

($1 \le v \le n$, *n* is the number of devices in a cell)

$g_i(r)=f_i(r)/w_i(ID_{D_V})$

New Device Join in the Service Group For backward secrecy, group keys need to be rekeyed when new nodes join the IPTV service group. When eNB reports ISC that a new device will be added to the service group, ISC checks the rekeying condition, and decides whether rekey the whole group key or just adopt temporary key for newly joining users. If the ISC decides the latter one, it prepares a polynomial as in (7) and (8), and sends the newly generated group key and $e_i(x, ID_{new})$ encrypted with the pairwise key between eNB and the new device to individual new joining nodes. When eNB confirms that the new node is authentic one with the help of ISC, it unicasts the new group key and $e_i(x, ID_{new})$ encrypted with the pairwise key between eNB and the new device to the newly joining node. After receiving this information, the new node sends the confirm message encrypted with the new group key to eNB. This message can be decrypted by all original members. They can also confirm the new member has joined the group.

As described in previous subsection, group key rekeying is composed of many steps and could cause serious computation and communication overhead if group key rekeying is started on every membership changes. When some nodes frequently change the subscription or when some nodes just join or leave the service group right after the periodic rekeying, the efficiency is decreased. To deal with this situation, after getting the membership change report from the eNBs, the ISC checks whether normal group key rekeying is required or not. The threshold value for deciding to start rekeying process or not is computed as follows.

$$TH = \alpha \cdot (Acc\_users/Tot\_uers)(1+\beta \cdot (Spent\_Time/Rekeying\_Time)), \quad (9)$$

where Acc_users is the accumulated number of users who have changed their membership by leaving from or joining in the service group, and Tot_users is the number of total users who are subscribing the channel service. Spent_Time is the time since the latest rekeying time, and Rekeying_Time is the normal rekeying time period. It means that more than certain number of users changed their memberships and certain amount of time has spent after the periodic group key rekeying. α and β are the system parameters and can be adjusted between 0 and 1according to system environment. When α is big, the number of membership changing users is more importantly considered, while even if β is big, it cannot trigger group key rekeying if there is no membership change at all. Basically, the number of membership changing users is much more important in normal situations. The overview of our proposed system flow is shown in Fig. 7.

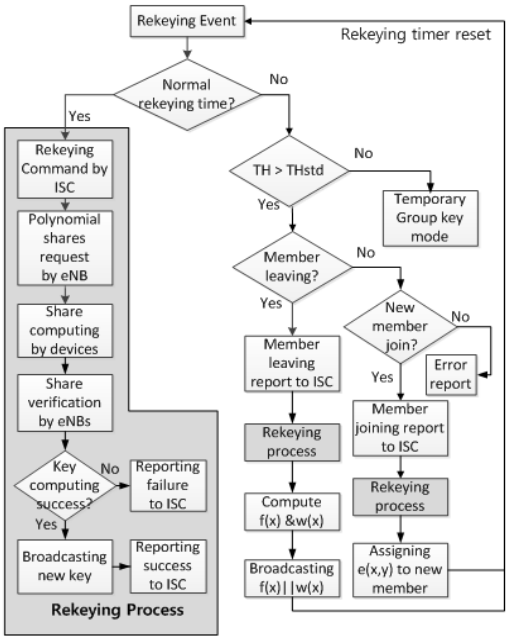


Fig. 7. Proposed Group Key Rekeying System Flows

3.5. Other Considerations

Service Member Mobility Management Different from the traditional IPTV service, in mobile IPTV, devices can move from one cell to another. They still want seamless service while they move. The important point here is that we need to supply them with the same quality of service while providing the new group key even though their locations change. Based on the IPTV systems, handoffs may occur or not. We don't consider the location update process here, and what we want to focus is that the mobile device is still the member of IPTV service, which means we don't need group key rekeying. The device just moved in a cell still has the old polynomial share from the old eNB. The device can join the rekeying process because new group key is encrypted with the old group key if membership change does not occur. At the first group key rekeying time in the new cell, the device gets its own polynomial share from the eNB of the new cell. With the share, the device can contribute its own share in the next rekeying process.

eNB Cooperation for Rekeying When the number of devices that receive the channel service is less than μ+1, the eNB cannot gather enough shares and hence cannot compute new group key for its own cell. In this case, more than one eNB need to cooperate and exchange the shares with each other.

When the cell is isolated and the eNB has difficulty for finding the other eNBs with which it can cooperate, the eNB notifies this to ISC and ISC can send the new group key for the eNB and the related mobile devices.

Group key polynomial update In traditional CAS, AK is regenerated by the system parameter for security purpose. In our system, ISC generates new $g_i(x)$ for each channel i according to the membership changes and the number of subscribers. When membership changes occur often, group key rekeying frequency is influenced more by the member leaving or joining events than by rekeying cycle. And in this situation, the lifetime of $g_i(x)$ for channel i is getting shorter, which means ISC needs to changes the $g(x)$ more often.

4. Key Management for Securing Mobile Open IPTV Services

For mobile open IPTV service, each devices need to subscribe the service not only to receive the contents but also to provide the contents of themselves. For secure communication between the devices, they need pairwise keys with each other. These pairwise keys can be generated by eNBs or by ISC according to the locations of the devices.

Pairwise key establishment between subscribers in the same cell When a device wants to subscribe the open IPTV service, it needs to request the service with the contents list it has for the IPTV Service Provider can manage the contents list. After requesting the service, the device can get the list from the ISP and can provide contents to or receive contents from other devices.

When a device requests some contents from a device in the same cell, this is notified to the eNB, and the eNB generates the pairwise key for the pair of devices and distribute the key encrypted with respective symmetric keys. With this key, the two devices can communicate with each other as in Fig. 8.

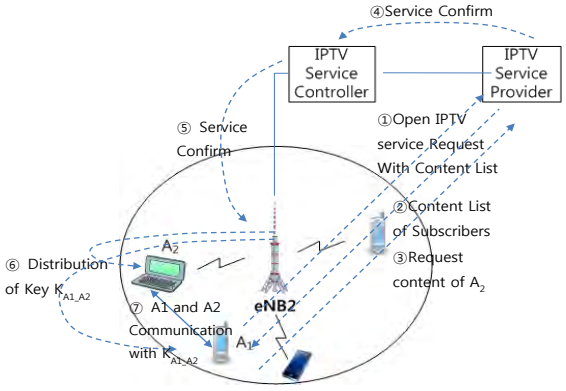


Fig. 8. Pairwise key establishment between devices for open IPTV in the same cell

Pairwise key establishment between subscribers in different cells When a device wants to communicate with the other device from different cell, eNBs cannot generate the pairwise key between them. In this case, ISC generates the pairwise key and distributes it to each eNBs of the respective cells. After getting the pairwise key between two devices, each eNBs encrypt the pairwise keys and redistribute it to each device. The pair can communicate with each other with the key in secure manner. As mentioned in the assumption, routing is out of scope in our work. The steps are as follows.

(1) When a user A_1 wants to get the open IPTV service, which means he or she wants to get any content from the other user, s/he needs to send the request message encrypted with pairwise key between the device and the eNB to the regional eNB, and this message is delivered to ISC and then to ISP. With the request message the contents list of the device can be reported to the ISP for the other users to request the content from the device.

(2) After checking the authenticity of the device, ISP sends the confirm message and the content list it manages to the requesting user.

(3) When A_1 decides some contents from the list, it requests the contents to the ISP. This message is also encrypted with the pairwise key between A_1 and eNB2.

(4) After receiving content lists from A_1, ISP checks if the content holders are in the same cell or not, and delivers the information to ISC.

(5) If they are in the same cells, ISP gives the right to generate pairwise keys to the eNB as in Fig. 8. If they are located in different cells, ISC generates the pairwise keys for A_1 and B_1 and delivers the keys to each eNBs to redistribute them to individual devices. These keys are also encrypted with pairwise keys between eNB and the devices. Finally, the devices can exchange the contents in secure manner. This process is in shown in Fig. 9.

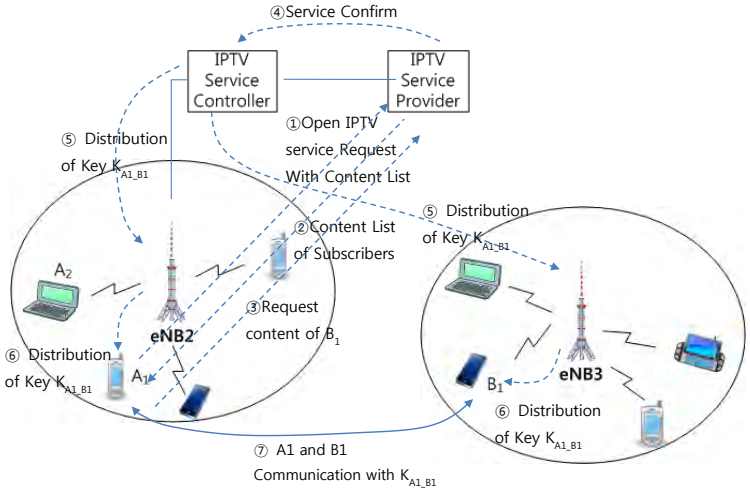


Fig. 9. Pairwise key establishment between devices for open IPTV in different cells

5. Performance Analysis

In this section, we analyze communication, computation, and storage overhead of our proposal. There is no proper existing work that we can compare with in the same structure in the mobile open IPTV domain where group memberships change very often while each pairs of users in the group can communicate with each other, i.e., open IPTV, so we compared our proposal with Blundo's mechanism and the basic PCGR mechanism to show how much ours can decrease the overhead efficiently.

5.1. Overhead Analysis

Communication Overhead Our proposal has less overhead compared to other approaches. In centralized scheme such as LKH [21] or SKDC [22], central controller sends a new key to each trusted node individually. In ours, each eNB computes the new group key and reports this to ISC to confirm. In addition, eNB broadcasts the new group key encrypted with the old group key to further decrease the communication messages. In PCGR, the overhead increases with the number of nodes that distribute the group keys. In our mechanism, the total messages for rekeying is two broadcast messages, one for share request, the other one for new key broadcasting, and one unicast message of each device for sending the key share to respective eNB. Because each device has energy constraint owing to the mobility, decreasing the communication overhead of mobile device is very important for mobile IPTV service. Because two broadcast messages are delivered to all devices, the devices check if the message is for itself or not, and can ignore it when the message is not for itself. Especially, when the number of devices increases, communication overhead in centralized or PCGR rises in accordance with the increasing number of devices, while our proposal only requires as many unicast messages as the number of additional devices no matter how many devices exist. It means that our proposal has advantage in large scale network. In temporary group key method, until normal group key rekeying is triggered, the very small number of messages is required, and this further decreases the communication overhead.

Every node in PCGR has to gather the key shares from neighbor nodes, as well as returning the share of its own to every neighbor node, and every one of them needs to compute the group key for itself. In our proposal, eNBs request key shares to their member nodes periodically or on membership change, and the neighbor nodes reply to this request. After eNB verifies the shares from member nodes, it computes the new group key and rebroadcasts it. When the number increases, our mechanism takes less time than [12] or [14], whose rekeying time increases in proportion to the number of nodes as in Fig. 10. This is very efficient when the scale of the network spans.

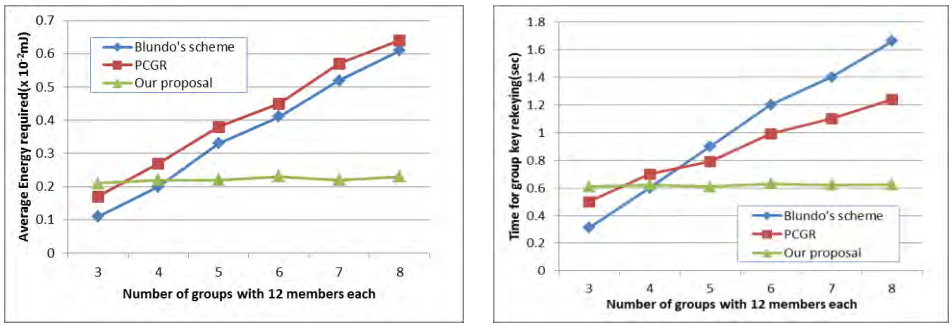


Fig. 10. Group key rekeying overhead (a) Energy consumption (b) Time overhead

In open IPTV, the subscribers communicate with each other for content sharing with or without the help of the other devices such as eNB. Fig. 11(a) shows that the communication time for a pair of users in open IPTV service. Communication time varies according to where each subscriber is located. The overhead when they are located in the same cell is getting shorter, and if they are in their direct M2M communication with each other, it is drastically short. In Fig. 11(b), we can see that the number of packets for rekeying is getting smaller as the rekeying period gets longer, which further decreases the overhead.

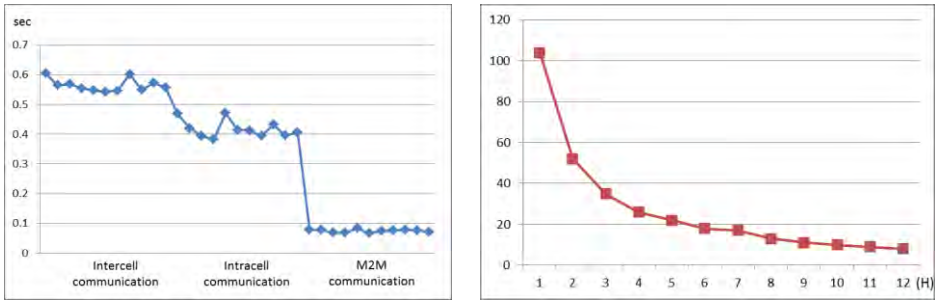


Fig. 11. Communication and rekeying overhead (a) Time for communication between devices as a function of distance (b) Number of packets as a function of group key rekeying frequency in 24 hours.

Computation Overhead eNBs need relatively more computation than that of the devices. eNBs need one decryption and one encryption and key computation for one rekeying process. Decryption is for getting the share from the devices and encryption is required when distributing the new group key. Share verification does not cost much because only simple mod operation is required for each share. Another important cost is for key computation. After receiving $\mu+1$ or more shares, the eNB needs to solve a ($\mu+1$)-variable linear equation system to compute $e(c,u)$, and the computational

complexity of using Gaussian elimination to solve the equation system is $O(\mu^3)$ multiplication/divisions. Devices require only one encryption and one decryption. Encryption is required before sending the share of each device, and decryption is for getting the new computed group key delivered by eNB. For the encryption and decryption, any kind of light encryption/decryption algorithm can be used. Our proposal especially requires less computational overhead for devices, which is proper for mobile IPTV service. Still, using temporary group key, much computation overhead can be decreased at the expense of temporal security degradation.

Storage Overhead In our proposal, each node stores as many e-polynomials as the number of channel groups it is being serviced. Each eNB stores as many *g'(x)* as the number of channels that the devices in its own cell are subscribing to. The number of channels that the eNB needs to support may differ from the number of devices in the cell. If many devices with small number of channels exist, the storage performance degrades. If the length of the coefficient is *L* and the number of channels a device watching is *n*, the node needs *L*n*(t+1)* bits. In the same sense, the storage requirement for an eNB is *L*n'*(t+1)* when *n'* is the number of channels that the eNB needs to relay. In basic PCGR, the node in each group requires *g'(x)*, which is *L*(t+1*) bits, and for shares from the neighbor nodes, it needs *n*L*(t+1)*.When the number of nodes is *N*, and the storage overhead is *N*L*(t+1)(n+1)*.

5.2. Security Analysis

Security Level When temporary group key is adopted, security level becomes temporarily low and these keys cannot be used very long. Because of the energy efficiency, when small number of devices change their memberships, temporary group keys are used as in subsection 3.4. In that case, as membership change ratio increases, the security vulnerability decreases. So THstd setup is very import to keep the security level at moderate level.

Access Control Basically, every node needs to register to get the IPTV service at the initial stage, and all communication is secured using group keys depending on each channel. When membership changes, new group keys are generated and distributed for secure and proper service for authentic users.

Intrusion Detection Security system should detect when devices or eNB are attacked by adversaries. Our proposal can identify if the devices are compromised by verifying the shares. Of course, in PCGR, group key rekeying nodes are not determined and if compromised nodes are requested the secret share, they return the information they just have. However, if in that case, they can be clever enough not to make any response for not being detected by their neighbor nodes (PCGR - selective reply). Then we cannot detect if they are compromised. Our proposal detects the compromised

nodes, because the eNB randomly selects the devices to answer the requests and checks the authenticity of the device as in 3.3. Compromised nodes can be detected much better via our proposal. In our proposal, however, the success ratio also drops less than 80%, when there are more than 40% adversaries, which is not normal situation.

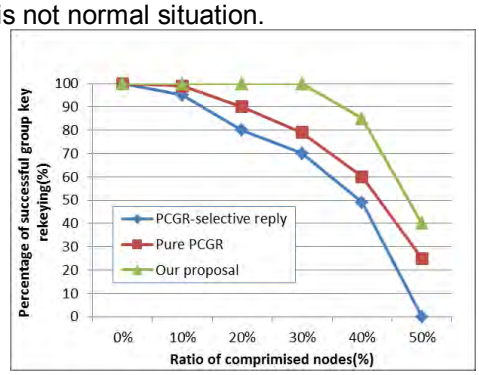


Fig. 12. Percentage of successful group key rekeying

Forward and Backward Secrecy Our proposal guarantees forward and backward secrecy. For backward secrecy, when some devices newly join the channel group, the eNBs report this to ISC; they respond this situation with adopting temporary group keys until normal group rekeying is triggered. For forward secrecy, when a node leaves the group, the temporary group key is also rekeyed, and the leaving node cannot decrypt the messages generated after it leaves.

Availability By filtering the wrong shares from neighbor sensor nodes, clusterheads can get the authentic shares and generate a proper new group key. With this filtering process, we can prevent wrong group key rekeying and wasting unnecessary system resources. In open IPTV, we can further prevent replay attack by adopting timestamp in the packet to protect availability.

Man-In-The-Middle (MITM) Attack In MITM attack, the attacker intercepts the messages between two endpoints and forges or modified them. In our proposal, every pair of devices share pairwise keys with each other, and even the adversaries capture the information in the middle, they cannot forge or modify the data. Even if an attacker captures encrypted data, it is very hard to decrypt them without knowing the pairwise keys.

6. Conclusion

In this paper, we proposed an enhanced group key management mechanism for securing mobile open IPTV. When mobility and openness are added to IPTV technology, key management for traditional IPTV is not proper to apply. Especially, when memberships change often, key updates are required more often. Our proposal basically supports device mobility and membership

changes in providing security to IPTV service. Based on the assigning channel group keys to each channel service in cellular environment and updating the keys considering the membership status and user mobility, we additionally enhance the mechanism considering the group key rekeying conditions based on the threshold. Our proposal also provides secure open IPTV service communication by establishing pairwise keys between devices. For our future work, we are planning to simulate our proposal and additionally analyze the security aspects.

Acknowledgment. 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 (2012R1A1A3019459). We thank Li Shi for her contribution in the experiment.

References

1. A. Pinto, M. Ricardo, “On performance of group key distribution techniques when applied to IPTV services,” Computer Communications, Elsevier, 2011.
2. A. N. El-Kassar, R.A. Haraty, “ElGamal Public-Key Cryptosystem in Multiplicative Groups of Quotient Rings of Polynomials over Finite Fields,” Vol.2, ComSIS, June, 2005.
3. I. Doh, J. Lim, K. Chae, “An Improved Security Approach based on Kerberos for M2M Open IPTV System,” In Proceedings of the NeoFusion, Sep., 2012.
4. M. Cedervall, U. Horn, Y. Hu, I. M. Lvars and T. Nasstrom, Open IPTV forum - Toward an open IPTV standard, Ericsson Review, no.3, 2007.
5. F. K. Tu, C. S. Laih and H. H. Tung, “On key distribution management for conditional access system on pay-TV system,” IEEE Trans. Consumer Electron., vol. 45, no. 1, pp. 151-158, Feb., 1999.
6. F. Hartung, S. Kesici, D. Catrein, “DRM protected dynamic adaptive HTTP streaming,” 2nd annual ACM conference on Multimedia systems, pp. 23–25, Feb., 2011.
7. I. Doh, J. Lim, M. Y. Chung, “Group Key Management for Secure Mobile IPTV Service,” In Proceedings of the IMIS, July, 2012.
8. S. O. Hwang, “Content and service protection for IPTV,” IEEE Trans. Broadcasting, vol. 55, no. 2, June, 2009.
9. D. Proserpio, D. Diaz-Sanchez, F. Almenárez, A. Marín, and R. S. Guerrero, “Achieving IPTV Service Portability through Delegation,” IEEE Trans. Consumer Electron., vol. 57, no. 2, pp.492-498, May, 2011.
10. D. Diaz-Sanchez, A. Marín, F. Almenarez and A. Cortes, “Sharing conditional access modules through the home network for Pay TV Access,” IEEE Trans. Consumer Electron., vol. 55, no. 1, pp.88-96, Feb., 2009.
11. D. Diaz-Sanchez, F. Sanvido, D. Proserpio and A. Marín, “DLNA, DVB-CA and DVB-CPCM integration for commercial content management,” IEEE Trans. Consumer Electron., vol. 56, no. 1, pp.79-87, Feb., 2010.
12. Carlo Blundo, Alfredo De Santis, Amir Herzbeerg, Shay Kutten, Ugo Vaccaro, Moti Yung, “Perfectly-Secure Key Distribution for Dynamic Conferences,” Information and Computation, 1995.

13. Y. Wang, B. Ramamurthy, Y. Xue, “Group Rekeying Schemes for Secure Group Communication in Wireless Sensor Networks,” In Proceedings of the IEEE International Conference on Communications 2007.
14. W. Zhang, G. Cao, “Group Rekeying for Filtering False Data in Sensor Networks: A Predistribution and Local Collaboration Based Approach,” IEEE Infocom 2005.
15. J. H. Huang, J. Buckingham, R. Han, “A Level Key Infrastructure for Secure and Efficient Group Communication in Wireless Sensor Networks,” In Proceedings of the International Conference on Security and Privacy for Emerging Areas in Communications Networks 2005.
16. S. Zhu, S. Setia, S. Jahodia, “LEAP+: Efficient Security Mechanisms for Large-Scale Distributed Sensor Networks,” ACM Transactions on Sensor Networks 2006.
17. P. Adusumilli, X. Zou, B. Ramamurthy, “DGKD: Distributed Group Key Distribution with Authentication Capability,” In Proceedings of the IEEE Workshop on Information Assurance and Security 2005.
18. R. Aparna, B.B. Amberker, “Key management scheme for multiple simultaneous secure group communication,” In Proceedings of the IEEE Internet Multimedia Services Architecture and Applications (IMSAA) 2009.
19. Y. Kim, A. Perrig, G. Tsudik, “Tree-based group key agreement,” ACM Transactions on Information and System Security (TISSEC) 2004.
20. Z. Yu, Y. Guan, “A Robust Group-based Key Management Scheme for Wireless Sensor Networks,” In Proceedings of the IEEE Communications Society 2005.
21. D. Wallner, E. Harder, and R. Agee, “Key Management for Multicast: Issues and Architectures,” RFC 2627, June, 1999.
22. H. Hugh, C. Muckenhirn, and T. Rivers, “Group Key Management Protocol Architecture,” RFC 2093, Internet Engineering Task Force, Mar., 1997.

Inshil Doh received the B.S. and M.S. degrees in Computer Science at Ewha Womans University, Korea, in 1993 and 1995, respectively, and received the Ph.D. degree in Computer Science and Engineering from Ewha Womans University in 2007. From 1995-1998, she worked in Samsung SDS of Korea to develop a marketing system. She was a research professor of Ewha Womans University in 2009~2010 and of Sungkyunkwan University in 2011. She is currently an assistant professor of Computer Science and Engineering at Ewha Womans University, Seoul. Her research interests include wireless network, sensor network security, and M2M network security.

Jiyoung Lim received the B.S. and M.S degrees in Computer Science at Ewha Womans University, Korea, in 1994 and 1996, respectively and received the Ph.D. degree in Computer Science and Engineering from Ewha Womans University in 2001. She is currently an associate professor of Computer Software at Korean Bible University, Seoul, Korea. Her research interests include wireless/sensor network security, and M2M network security.

Kijoon Chae received the B.S. degree in mathematics from Yonsei University in 1982, an M.S. degree in computer science from Syracuse University in 1984, and a Ph.D degree in Electrical and computer engineering from North Carolina State University in 1990. He is currently a professor of Computer Science and Engineering at Ewha Womans University, Seoul, Korea. His research interests include network security, sensor network, network protocol design and performance evaluation.

Received: September 17, 2012; Accepted: March 22, 2013

DOI: 10.2298/CSIS120921038B

Benefiting From the Community Structure in Opportunistic Forwarding

Bing Bai[1], Zhenqian Feng[1], Baokang Zhao[2], Jinshu Su[2]

Department of Computer,
National University of Defense Technology,
Changsha, China
[1]{nudt.bb, fengzhenqian1983}@gmail.com, [2]{zbk, sjs}@nudt.edu.cn

Abstract. In Delay Tolerant Networks (DTNs), an end-to-end connectivity cannot be assumed for node mobility and lack of infrastructure. Due to the uncertainty in nodal mobility, routing in DTNs becomes a challenging problem. To cope with this, many researchers proposed opportunistic routing algorithms based on some utilities. However, these simple metrics may only capture one facet of the single node's mobility process, which cannot reflect the inherent structure of the networks well. Recently, some researchers introduce the Complex network analysis (CNA) to formulate and predict the future contact in DTNs. The community structure is one of the most important properties of CNA. And it reveals the inherent structure of the complex network. In this paper, we present a community-based single-copy forwarding protocol for DTNs routing, which efficiently utilizes the community structure to improve the forwarding efficiency. Simulation results are presented to support the effectiveness of our scheme.

Keywords: Social Network, Forwarding, Delay Tolerant Network, Community

1. Introduction

In Delay Tolerant Networks (DTNs) [1], an end-to-end connectivity cannot be assumed for node mobility and lack of infrastructure. In such environment, two nodes can transmit messages between each other only when they are in contact (i.e., move within transmission range). Due to the uncertainty in nodal mobility, routing in DTNs becomes a challenging problem. To cope with this, many researchers proposed opportunistic routing algorithms[2][3], in which messages are forwarded between mobile nodes opportunistically upon contacts; and a relay selection is determined separately in each hop, aiming to get higher delivery probability.

To cope with the inherent unpredictability of future contact opportunities, many protocols[4][5] forward multiple copies of the same messages to achieve short latency and high delivery probability. However, many studies[6]

have shown that node mobility in DTNs is not entirely random, and instead, some patterns can be found more or less. Based on this observation, some researchers proposed utility-based routing protocols[7][33][34], in which messages are forwarded to the nodes with higher probability to deliver it to the destination. In the utility-based routing protocols, many methods of the utility computing have been proposed [8][9][10]. Among them, a number of schemes implicitly utilize the social properties of nodes. For example [11] uses time of last encounter and [12] uses contact frequency to make the prediction of the future of the network, both of which are the social properties called *similarity* in fact. However, these simple metrics may only capture one facet of the single node's mobility process, which cannot reflect the inherent structure of the networks well.

Recently, some researchers introduce the Complex network analysis[14] (CNA) to formulate and predict the future contact in DTNs. For example, SimBet[16] uses the combination of nodes' centrality and similarity as the utility to conduct the forwarding of messages. And BubbleRap[15], defines the node's social properties as their rankings to determine the forwarding priority of nodes. The community structure is one of the most important properties of CAN. And it reveals the inherent structure of the complex network. In this paper, we present a community-based single-copy forwarding protocol for DTNs routing, which efficiently utilizes the community structure to improve the forwarding efficiency.

To utilize the inherent community structure of DTNs, we decompose the problem into four steps:

1) *Mapping of contacts to social graphs;* In DTNs, nodes usually have the knowledge of their contacts, also called encounter history, which can be depicted by a series of time-node couple. Our work is utilizing these encounter history to generate a social graph, in which the vertexes denote the nodes and the weighted edges represent the encounter history of node pairs.

2) *Detecting the communities on social graphs*; Many studies have been done for the community detecting in social network[28][29][30][31][38]. And the new community detection algorithm in DTNs has also been proposed[35]. It is our future research direction. So it is not discussed in this paper. We use Newman's community detection algorithm[30] in our simulation. The algorithm is offline and the community information is distributed by the profiles.

3) *Virtualizing social graphs to community graphs*; To make the community structure simple from the node view, we define community graph, in which the vertexes denote the communities and the weighted edges represent the encounter history of community pairs.

4) *Routing with the community information;* Our routing protocol is based on the community graph. So it includes inter- and intra- community routing.

Our approach is based on the weighted network model for DTNs. Although the unweighted network model has been discussed in DTNs[17], we believe that the edges with different numeric can reflect the nodes' relation better than the ones with only 0 or 1.

The rest of this paper is as follows. Section 2 introduces the related work of current DTNs routing protocols and community detection in weighted networks. Section 3 proposes our community-based routing protocol. Section 4 discusses our simulation method and results. Finally, Section 5 presents conclusions and future research directions.

2. Related Work

In recent years, a lot of routing protocols have been proposed to cope with the challenge environment in DTNs. According to the number of copies of each message that can coexist in the network, current DTNs routing protocols can be classified into two categories: single-copy routing and multi-copy routing.

The single-copy routing schemes keep only one copy of a message in the network. The simplest case of such schemes is that the source node holds the message and forwards it only to its destination. This scheme obviously has minimal overhead, but the delivery delay of a message could be unbounded [19]. Researchers proposed many schemes [2][7][26][27] to forward the message to its destination by intermediate nodes. Usually, the forwarding decisions are made according to the estimation of the node candidates. In [2], four knowledge oracles are defines to represent the amount of knowledge about the network topology. And for different oracles available, the authors present corresponding routing. The lack of this approach is that each node must know the accurate oracle. To overcome this weakness, [7] proposes minimal estimated expected delay (MEED) routing, which computes the expected delay only using the observed contact history instead of the knowledge oracles. Also utilizing the expected delay to make the local forwarding decisions, T. Spyropoulos et al. [20] give analysis of the random walk model and the estimation function of expected delay based on the distance between nodes.

In contrast, the multi-copy routing protocols may generate multiple copies of each message that can be forwarded to increase the message delivery rate. As mentioned before, epidemic routing is the straightforward idea of this case but with huge resource consumptions. Some researchers use history or predication-based approaches to reduce the number of copies spreading in the networks [8][13][21][22]. PROPHET [21] is a probabilistic routing protocol which defines a delivery predictability metric, reflecting the history of node encounters and transitive and time dependent properties of that relation. Another method of multi-copy routing, also called quota-based routing [10], is limiting the number of copies of each message that can be spread in the networks when message is created. Different algorithms have been introduced to efficiently forward the limited copies. Spray and Wait routing [5] is one of the most famous quota-based protocols in which the message holder forwards half copies to the encountered node per contact until one copy left, and delivers the left copy only to the destination. [10] uses the

average encounter times to distribute the copies between nodes. [9] models the message forwarding as an optimal stopping rule problem and proposes a variation of quota-based routing, hop-count-limited forwarding, to maximize the expected delivery rate while satisfying the hop count limited condition. Erasure coding techniques have also been proposed for DTNs routing[23][24]. The basic idea of erasure coding is to encode an original message into a large number of coding blocks. Once sufficiently large subset of the generated code blocks are received, the original message can be successfully decoded.

Recently, researchers have found that in many applications communication devices are taken by human beings, and so conforming to the characteristics of social networks. So several social network metrics, which are measured based on nodes' direct or indirect observed encounters, are used to guide the packet forwarding in [15][16][25]. SimBet Routing [16] introduces the ego-centric centrality and similarity in social network to guide the DTNs routing. However, these metrics may only capture the single node's mobility process, which cannot reflect the inherent structure of the whole networks well. LocalCom [37] is a social-based epidemic routing algorithm, in which messages are flooding inter-communities but forwarding intra-communities. In this paper, we propose a community based single-copy routing algorithm in DTNs, which utilizes the community structure to improve the forwarding efficiency

3. Community-Based Routing (CBR)

As described in section 1, the CBR protocol, which utilizes the inherent social community structure to facilitate packet forwarding in DTNs, has four main steps: mapping of contacts to social graphs, detecting the communities on social graphs, virtualizing social graphs to community graphs, and routing with the community information.

For the community detection is not our contribution, we just introduce three steps of our scheme without the second step (detecting the communities on social graphs).

3.1. Mapping of contacts to Social Graphs

To map the node contact history to a social graph, we first need to determine the meaning of the weight of each edge in the graph. Similar as [9], we use the meeting probability of two nodes as the weight of the edge between them in the graph.

To calculate the meeting probability[9], we use a discrete residual time-to-live Tr for each message, with time-slot size U. Tr is a measurement in clock time. Let T_{max} be the maximum possible time-to-live of any message, the range of Tr is between 0 and T_{max}/U. Our delivery probability metric is a

function of Tr, and it is calculated using an inductive method. The amount of computation for our delivery probability metric is inversely proportional to the length of U, but its accuracy decreases as U increases. In each time-slot Tr, a node can either meet or not meet another node. A node has the probability to meet several other nodes during the same time-slot, and we simply assume that all meetings start at the beginning of some time-slot. This assumption holds when U is smaller than any meeting duration, and we truncate all meeting durations so that the starting time of them are aligned in the beginning of their respective time-slots. The meeting probability of two nodes in any time-slot of length U is estimated under the assumption of exponential inter-meeting time by

$$p_{i,j} = 1 - \exp(-\frac{U}{I_{i,j}}) \ . \tag{1}$$

Where U is the length of time-slot, and $I_{i,j}$ is the mean inter-meeting time between node i and node j. The Newman's community detection algorithm require the weight of edges is integer, so what we use as the weight of edges is

$$w_{i,j} = \lfloor C \times \ln(p_{i,j}) \rfloor \ . \tag{2}$$

where C is the *integralization* factor.

Note that the calculation of $p_{i,j}$ itself does not rely on the assumption of exponential inter-meeting times. Using a particular estimation that is more realistic for a network in question should result in better routing performance.

3.2. Virtualizing social graphs to community graphs

For we skip the second step of scheme introduce, so we assume that the community detection has been finished now.

Since the community information has been known by each node, we can simplify the social graphs. Based on the social graphs, whose communities have been detected, we further define the community graph, in which the vertexes denote the communities and the weighted edges represent the encounter history of community pairs. By this way, each node can get very simple view of the whole network, in which only the node in the same community with it and the other communities. Here, the other communities are treated as "big" nodes.

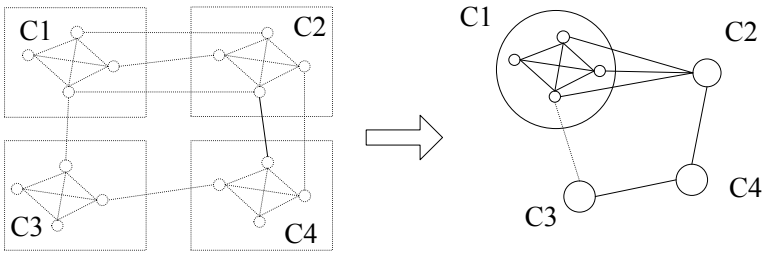


Fig. 1. Virtualizing social graphs to community graphs

Figure 1 gives an example. The left graph is the social graphs, whose communities have been detected. And the right one is the community graph for each node in community 1 (C1).

A. Aggregate the edges between community pairs

In figure1, we can see that in the social graph, there are two edges between C2 and C4, while in the community one, it reduces to one edge. To relief the depression of precision caused by the edge aggregation, we recalculate the weight of the aggregated edges between communities by

$$w_{C_m,C_n} = \lfloor C \times \ln(P_{C_m,C_n}) \rfloor . \tag{3}$$

where Cm , Cn are adjacent communities, C is still the *integralization factor*, and P_{C_m,C_n} is

$$P_{C_m,C_n} = 1 - \prod_{i \in C_m, j \in C_n} (1 - p_{i,j}) . \tag{4}$$

where P_{C_m,C_n} represents the probability of any node in Cm meet any node in Cn based on the contact history.

Thus, we have finished the aggregation of the edges between community pairs. But treat communities Cm and Cn as normal node is still not proper. This issue will be disused in next part.

B. Evaluate the intra-closeness of each community

As described above, treat communities Cm and Cn as normal node is not proper, for the intra-community cost has been ignored.

Although many studies has shown that the nodes in same community is “closer” than the outsiders, its cost is still cannot be ignored especially in DTNs. So we have to evaluate the intra-closeness of each community. We use the expectation of the closeness between each node-pair in a community, which is the average longest path length

$$w_m = \frac{1}{\frac{1}{2} N_m (N_m - 1)} \sum_{i,j \in C_m, i>j} d_{i,j} \quad . \tag{5}$$

where Nm is the node number in the community Cm, $d_{i,j}$ is the longest path length between node i and j in Cm.

Thus, we have finished the virtualization of social graphs to community graphs, both the inter- and intra- community edges. Based on this, we can design our CBR protocol.

3.3. Routing with the community information

Based on the above work, we propose our community based routing protocol. The key problem of routing is how to select the next relay. Here, we define a Closeness function to help make decision.

As shown in Algorithm 1, each message carries the destination node address Dm and its community address Cm. When node i meet node j, for each message Mm held by node i, calculate the *Closeness* value u of i and j according to their community address. If ui<uj, send Mm to node j.

For node j, the progress is the same.

Algorithm 1 community-based routing

Let $N_1, \ldots, N_N$ be nodes
Let $M_1, \ldots, M_M$ be messages
each message carries the destination node address D_m and its community address C_m
On contact between N_i and node N_i :
for every M_m held by N_i **do**
if $C_i = C_m$ **do** u_i <— Closeness (N_i , D_m)
else u_i <— Closeness (N_i , C_m)
end if
if $C_i = C_m$ **do** u_i <— Closeness (N_i, D_m)
else u_i <— Closeness (N_i, C_m)
end if
if $u_i < u_j$ **do**
Send M_m to N_i
end if
end for

4. Evaluation

To evaluate our protocol, CBR, we use the Opportunistic Network Environment simulator (ONE) [19], which is a specifically designed simulation tool for delay tolerant networks. Simulation results show that CBR improves delivery ratio by utilizing the inherent community structure in DTNs.

4.1. Simulation setup

We compare CBR against other routing protocols using the dataset, Haggle project [36]. In Haggle project, about fifty devices were distributed to students attending Infocom 2005 student workshop. And the contacts were logged and provided. We divide the original trace files into discrete sequential contact events as the inputs of the simulator. Each contact record includes the start time, end time, and ID of the nodes in contact.

Before the simulation starts, we first map the contacts to a social graph, using the formula 1 and formula 2 mentioned in section 3.1.Here the *integralization* factor C is 100. And then we use the Newman's community detection algorithm[30] to divide the communities. At the third step, we use the formula 3,4,5 to virtualize the social graph to community graph. After these work, the information calculated (for example, the intra-closeness of each community) is distributed to each node. And all of this is worked offline.

In our simulations, we primarily focused on two parameters:1) *Delivery ratio*: the proportion of packets that arrived at the destination within the delay requirement; 2) *Average Delay*: although it is not considered so important in DTNs; and 3) Goodput[10]: the number of messages delivered divided by the total number of messages transferred (including those transfers that did not result in a delivery).

For each round of simulation, 1000 messages are created, uniformly sourced between all node pairs. The packet size constant at 25KB, and the buffer space constant at 1MB. Besides, we run each simulation 10 times with different random seeds of events creator for statistical confidence.

4.2. Performance Results

To evaluate our CBR protocol, we compared it with three other famous protocols: Epidemic, Prophet, and LocalCom [37]. Epidemic is the extreme case of multi-copy routing protocols, which is just used as a bound. The Prophet is one of the most famous utility-based routing, representing the traditional routing protocols in DTNs. And LocalCom is also a community-based forwarding algorithm, which is however a flooding one.

In this evaluation, we are trying to answer two questions: 1) Is CBR better than the traditional utility-based routing algorithm in DTNs? 2) Can CBR get

better performance than the other existing community-based routing protocols in DTNs.

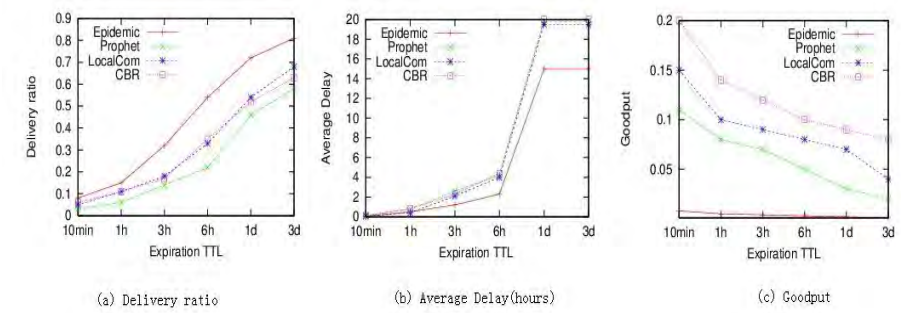


Fig. 2. Simulation results

As shown in Figure 2, CBR reaches higher performance than Prophet in both metrics. Also CBR can reach nearly the same delivery ratio as LocalCom, but better goodput than it. The delivery ratio and cost of the Epidemic scheme represent the upper bound in all three cases. Since the simple flooding scheme utilizes all the possible paths over time to forward the packet, if a path that can satisfy the delay requirement exists, it will be included.

In Fig.2(a), the delivery ratio of CBR is very close to the LocalCom during the whole simulation. And CBR can reach totally higher performance than Prophet. The reason for this is that the CBR utilize the community structure which reflects the inherent property of the whole network, while the Prophet only knows the local information of the nodal movement history.

In Fig.2(b), we can see that compared to LocalCom and Prophet, the average delay of CBR is very close. And it is very amazing that when the Expiration TTL of each message is increased from 1 day to 3 days, the average delay of all the four protocols did not change. The reason for this is that some of the messages sent from source nodes are inherently unreachable. For these messages, the increasing of TTL is useless.

In Fig.2(c), the Goodput of CBR is higher than Prophet and LocalCom. The reason of this is that LocalCom is based on community level broadcast and Prophet is multi-copy scheme, while CBR is a single-copy routing scheme. So in the same time, there is only one copy of each messages in networks in CBR. But there may be multiple copies in the other two schemes.

In summation, CBR outperforms the Prophet in terms of delivery ratio, and goodput, while the average delay is close. Although the delivery ratio of CBR is sometimes not better than LocalCom, the gap is acceptable. And the goodput of CBR is clearly outperforms the LocalCom. So CBR can reach better performance than both the traditional utility-based routing algorithms and the other existing community-based routing protocols

5. Conclusions and future work

Social network properties are observed in many DTNs and tend to be stable over time. In this paper, we seek to utilize the community structure, which is based on social network properties, to improve routing performance. We define the social graph and community graph based on nodes' encounter history to depict the neighboring relationship between nodes. The simulation based on real mobility traces shows that CBR can get good performance. In the future, we plan to study the multi-copy routing based on the community structure and the community detection in DTNs.

References

1. V. Cerf, S. Burleigh, A. Hooke, L. Torgerson,R. Durst, K. Scott, K. Fall, and H. Weiss. Delay Tolerant Networking Architecture. In Internet draft: draft-irrf-dtnrg-arch.txt, DTN Research Group. (2006)
2. S. Jain, K. Fall, and R. Patra. Routing in a Delay Tolerant Network. In Proc. ACM Sigcomm. (2004)
3. L. Pelusi, A. Passarella, and M. Conti, "Opportunistic networking: Data forwarding in disconnected mobile ad hoc networks," IEEE Communications Magazine. (2006)
4. A. Vahdat and D. Becker. Epidemic routing for partially connected ad hoc networks. Technical Report CS-200006, Duke University. (2000)
5. T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Spray and wait: An efficient routing scheme for intermittently connected mobile networks. In Proceedings of WDTN. (2005)
6. A. Chaintreau et al. Impact of human mobility on the design of opportunistic forwarding algorithms. In Proc. of INFOCOM. (2006)
7. E. P. C. Jones, L. Li, and P. A. S. Ward. Practical Routing for Delay Tolerant Networks. In SIGCOMM DTN Workshop. (2005)
8. J. Burgess, B. Gallagher, D. Jensen, and B. N. Levine .MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networking. In Proc. of IEEE INFOCOM. (2006)
9. C. Liu and J. Wu. An Optimal Probabilistic Forwarding Protocol in Delay Tolerant Networks. In Proc. of ACM MobiHoc. (2009)
10. S. C. Nelson, M. Bakht, and R. Kravets. Encounter-Based Routing in DTNs. In INFOCOM. (2009)
11. H. Dubois-Ferriere, M. Grossglauser, and M. Vetterli, "Age matters:efficient route discovery in mobile ad hoc networks using encounter ages," in ACM MobiHoc. (2003)
12. V. Erramilli, A. Chaintreau, M. Crovella, and C. Diot, "Diversity of forwarding paths in pocket switched networks," in IMC. (2007)
13. V. Erramilli, M. Crovella, A. Chaintreau, and C. Diot, "Delegation forwarding," in ACM MobiHoc. (2008)
14. M. E. J. Newman, "The structure and function of complex networks," SIAM Review, vol. 45, pp. 167–256. (2003)
15. P. Hui, J. Crowcroft, and E. Yoneki. BUBBLE Rap:Social-based Forwarding in Delay Tolerant Networks.In Proc. of ACM MobiHoc. (2008)

16. D. Elizabeth and H. Mads. Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs.In Proc. of ACM MobiHoc. (2007)
17. T. Hossmann, T. Spyropoulos, and F. Legendre, “Know thy neighbor:Towards optimal mapping of contacts to social graphs for DTN routing,” in IEEE Infocom 2010. (2010)
18. A. K, J. Ott, and T. K, The one simulator for dtn protocol evaluation. In SIMUtools. (2009)
19. M. Grossglauser and D. Tse. Mobility Increases the Capacity of Ad Hoc Wireless Networks. In Proc. of IEEE INFOCOM. (2001)
20. T. Spyropoulos, K. Psounis, and C. S. Raghavendra. Single-copy Routing in Intermittently Connected Mobile Networks. In Proc. of IEEE SECON. (2004)
21. A. Lindgreny, A. Doria, and O. Schel\'en. Probabilistic Routing in Intermittently Connected Networks, Mobile Comp. and Commun. Rev., vol.7, no. 3. (2003)
22. B. Burns, O. Brock Brian, N. Levine. MV Routing and Capacity Building in Disruption Tolerant Networks. In Proc. of IEEE INFOCOM. (2005)
23. S. Jain, M. Demmer, R. Patra, K. Fall. Using Redundancy to Cope with Failures in a Delay Tolerant Network. In Proc. ACM SIGCOMM. (2007)
24. Y. Wang, S. Jain, M. Martonosi, K. Fall. Erasure-Coding Based Routing for Opportunistic Networks. In SIGCOMM DTN Workshop. (2005)
25. A. Chaintreau, P. Hui, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. IEEE Transactions on Mobile Computing, 6(6):606–620. (2007)
26. N. Djukic, M. Piorkowski, and M. Grossglauser. Island hopping: Efficient mobility-assisted forwarding in partitioned networks. In Proc. of IEEE SECON (2006)
27. W. Zhao, M. Ammar, et al. A message ferrying approach for data delivery in sparse mobile ad hoc networks. InProc. Of ACM MOBIHOC. (2004)
28. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks.Physical Review E, 69. (2004)
29. M. E. J. Newman. Analysis of weighted networks. Physical Review E, 70:056131. (2004)
30. M. E. J. Newman. Detecting community structure in networks.Eur. Phys. J. B, 38:321–330. (2004)
31. G. Palla, I. Derenyi, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 435(7043):814–818. (2005)
32. S. Okasha. Altruism, group selection and correlated interaction. British Journal for the Philosophy of Science, 56(4):703–725. (2005)
33. J. Leguay, A. Lindgren, et al. Opportunistic content distribution in an urban setting. InACM CHANTS, pages205–212 (2006)
34. J. Lebrun, C.-N. Chuah, et al. Knowledge-based opportunistic forwarding in vehicular wireless ad hoc networks.IEEE VTC, 4:2289–2293. (2005)
35. P. Hui, E. Yoneki, et al. Distributed community detection in delay tolerant networks. InSigcomm Workshop MobiArch’07. (2007)
36. J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chain-treau. CRAWDAD data set cambridge/haggle (v. 2006-09-15). http://crawdad.cs.dartmouth.edu /cambridge/haggle. (2006)
37. F. Li and Jie Wu. LocalCom: A Community-based Epidemic Forwarding Scheme in Disruption-tolerant Networks. In Proceeding Secon’09, pages574–582 (2009)
38. Liu, X., Yong, X., Lin, H.: An Improved Spectral Clustering Algorithm Based on Local Neighbors in Kernel Space. Computer Science and Information Systems, Vol. 8, No. 4, 1143-1157. (2011)

Bing Bai received the B.S degree in Comupter Science from National University of Defense Technology, Changsha, China, in 2005. He is currently a PhD student in School of Computer at the National University of Defense Technology, Changsha, China. His research interests include Routing in Delay Tolerant Networks and WSN.

Zhenqian Feng is an Assistant Professor in the School of Computer Science, National University of Defense Technology. He recevied his Ph.D. degree in Computer Science from National University of Defense Technology, in 2012. His current research interests include Cloud Computing and Data Centre Networks.

Baokang Zhao is an Assistant Professor in the School of Computer Science, National University of Defense Technology. He recevied his Ph.D. degree in Computer Science from National University of Defense Technology, in 2009. He served as a program committee member for several international conferences and a reviewer for several international journals. He serves on the editor board of Journal of Internet Sevices and Information Security(JISIS). His current research interests include security and privacy in wireless networks, algorithms and protocols in computer networks, design and optimization in embedded systems. He is a member of the ACM, IEEE and CCF.

Jinshu Su received the B.S degree in Mathematics from Nankai University, Tianjin, China, in 1983, the M.S. degree in Computer Science, National University of Defense Technology, changsha, China, in 1989, and the PhD degree in Computer Science, National University of Defense Technology, changsha, China, in 1999. He is a full professor at the School of Computer Science, National University of Defense Technology, and serves as head of the Institute of network and information security, NUDT. He is the academic leader of the State Innovative Research Team in University ("Network Technology" Innovative Team) awarded by the Ministry of Education, CHINA. He has lead several national key projects of CHINA, including national 973 projects, 863 projects and NSFC Key projects. His research interests include high performance routers, internet routing, high performance computing, wireless networks and information security. He is a member of the ACM and IEEE.

Received: September 21, 2012; Accepted: March 18, 2013

DOI: 10.2298/CSIS120722039W

Wiener-based ICI Cancellation Schemes for OFDM Systems over Fading Channels

Jyh-Horng Wen[1], Yung-Cheng Yao[2], and Ying-Chih Kuo[3]

[1] Department of Electrical Engineering, Tunghai University
No. 181, Section 3, Taichung Port Road, Taichung City 40704, Taiwan
jhwen@thu.edu.tw
[2] Department of Electrical Engineering, National Chung Cheng University
168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan
yaoyc@thu.edu.tw
[3] Institute of Communications Engineering, National Chung Cheng University
168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan
ying@mail.ktc.com.tw

Abstract. The subcarriers of orthogonal frequency division multiplexing (OFDM) systems may fail to keep orthogonal to each other under time-varying channels. The loss of orthogonality among the subcarriers will degrade the system performace, and this effect is named intercarrier interference (ICI). In this paper, a Wiener-based successive interference cancellation (SIC) scheme is proposed to detect the OFDM signals. It provides good ICI cancellation performance; however, it suffers large computation complexity. Therefore, a modified Wiener-based SIC scheme is further proposed to reduce the computation complexity. Simulation results show the performance of the Wiener-based SIC scheme is better than those of zero forcing, zero forcing plus SIC and original Wiener-based schemes. Furthermore, with the modified Wiener-based SIC scheme, the performance is still better than the others. Although the performace of the modified Wiener-based SIC scheme suffers little degradation compared to Wiener-based SIC scheme, the computation complexity can be dramatically reduced.

Keywords: Orthogonal frequency division multiplexing (OFDM), fading channels, intercarrier interference (ICI), Wiener-based, successive interference cancellation (SIC).

1. Introduction

Orthogonal frequency division multiplexing (OFDM) has been applied in many digital transmission systems, such as digital audio broadcasting (DAB) system, digital video broadcasting terrestrial TV (DVB-T) system, asymmetric digital subscriber line (ADSL), IEEE 802.11a/g wireless local area network (WLAN), IEEE 802.16 worldwide interoperability for microwave access (WiMax) systems, and ultra-wideband (UWB) systems [1-6]. It can also be

applied to cooperative communication systems [7]. OFDM systems split a high-rate data stream into numbers of low-rate data stream. Since the available channel is divided into several narrowband subchannels, OFDM systems have such advantages: immunity to delay spread, resistance to frequency selective fading, simple equalization, and efficient bandwidth usage. However, OFDM systems have several disadvantages: the problem of synchronization; hardware complexity of FFT units at transmitter and receiver; the problem of high peak to average power ratio (PAPR); intercarrier interference (ICI) effect. The performance degrades significantly for intercarrier interference, and several methods have been proposed to mitigate the ICI effect with different efficiency and complexity.

The remainder of the paper is organized as follows. Related work is given in Section 2. In section 3, channel model of OFDM system is introduced. In section 4, signal detection and interference cancellation schemes are introduced. The simulation results are shown in section 5. Finally, the conclusion is given in section 6.

2. Related Work

Carrier frequency offset, caused by Doppler shift, and time-varying channel bring the intercarrier interference. Several ICI cancellation schemes have been proposed, and ZF (zero forcing) detection scheme is one of them. Although conventional ZF detection scheme is widely used in noise free environment, the noise enhancement occurs while suppressing the ICI effect. Wiener solution has been proved to be able to detect signals without noise enhancement [8]. On the other hand, successive interference cancellation scheme has been successfully used in MC-CDMA and OFDM systems to mitigate multiple access interference and intercarrier interference respectively [9-10]. In this paper, we first study the performance of Wiener-based SIC for OFDM systems over fading channels. Although the Wiener-based SIC scheme can provide good ICI cancellation performance, its computation complexity increases as number of subcarriers increases [11]. This is a trade-off between bit error rate (BER) performance and computation complexity. Therefore, we further study a modified Wiener-based SIC ICI cancellation scheme to reduce computation complexity without reducing BER performance or with minor BER performance degradation.

3. Channel Model

The block diagram of OFDM system shown in Fig. 1 has several propagation paths between transmitter and receiver. The schematic of multipath communication environment is shown in Fig. 2. Each path introduces different phase, amplitude attenuation, delay and Doppler shift to the signal.

Since the transmission environment is time-varying; therefore, the phase, attenuation, delay and Doppler shift of the signal are random variables.

For a time-varying multipath channel, the impulse response could be expressed as:

$$h(t,\tau)=\sum_{l=0}^{L-1}h_l(t)\delta(\tau-\tau_l), \tag{1}$$

where the amplitude of $h_l(t)$ is modeled as Rayleigh distribution with the maximum Doppler shift f_d, and it denotes the channel impulse response as l-th delay path at the time t. According to (1) the time delay and the attenuation are function of time. The Fig. 3 shows the time-varying channel. In the mobile radio channels, the Rayleigh distribution is commonly used to describe the statistical time-varying channel. It is well known the envelope of sum of two quadrature Gaussian noise signals obeying a Rayleigh distribution. Fig. 4 shows a Rayleigh distributed signal envelope as a function of time.

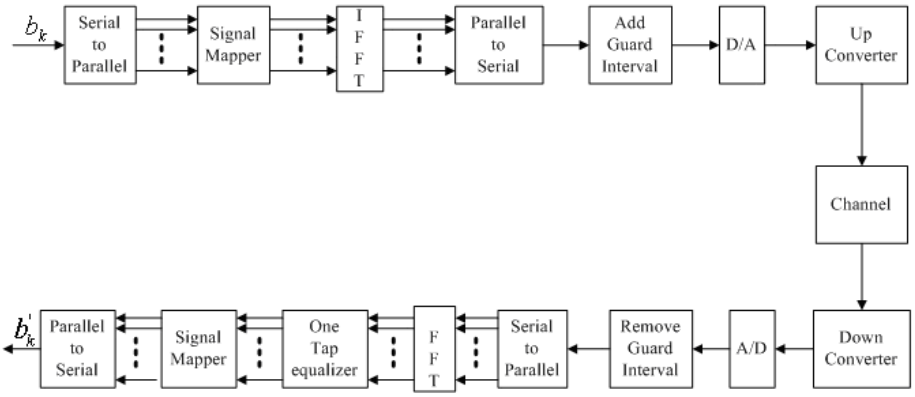


Fig. 1. Block diagram of OFDM systems

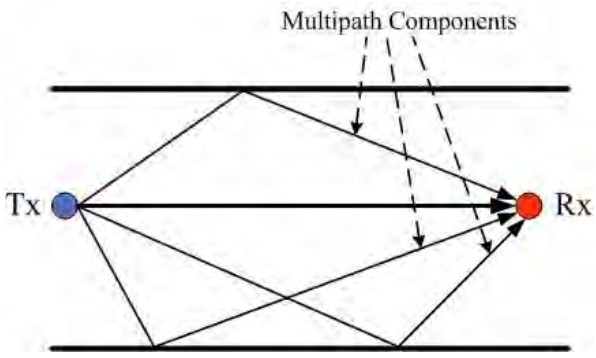


Fig. 2. The multipath environment

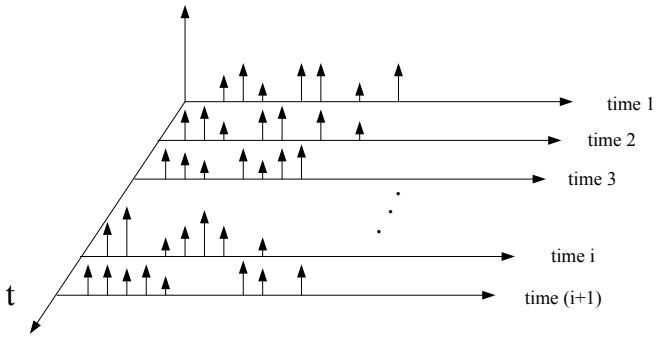


Fig.3. The impulse response of the time-varying channel

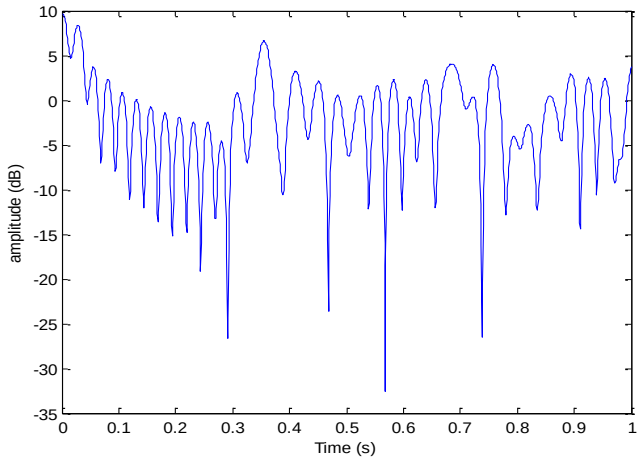


Fig. 4. Rayleigh distributed signal envelope

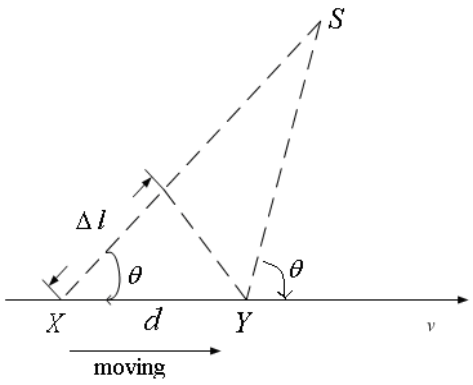


Fig. 5. Illustration of the Doppler effect

As shown in Fig. 5, consider a mobile moving at a constant velocity v, along a path with length d between point X and Y, it receives signal from a remote source S. The difference in path lengths traveled by the signal wave from source S to the mobile at point X and Y is $\Delta l = d\cos\theta = v\Delta t\cos\theta$, where Δt is the time required for the mobile to travel from X and Y, and the angles θ are assumed to be the same at points X and Y since the source is assumed to be very far away from the mobile. Therefore, the phase change in the received signal due to the difference in path lengths is

$$\Delta\phi = \frac{2\pi\Delta l}{\lambda} = \frac{2\pi v\Delta t}{\lambda}\cos\theta, \tag{2}$$

and the apparent change in frequency, or Doppler shift is given by f_m, where

$$f_m = \frac{v}{\lambda}\cos\theta = \frac{v}{c}f_c\cos\theta, \tag{3}$$

where c is velocity of light and f_c is carrier frequency. The Doppler shift f_d could be maximized when $\cos\theta$ is equal to 1.

The time-varying channel is expressed in (1), and the time-varying path gain $h_l(t)$ is generally represented by a Rayleigh random process. For the classical Doppler spectrum [12], the spectral density of $h_l(t)$ is

$$p(f) = \begin{cases} \dfrac{1}{\pi f_d\sqrt{1-\left(\dfrac{f}{f_d}\right)^2}} & \text{, if } |f| < f_d \\ 0 & \text{, else} \end{cases}, \tag{4}$$

where f_d is the maximum Doppler frequency. Therefore, if the channel is time-varying, the ICI would be occurred.

In the orthogonal frequency division multiplexing (OFDM) system, the transmission bandwidth is divided into many narrow subchannels, and they are transmitted in parallel. Because the bandwidth of subchannel is very narrow, channel response could be seen constant. In contrast to time domain, the symbol duration increases, such that the intersymbol interference (ISI) would be happened. If the guard interval is greater than the maximum delay path, the ISI will be removed. This is the reason why OFDM could against the frequency selective fading. Increase in the symbol duration makes it much more vulnerable to time selective fading due to the Doppler spread effect.

The output of IFFT (inverse fast Fourier transform) in OFDM system could be expressed as:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1}X(k)e^{j\frac{2\pi}{N}nk} \qquad , n = 0,1,2,...,N-1 \tag{5}$$

The transmitted signal could be represented as:

$$x(t)=\frac{1}{N}\sum_{k=0}^{N-1}X(k)e^{j\frac{2\pi}{T}kt} \qquad , -T_g \le t < T \tag{6}$$

where the interval $-T_g \leqq t < 0$ is the guard interval for opposing to the intersymbol interference (ISI).

Then the received signal $r(t)$ could be obtained as:

$$\begin{aligned} r(t) &= (x(t)e^{j2\pi f_c t}) * h(t,\tau) + w'(t) \\ &= \int_{-\infty}^{\infty} h(t,\tau')x(t-\tau')e^{j2\pi f_c(t-\tau')}d\tau' + w'(t) \\ &= \int_{-\infty}^{\infty}\sum_{l=0}^{L-1} h_l(t)\delta(\tau'-\tau_l)x(t-\tau')e^{j2\pi f_c(t-\tau')}d\tau' + w'(t) \\ &= \sum_{l=0}^{L-1} h_l(t)x(t-\tau_l)e^{j2\pi f_c(t-\tau_l)} + w'(t). \end{aligned} \tag{7}$$

where f_c is the carrier frequency.

The signal in the output of down converter is

$$\begin{aligned} y'(t) &= r(t)e^{-j2\pi(f_c-\Delta f')t} \\ &= \left[\sum_{l=0}^{L-1} h_l(t)x(t-\tau_l)e^{j2\pi f_c(t-\tau_l)} + w'(t)\right]e^{-j2\pi(f_c-\Delta f')t} \\ &= \sum_{l=0}^{L-1} h_l(t)x(t-\tau_l)e^{-j2\pi f_c\tau_l}e^{j2\pi\Delta ft} + w'(t)e^{-j2\pi(f_c-\Delta f')t} \end{aligned} \tag{8}$$

where Δf is the carrier frequency offset. After passing through the lowpass filter, the signal could be obtained as:

$$y(t)=\sum_{l=0}^{L-1} h_l(t)x(t-\tau_l)e^{j2\pi\Delta ft} + w(t) \tag{9}$$

Sampling the received signal $y(t)$ with the rate N/T and removing the portion of cyclic prefix, the received signal could be obtained as $y(n)=y(T/N)$, $n=0, 1, 2,\ldots, N-1$, within one symbol interval. The received signal could be rewritten as:

$$\begin{aligned} & y(n\frac{T}{N})=\sum_{l=0}^{L-1} h_l(n\frac{T}{N})x(n\frac{T}{N}-l\frac{T}{N})e^{j2\pi\Delta f(n\frac{T}{N})} + w(n\frac{T}{N}) \\ & \Rightarrow y(n)=\sum_{l}^{L-1} h_l(n)x(n-l)e^{j\frac{2\pi}{N}\varepsilon n} + w(n), \quad n=0,1,2,\ldots,N-1 \end{aligned} \tag{10}$$

where $|\tau_{l+1}-\tau_l|$ is assumed equal to T/N, and the normalized frequency offset is represented as $\varepsilon=\Delta f\times T=\Delta f/f$ in which T is the symbol duration, and f is the subcarrier spacing.

The FFT of $y(n)$ could be expressed as:

$$
\begin{aligned}
Y(m) &= \sum_{n=0}^{N-1} y(n) e^{-j\frac{2\pi}{N}nm} \\
&= \sum_{n=0}^{N-1}\left(\sum_{l=0}^{L-1} h_l(n)x(n-l)e^{-j\frac{2\pi}{N}\varepsilon n} + w(n)\right)e^{-j\frac{2\pi}{N}nm} \\
&= \sum_{n=0}^{N-1}\left(\sum_{l=0}^{L-1} h_l(n)\left(\sum_{k=0}^{N-1}\frac{1}{N}X(k)e^{j\frac{2\pi}{N}k(n-l)}\right)e^{-j\frac{2\pi}{N}\varepsilon n} + w(n)\right)e^{-j\frac{2\pi}{N}nm} \\
&= \sum_{n=0}^{N-1}\left(\sum_{l=0}^{L-1} h_l(n)\left(\sum_{k=0}^{N-1}\frac{1}{N}X(k)e^{j\frac{2\pi}{N}k(n-l)}\right)e^{-j\frac{2\pi}{N}\varepsilon n}\right)e^{-j\frac{2\pi}{N}nm} + W(m) \\
&= \sum_{n=0}^{N-1}\left(\sum_{l=0}^{L-1}\left(\frac{1}{N}\sum_{k=0}^{N-1} h_l(n)e^{j\frac{2\pi}{N}(k+\varepsilon-m)n}\right)e^{-j\frac{2\pi}{N}kl}\right)X(k) + W(m)
\end{aligned}
\tag{11}
$$

where the ICI term is defined as:

$$
\begin{aligned}
\mathrm{ICI} &= \sum_{l=0}^{L-1}\left(\frac{1}{N}\sum_{k=0}^{N-1} h_l(n)e^{j\frac{2\pi}{N}(k+\varepsilon-m)n}\right)e^{-j\frac{2\pi}{N}kl} \\
&= \left(\sum_{l=0}^{L-1} h_l(n)e^{-j\frac{2\pi}{N}kl}\right)\sum_{n=0}^{N-1}\frac{1}{N}e^{j\frac{2\pi}{N}(k+\varepsilon-m)n}
\end{aligned}
\tag{12}
$$

According to the (12), it is clear that if there is no frequency offset, ε=0, and the channel is stationary, $h_l(n)$=h_l, then the ICI= $h(k)\delta(m\text{-}k)$, there will be no intercarrier interference. If there is no frequency offset, ε=0, due to the time-varying channel fading characteristic of the mobile channel, ICI would exist in OFDM systems for the mobile application. In contrarily, the channel is stationary but the frequency offset is not equal to zero, ICI would still exist in OFDM systems.

In this paper we focus on the time-varying channel fading characteristic of the mobile channel, so we set frequency offset ε equal to zero. In time-varying channels, the ICI term is defined as:

$$
\begin{aligned}
\mathrm{ICI} &= \sum_{l=0}^{L-1}\left(\frac{1}{N}\sum_{k=0}^{N-1} h_l(n)e^{j\frac{2\pi}{N}(k-m)n}\right)e^{-j\frac{2\pi}{N}kl} \\
&= \sum_{l=0}^{L-1}\left(H_l(m-k)\right)e^{-j\frac{2\pi}{N}kl}.
\end{aligned}
\tag{13}
$$

where $H_l(m\text{-}k)$ is the ICI effect of the k-th subcarrier to the m-th subcarrier. Then output of FFT (fast Fourier transform) is also written as:

$$
\begin{aligned}
Y(m) &= \sum_{k=0}^{N-1}\left(\sum_{l=0}^{L-1}\left(H_l(m-k)\right)e^{-j\frac{2\pi}{N}kl}\right)X(k) + W(m) \\
&= \sum_{l=0}^{L-1}\left(H_l(0)\right)e^{-j\frac{2\pi}{N}kl}X(m) + \sum_{\substack{k=0 \\ k\neq m}}^{N-1}\left(\sum_{l=0}^{L-1}\left(H_l(m-k)\right)e^{-j\frac{2\pi}{N}kl}\right)X(k) + W(m)
\end{aligned}
\tag{14}
$$

where the first term is the desired signal and the second term is the ICI component.

For a time-varying fading channel, the channel variations would lead to the loss of orthogonality between subcarriers, hence the ICI effect could be occurred. The ICI effect in OFDM systems would result in an error floor. In next section, we would propose a scheme to suppress ICI.

4. Signal Detection and Interference Cancellation Schemes

4.1. Zero-forcing Detection Scheme

The received signals $y(n)$ in the time-varying channel could be obtained as:

$$y(n)=\sum_{l=0}^{L-1}h_l(n)x(n-l)+w(n) \qquad ,n=0,1...,N-1 \tag{15}$$

The received signals can be represented as a matrix form, and it is represented as:

$$\mathbf{y}=\mathbf{hx}+\mathbf{w}. \tag{16}$$

Each element of the received signal $\mathbf{y}$, the channel matrix $\mathbf{h}$, the transmitted signal $\mathbf{x}$, and the AWGN (additive white Gaussian noise) $\mathbf{w}$ can be expressed as:

$$\begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(N-1) \end{bmatrix} = \begin{bmatrix} h_0(0) & 0 & \cdots & 0 & h_{L-1}(0) & h_{L-2}(0) & \cdots & h_1(0) \\ h_1(1) & h_0(1) & & \cdots & 0 & h_{L-1}(1) & & h_2(1) \\ \vdots & \ddots & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & h_{L-1}(N-1) & h_{L-2}(N-1) & \cdots & \cdots & h_0(N-1) \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} + \begin{bmatrix} w(0) \\ w(1) \\ \vdots \\ w(N-1) \end{bmatrix} \tag{17}$$

The element $h_l(n)$ in $\mathbf{h}$ denotes the channel response of the l-th path at the n-th sample time. If N is the number of subcarriers, then $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{w}$ can be expressed as an N-by-1 vector, and $\mathbf{h}$ is an N-by-N matrix.

The Fourier Transform of the received signal, $\mathbf{y}$, can be multiplied by $\mathbf{F}$ on both sides of (16). Hence, the received signal vector $\mathbf{Y}$ in the frequency domain will be expressed as:

$$\mathbf{Y}=\mathbf{HX}+\mathbf{W}, \tag{18}$$

where $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{W}$ denote the Fourier series of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{w}$, respectively. $\mathbf{H}$ is defined as the frequency domain channel matrix, and it can be expressed in terms of matrix $\mathbf{h}$. The frequency domain channel matrix $\mathbf{H}$ can be obtained as $\mathbf{FhF}^H$, where $\mathbf{F}$ denotes the N-by-N Fourier Transform matrix,

which can be seen in (19), and $\mathbf{F}^H$ denotes applying the Hermitian operation on $\mathbf{F}$.

$$\mathbf{F}=\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{-j\frac{2\pi}{N}} & e^{-j\frac{2\pi}{N}*2} & \cdots & e^{-j\frac{2\pi}{N}*(N-1)} \\ 1 & e^{-j\frac{2\pi}{N}*2} & e^{-j\frac{2\pi}{N}*4} & \cdots & e^{-j\frac{2\pi}{N}*2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{-j\frac{2\pi}{N}*(N-1)} & e^{-j\frac{2\pi}{N}*2(N-1)} & \cdots & e^{-j\frac{2\pi}{N}*(N-1)(N-1)} \end{bmatrix}. \quad (19)$$

In order to detect the signals in (18), the zero forcing (ZF) detection scheme can be used by inversing the channel matrix $\mathbf{H}$. If the matrix $\mathbf{H}$ is not a square matrix, the inverse of the matrix will be replaced with the pseudo-inverse operation. Hence, the detected signals can be obtained as $\hat{\mathbf{X}}=\mathbf{H}^{+}\mathbf{Y}$, where the matrix $\mathbf{H}^{+}=(\mathbf{H}^{H}\mathbf{H})^{-1}\mathbf{H}^{H}$ is the pseudo-inverse of $\mathbf{H}$. Noise enhancement will occur when the ZF method is used, because the detected signals will be obtained as $\hat{\mathbf{X}}=\mathbf{H}^{+}\mathbf{Y}$.

$$\hat{\mathbf{X}}=\mathbf{H}^{+}\mathbf{Y}=\mathbf{H}^{+}(\mathbf{HX}+\mathbf{W})=\mathbf{IX}+\mathbf{H}^{+}\mathbf{W} \quad (20)$$

In (20), the first term becomes an identical matrix multiplied with the signal $\mathbf{X}$, and the second term cannot become a zero vector. Then, the second term $\mathbf{H}^{+}\mathbf{W}$ may enhance the noise term if some components in $\mathbf{H}^{+}$ become large due to the operation of $(\mathbf{H}^{H}\mathbf{H})^{-1}$. In the noise free environment, ZF detection will be widely used. However, noise enhancement occurs when the ZF detection is used.

4.2. Wiener-based Detection Scheme

In the adaptive theory [8], the Wiener filter is useful for communication systems. The Wiener filter theory is formulated for the general case of a complex valued stochastic process with the filter specified in terms of its impulse response.

In frequency domain, the received signals are $\mathbf{Y}=\mathbf{HX}$ where $\mathbf{X}$ is the transmitted signal. In order to estimate signals at the receiver, the estimated signal $\hat{\mathbf{X}}$ can be obtained by the Wiener solution. The Wiener solution $\mathbf{K}$ can be obtained by the following algorithm.

In order to find the Wiener solution, we must minimize the cost function. We define the cost function as the mean square error between the estimated signal $\hat{\mathbf{X}}$ and the transmitted signal $\mathbf{X}$. Then, the cost function can be expressed as

$$C = E\left(\left(\mathbf{X} - \hat{\mathbf{X}}\right)^2\right) \tag{21}$$

The Wiener solution is shown as follow:

$$\mathbf{K} = \arg\min_{\mathbf{Q}} C = \arg\min_{\mathbf{Q}} E\left(\left\|\mathbf{X} - \mathbf{QY}\right\|^2\right), \tag{22}$$

The Wiener solution can be achieved by the "orthogonal principle," and the geometric interpretation is presented in Fig. 6.

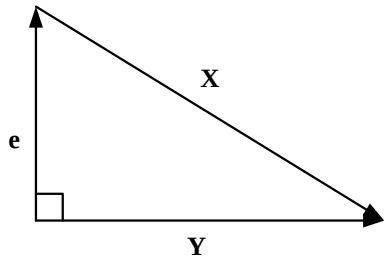


Fig. 6. Geometric interpretation of the relationship between the desired signal $\mathbf{X}$, the output of FFT $\mathbf{Y}$, and the mean square error $\mathbf{e}$

To achieve the orthogonal principle, the inner product between $\mathbf{Y}$ and $\mathbf{e}$ is held to zero

$$E\left(\left(\mathbf{X} - \hat{\mathbf{X}}\right)^H \mathbf{Y}\right) = 0 \Rightarrow E\left(\mathbf{X}^H \mathbf{Y}\right) = E\left(\mathbf{Y}^H \mathbf{Y}\right)\mathbf{K} \tag{23}$$

The Wiener solution can be determined by (23). Then, this solution is expressed as follows:

$$\mathbf{K} = \left(E\left(\mathbf{Y}^H \mathbf{Y}\right)\right)^{-1} E\left(\mathbf{X}^H \mathbf{Y}\right) = \left(\mathbf{H}^H \mathbf{H} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}^H \tag{24}$$

Therefore, the detected signal $\hat{\mathbf{X}}$ can be obtained by the Wiener solution as $\hat{\mathbf{X}} = \mathbf{KY}$.

4.3. Wiener-based ICI Cancellation Scheme

Since the Wiener solution can detect a reliable signal, the results will be applicable to signal detection in the successive interference cancellation scheme.

In the successive interference cancellation scheme, in order to utilize ICI as a source of diversity, both reliable signal detection and an efficient ICI cancellation are needed. First, in order to achieve reliable signal detection, we utilize the method of ordering received signals based on signal-to-interference and noise ratio (SINR).

In order to fully utilize the time diversity while suppressing the residual interference and the noise enhancement, the signal detection is successive, but not detecting all the signals simultaneously.

The detection orders with subcarriers in the SIC scheme are decided by the SINR. The SINR can be obtained in the following manner. The vector of the received signal $\mathbf{Y}$ can be expressed as $\mathbf{Y}=\mathbf{HX}+\mathbf{W}$. The received signal can also be represented as:

$$\mathbf{Y}=\begin{bmatrix}\vec{h}_0 & \vec{h}_1 & \cdots & \vec{h}_{N-2} & \vec{h}_{N-1}\end{bmatrix}\begin{bmatrix}X(0)\\X(1)\\\vdots\\X(N-2)\\X(N-1)\end{bmatrix}+\mathbf{W}, \tag{25}$$

where $\vec{h}_k$ is the k-th column vector of the channel matrix $\mathbf{H}$, and the $X(k)$ is the k-th subcarrier signal at the input of IFFT.

Then, the vector of the received signal can be rewritten as the following form:

$$\mathbf{Y}=\vec{h}_0X(0)+\vec{h}_1X(1)+\cdots+\vec{h}_{N-1}X(N-1)+\mathbf{W}, \tag{26}$$

Then, we can use the Wiener solution to detect the signal $\hat{\mathbf{X}}$. Therefore, the k-th signal of the k-th subcarrier can be detected by $\hat{X}(k)=\vec{k}_k\mathbf{Y}$, where $\vec{k}_k$ is the k-th row vector of the Wiener solution $\mathbf{K}$, and the k-th signal of the k-th subcarrier can be obtained as:

$$\vec{k}_k\mathbf{Y}=\vec{k}_k(\vec{h}_0X(0)+\vec{h}_1X(1)+\cdots+\vec{h}_kX(k)+\cdots+\vec{h}_{N-1}X(N-1)+\vec{h}_NX(N))+\vec{k}_k\mathbf{W} \tag{27}$$

In (27) the desired signal is denoted as $\vec{k}_k\vec{h}_kX(k)$, and the others are the ICI and noise component. Hence, for the particular subcarrier k, the $SINR_k$ is defined as:

$$SINR_k=\frac{E\left[\left\|\vec{k}_k\vec{h}_kX(k)\right\|^2\right]}{E\left[\sum_{\substack{q=0\\q\neq k}}^{N-1}\left\|\vec{k}_k\vec{h}_qX(q)\right\|^2\right]+E\left[\left\|\vec{k}_k\mathbf{W}\right\|^2\right]}. \tag{28}$$

Each subcarrier's SINR can be obtained by (28). Hence, the subcarrier with the highest SINR is decided, following which we first detect the signal $\hat{X}(k)$. Equivalently, we choose the k-th row vector of $\mathbf{K}$ to detect the signal of the k-th subcarrier. After making a hard decision, the detected signal of the k-th subcarrier $\hat{X}(k)$ is reconstructed as the ICI component of the k-th subcarrier. Then, ICI effect for the k-th subcarrier will be cancelled in the received signal.

Following this, after canceling the ICI for the k-th subcarrier, the received signal $\mathbf{Y}$ will be obtained to

$$\mathbf{Y}^{(j+1)} = \mathbf{Y}^{(j)} - \vec{h}_k \hat{X}(k), \tag{29}$$

where $\hat{X}(k)$ is the hard decision signal of the k-th subcarrier. $\vec{h}_k \hat{X}(k)$ is the ICI term corresponding to *k*-th subcarrier. As long as this hard decision data is correct, the new vector $\mathrm{Y}^{(j+1)}$ has less interference. After this operation, the ICI

term of the detected signal will be removed and the channel matrix $\mathbf{H}$ should be reconstructed by removing the k-th column vector and k-th row vector. The column number and row number of the new channel matrix is then reduced. Therefore, the proposed Wiener-based SIC scheme repeats these steps until all the subcarriers are detected completely. According to this process, we will be able to detect all signals completely. The simulation results will be shown in section 5.

4.4. Modified Wiener-Based SIC ICI Cancellation Scheme

From above, we clearly know that the computation complexity of the Wiener-based successive interference cancellation scheme is very high. The size of the channel matrix $\mathbf{H}$ will be increased with the number of subcarriers. Hence, the computation complexity will increase. Therefore, in this section, we will propose an algorithm to reduce the computation complexity for the Wiener-based SIC scheme.

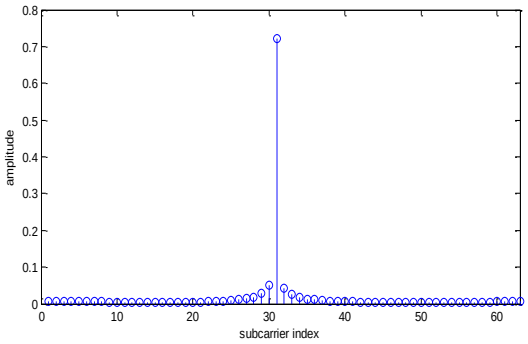


Fig. 7. The ICI amplitude for the desired signal at subcarrier 30

The traditional Wiener filter or a ZF equalizer is too complicated to be implemented, since it involves an N-by-N matrix inverse and matrix multiplication. N is usually fairly large, for example, N = 64 for the IEEE 802.11a standard, and N = 1024 for the IEEE 802.16 standard. As a matter of the fact, the ICI power arises from the neighboring subcarriers around the

subject subcarrier. If we only focus on the neighboring subcarriers around the subject subcarrier, the computation complexity can be reduced significantly. Consequently, a simple scheme is investigated in Fig. 7 and Fig. 8, which show the ICI amplitude and ICI power for the desired signal.

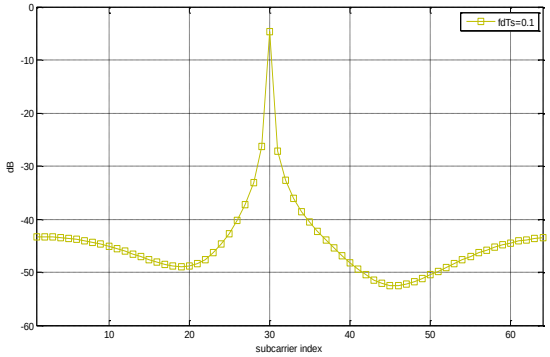


Fig. 8. The ICI power for the desired signal at subcarrier 30

Fig. 7 and Fig. 8 show the effective subcarriers that contribute the ICI to specific subcarrier are actually much smaller than the number of subcarriers in one OFDM symbol. Using this fundamental observation, we are going to focus on few subcarriers around the desired subcarrier. Now, if we only focus on q subcarriers around the desired subcarrier, we can rewrite the frequency-domain channel matrix $\mathbf{H}$ as $\bar{\mathbf{H}}$, and $\bar{\mathbf{H}}$ can be expressed as follows:

$$\bar{\mathbf{H}} = \begin{bmatrix} H_{0,0} & H_{0,1} & \cdots & H_{0,q-1} & 0 & \cdots & H_{0,N-q} & H_{0,N-q+1} & \cdots & H_{0,N-1} \\ H_{1,0} & H_{1,1} & & & & \ddots & & & & H_{1,N-1} \\ \vdots & & H_{2,2} & & & & & & & \vdots \\ H_{q-1,0} & & & \ddots & & & & & & H_{q-1,N-1} \\ 0 & & & & \ddots & & & & & 0 \\ \vdots & & & & & \ddots & & & & \vdots \\ H_{N-q,0} & \ddots & & & & & \ddots & & & H_{N-q,N-1} \\ H_{N-q+1,0} & & \ddots & & & & & & & H_{N-q+1,N-1} \\ \vdots & & & \ddots & & & & & & \vdots \\ H_{N-1,0} & H_{N-1,1} & \cdots & H_{N-1,q-1} & 0 & \cdots & H_{N-1,N-q} & H_{N-1,N-q+1} & \cdots & H_{N-1,N-1} \end{bmatrix} \quad (30)$$

The channel matrix $\bar{\mathbf{H}}$ is shown as (30). Hence, each signal on the subcarrier in the output of FFT can be rewritten as:

$$Y(m) = \sum_{k=m-q}^{m+q} H_{m,k} X(k), \quad (31)$$

where

$$H_{m,k} = \frac{1}{N} \sum_{l=0}^{L-1} e^{-j\frac{2\pi}{N}kl} \sum_{n=0}^{N-1} h_l(n) e^{-j\frac{2\pi}{N}(m-k)n}$$

Hence, for detecting the k-th subcarrier, we only used the partial element of $\mathbf{Y}$ to detect the signal. For example, if the value q is set as 2, two neighboring subcarriers (two on each side) are employed in the simplified equalizer; if the desired subcarrier is $X(3)$, then received signal signals $Y(1)$, $Y(2)$, $Y(4)$, and $Y(5)$ are all used. Therefore, if we want to detect the k-th subcarrier, the vector of the received signal and the vector of the channel matrix in the frequency-domain can be reduced as:

$$\mathbf{Y}_k = \mathbf{H}_k \mathbf{X}_k + \mathbf{W}_k, \quad k = 0,1,2,...,N-1 \tag{32}$$

$$\mathbf{Y}_k = \mathbf{Y}(k-q:k+q) \tag{33}$$

$$\mathbf{H}_k = \bar{\mathbf{H}}(k-q:k+q, k-2q:k+2q) \tag{34}$$

$$\mathbf{X}_k = \mathbf{X}(k-q:k+q) \tag{35}$$

where H_k indicates the partial matrix of $\bar{\mathbf{H}}$ that is the consecutive row vector from the $(k - q)$-th vector to the $(k + q)$-th vector and column vector from the $(k - 2q)$-th vector to the $(k + 2q)$-th vector. $\mathbf{Y}_k$ means the partial vector whose elements are consecutive from $Y(k\text{-}q)$ to $Y(k+q)$. Here, if k-q<0, the related (k-q)-th vector and (k-q)-th element is redefined as ((k-q) mod N)-th vector and ((k-q) mod N)-th element.

In the modified algorithm, the channel matrix H is reduced to H_k and the size is also reduced from N-by-N to $(2q + 1)$-by-$(4q + 1)$. Then, the Wiener solution will be obtained as:

$$\mathbf{G}_k = \left(\mathbf{H}_k{}^H \mathbf{H}_k + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}_k{}^H = [\vec{g}_0\ \vec{g}_1 \cdots \vec{g}_{2q}]^T, \tag{36}$$

where $\vec{g}_n$ is the n-th row vector of $\mathbf{G}_k$.

The detection scheme presented above is modified in the Wiener-based SIC scheme. If the size of the matrix is reduced, the computation complexity of matrix multiplication and matrix inverse is also reduced. In the modified scheme, first, we find the SINR for each subcarrier and sort each. The detection order is from the maximum SINR to the minimum SINR. Equivalently, we apply the modified detection scheme in the Wiener-based scheme. The algorithm of the modified scheme for reducing the computation complexity is shown in the following steps.

Step 1. Find the Wiener solution and compute the SINR for each subcarrier.

Step 2. Find the maximum $\mathrm{SINR_k}$ with the k-th subcarrier of undetected subcarriers, and find the Wiener solution.

$$\mathbf{G}_k = \left(\mathbf{H}_k{}^H \mathbf{H}_k + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{H}_k{}^H.$$

Step 3. Detect the k-th subcarrier signal that has the maximum SINR.

$\hat{X}(k) = decision(\vec{g}_q \mathbf{Y}_k)$ where $\vec{g}_q$ is the $q-$th row of $\mathbf{G}_k$

Step 4. Cancel the intercarrier interference for $\hat{X}(k)$.

$\mathbf{Y}^{(j+1)} = \mathbf{Y}^{(j)} - \vec{h}_k \hat{X}(k)$ where $\vec{h}_k$ is the $k-$th column of $\mathbf{H}$

Step 5. Let the k-th column vector of $\mathbf{H}$ equal zero. ($\vec{h}_k = 0$)

Step 6. If $\mathbf{H}$ becomes a zero matrix, stop the scheme; if not, return to Step 2.

According to the modified algorithm for reducing the computation complexity, we analyze the complexity for different methods, and compare the order of computation complexity in section 4.5. In section 5, we show the BER performance of the modified SIC scheme for a different value q.

4.5. Complexity Analysis

The evaluation of the computation complexity for matrix operations follows the rules in [13–14]. For an N-by-N matrix multiplication or inversion, its order of the computation complexity is equivalent to $O(N^{2.376})$. For an M-by-N matrix multiplication with an N-by-M matrix and an M-by-N matrix, it is equivalent to $O(N^{1.376+r})$ of computation complexity where $r = \log_M N$.

In the Wiener-based SIC scheme, it is necessary to undertake matrix multiplication and inverse operation for each iteration. The Wiener solution can be obtained as $\mathbf{K}=(\mathbf{H}^H\mathbf{H}+\sigma^2\mathbf{I})^{-1}\mathbf{H}^H$; hence, the computation complexity of the Wiener solution for each OFDM symbol is obtained as:

$$k^{(r+1.376)} + k^{2.376} + k^{(r+1.376)}, \tag{37}$$

where k is the size of matrix $\mathbf{H}$. The first term denotes the computation complexity of the matrix multiplication, $\mathbf{H}^H\mathbf{H}$, the second term denotes the computation complexity of the inverse operation, and third term denotes the computation complexity of the results that are caused by the inverse of $\mathbf{H}^H\mathbf{H}+\sigma^2\mathbf{I}$ multiplication with $\mathbf{H}^H$. According to the Wiener-based SIC scheme in section 4.3 , the complexity for an OFDM symbol can be computed as:

$$\sum_{k=2}^{N}[2(k^{(r+1.376)}) + k^{2.376}] = \sum_{k=2}^{N}[2(k^{2.376}) + k^{2.376}] = \sum_{k=2}^{N} 3(k^{2.376}). \tag{38}$$

In the modified Wiener-based SIC scheme, the Wiener solution is rewritten as $\mathbf{G}_k=(\mathbf{H}_k^H\mathbf{H}_k+\sigma^2\mathbf{I})^{-1}\mathbf{H}_k^H$; hence, the computation complexity for an OFDM symbol is also computed as:

$$\begin{aligned} &3N^{2.376} + N((2q+1)^{r+1.376} + (2q+1)^{2.376} + (2q+1)^{r+.376}) \\ &= 3N^{2.376} + N(2(2q+1)^{2.376} + (2q+1)^{r+1.376}), \end{aligned} \tag{39}$$

where $r = \log_{(4q+1)}(2q+1)$. Table 1. shows the complexity for each method.

Table 1. Order of computation complexity for the presented methods

Method	Multiplication	Results (N=64)
ZF	$3N^{2.376}$	58,696
Wiener solution	$3N^{2.376}$	58,696
ZF+SIC	$\sum_{k=2}^{N} 3(k^{2.376})$	1,142,237
Wiener+SIC	$\sum_{k=2}^{N} 3(k^{2.376})$	1,142,237
Reduced complexity scheme	$3N^{2.376} + 2N(2q+1)^{r+1.376} + N(2q+1)^{2.376}$	175,075 (for q=8) 85,002 (for q=4)

According to Table 1., we can discover that the computation complexity of the modified Wiener-based SIC scheme is less than that of the original Wiener-based SIC scheme. If the number of subcarriers is very large, the gap of computation complexity between the original Wiener-based SIC scheme and the modified Wiener-based SIC scheme will increase. In the next section, we show the performance for the different value q.

5. Simulation Results

In this section, we demonstrate the BER performance of our proposed scheme. We investigate the performance of the proposed scheme over Rayleigh fading channels. The environment of simulation is shown in Table 2.

Table 2. The environment of the simulation

Modulation scheme	QPSK
Number of subcarriers	64
Channel	Rayleigh fading multipath channel
Normalized Doppler frequency $f_d T_s$	0.1, 0.05
Path number	6

In Fig. 9, we simulated the BER performance of the OFDM system with different schemes. The simulation results show that the BER performance would be error floor when ICI occurs. The performance with the Wiener solution scheme is better than that with the ZF scheme, because the latter does not consider the noise term. Therefore, using the ZF method involves noise enhancement. The use of the Wiener solution to detect the signal in the Wiener-based SIC scheme ensures that the noise enhancement will be avoided. Hence, the Wiener-based SIC scheme's performance is also better than that of ZF-SIC scheme.

Fig. 10 shows that the BER performance with different normalized Doppler frequencies. The system has better BER performance when the normalized Doppler frequency is large. Meanwhile, as the $f_d T_s$ gets large, the Wiener-based SIC scheme achieves more diversity gain. This shows that the Wiener-based SIC scheme can utilize the ICI as a source of diversity.

According to the Fig. 11, the modified Wiener-based SIC scheme's performance is not better than that of the original Wiener-based SIC scheme. Comparing the modified Wiener-based SIC scheme with the original Wiener-based SIC scheme, the performance loss of the modified Wiener-based SIC scheme is 2 dB for q = 8 at BER = 10^{-3}. However, according to the Table 2, the computation complexity of the modified Wiener-based SIC scheme is much lower than that of the original Wiener-based SIC scheme. Therefore, a suitable value, q, is a trade-off problem between performance and computation complexity.

6. Conclusions

A refined SIC detection for OFDM systems under Rayleigh fading channels has been presented. The performance for a low SINR subcarrier can be significantly improved due to ICI reduction scheme. According to the simulation results, SIC detection is very suitable for high fading rate mobile communications, such as the high-speed rail communication systems. The algorithm and the BER performance for the Wiener-based SIC scheme have been presented. According to the simulation results, we could clearly realize the performance of the Wiener-based SIC scheme is better than the ZF-SIC scheme's. Because the detection scheme may have noise enhancement in ZF-SIC scheme, the performance would be degraded. Although the Wiener-based SIC scheme has better performance, it has high computation complexity. In order to reduce the computation complexity for the Wiener-based SIC scheme, the modified Wiener-based SIC scheme is proposed.

According to the analysis, the computation complexity of the modified Wiener-based SIC scheme with q = 8 is 15% of the original Wiener-based SIC scheme. According to complexity analysis and simulation results, the performance with a large q is better than the performance with a small q, but the computation complexity is higher. Hence, this is a trade-off problem between the system performance and the computation complexity.

The schemes studied in this paper require perfect channel state information. To obtain a precision channel state information becomes an important issue worthy of further studying. Besides, due to the population of MIMO-OFDM technology, applying the proposed scheme to MIMO-OFDM systems is also worthy to investigate.

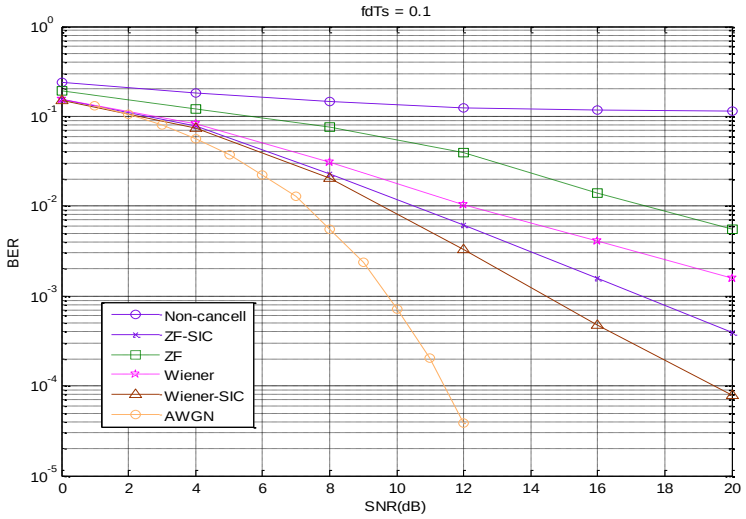


Fig. 9. BER versus SNR of four methods for $f_d T_s$ = 0.1

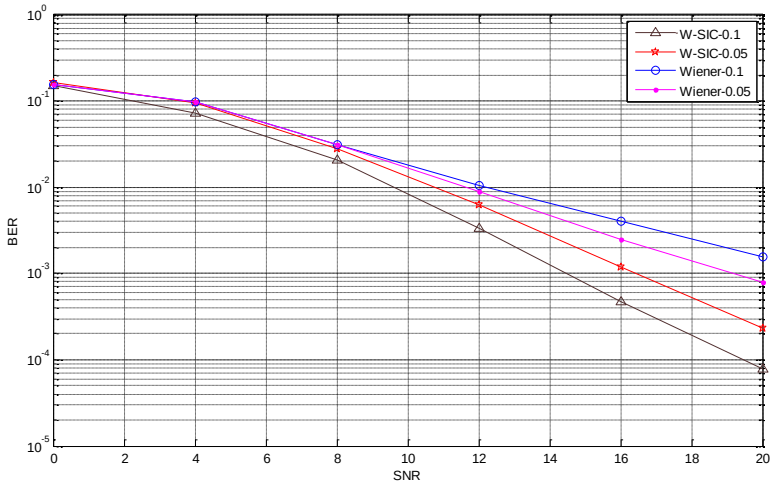


Fig. 10. BER versus SNR for different $f_d T_s$

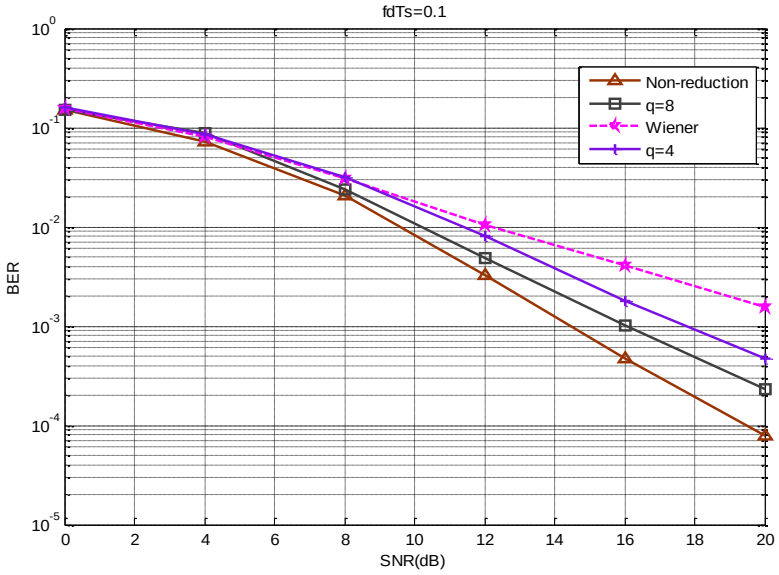


Fig. 11. BER versus SNR for different *q*

Acknowledgment. This work is partially supported by National Science Council, Taiwan, under Grant NSC 101-2221-E-029-020-MY3.

References

1. M. Engels, *Wireless OFDM Systems: How to Make Them Work?*, Kluwer Academic Publishers, 2002.
2. W.Y. Zou, and Y. Wu, "COFDM: an overview," *IEEE Transactions on Broadcasting*, vol. 41, no. 1, pp. 1-8, Mar. 1995.
3. L. Hanzo, W. Webb, and T. Keller, Single and Multi-carrier Quadrature Amplitude Modulation – Principles and Applications for Personal Communications, WLANs and Broadcasting, John Wiley & Sons Ltd., West Sussex, England, 2000.
4. T. Y. Al-Naffouri, K. M. Z.I Islam, N. Al-Dhahir, and S. Lu, "A Model Reduction Approach for OFDM Channel Estimation Under High Mobility Conditions," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, Apr. 2010.
5. A. Al-Habashna, O. A. Dobre, R. Venkatesan, and D. C. Popescu, "Second-Order Cyclostationarity of Mobile WiMAX and LTE OFDM Signals and Application to Spectrum Awareness in Cognitive Radio Systems," *IEEE Journal of Selected Topics In Signal Processing*,vol. 6, no. 1, Feb. 2012.
6. N. Uchida, K. Takahata, Y. Shibata and N. Shiratori, "Never Die Network Based on Cognitive Wireless Network and Satellite System for Large Scale Disaster," *JoWUA*, vol. 3, no. 3, pp. 74-93, Sep. 2012.

7. M. Čudanov, O. Jaško, and M. Jevtić, "Influence of Information and Communication Technologies on Decentralization of Organizational Structure," *Computer Science and Information Systems*, vol. 6, no. 1, pp. 93-109, Jun. 2009.
8. S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice-Hall, New Jersey, USA, 2002.
9. J.Y. Kim and H. Huh, "MAI mitigation by SIC for a multicarrier CDMA system," *IEEE Topical Conference on Wireless Communication Technology*, pp. 103-104, 2003.
10. J.H. Wen, Y.C. Yao, and Y.C. Kuo "Wiener-Based SIC Scheme for ICI Cancellation in OFDM Systems under Time-Varying Channels," *International Journal of Advanced Information Technologies*, pp. 93-99, Jun. 2012.
11. Y.C. Yao, Y.C. Kuo, and J.H. Wen, " A Simplified Wiener-Based SIC ICI Cancellation Scheme for OFDM Systems over Time-Varying Channels," *Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing,* pp. 475-480, 2012.
12. W.C. Jakes, Ed., *Microwave Mobile Communications*, 2nd ed., Wiley-IEEE Press, New York, USA, 1994.
13. X. Huang and and V. Y. Pan, "Fast retangular matrix multiplication and applications," *Journal of Complexity*, vol. 14, no. 2, pp. 257-299, June 1998.
14. D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Journal of Symbolic Computation*, vol. 9, no. 3pp. 251-280, Mar. 1990.

Jyh-Horng Wen is a professor in Department of Electrical Engineering, Tunghai University, Taiwan. His current research interests include computer communication networks and wireless broadband systems.

Yung-Cheng Yao and **Ying-Chih Kuo** are both graduate students of Professor Wen.

Received: July 2, 2012; Accepted: January 29, 2013

DOI: 10.2298/CSIS121102040T

Efficient Implementation for QUAD Stream Cipher with GPUs

Satoshi Tanaka[1], Takashi Nishide[2], and Kouichi Sakurai[2]

[1] Graduate School of Information Science and Electrical Engineering,
744 Motooka, Nishi-ku, Fukuoka, Japan
tanasato@itslab.inf.kyushu-u.ac.jp
[2] Faculty of Information Science and Electrical Engineering,
744 Motooka, Nishi-ku, Fukuoka, Japan
{nishide@inf, sakurai@csce}.kyushu-u.ac.jp

Abstract. QUAD stream cipher uses multivariate polynomial systems. It has provable security based on the computational hardness assumption. More specifically, the security of QUAD depends on hardness of solving non-linear multivariate systems over a finite field, and it is known as an NP-complete problem. However, QUAD is slower than other stream ciphers, and an efficient implementation, which has a reduced computational cost, is required.
In this paper, we propose an efficient implementation of computing multivariate polynomial systems for multivariate cryptography on GPU and evaluate efficiency of the proposal. GPU is considered to be a commodity parallel arithmetic unit. Moreover, we give an evaluation of our proposal. Our proposal parallelizes an algorithm of multivariate cryptography, and makes it efficient by optimizing the algorithm with GPU.

Keywords: stream cipher, efficient implementation, Multivariate Cryptography, GPGPU.

1. Introduction

1.1. Background

Nowadays cryptography is a necessary technology for network communication. Multivariate cryptography uses multivariate polynominals system as a public key. The security of multivariate cryptography is based on the hardness of solving non-linear multivariate polynomial systems over a finite field [2]. Multivariate cryptography is considered to be a promising tool for fast digital signature, because it requires just computing multivariate polynomial system.

QUAD is a stream cipher, which uses a multivariate quadratic system [4]. Symmetric ciphers are used to authentication schemes [8] and signatures [10]. The security of QUAD depends on the multivariate quadratic (MQ) problem. Therefore QUAD has provable security like public key cryptography though it is a symmetric cipher. QUAD has high security, but it is very slow compared with other symmetric ciphers. When QUAD stream cipher is accelerated, we can realize high security communication with QUAD.

1.2. Related Works

Berbain et al. [3] provided efficient implementation techniques for multivariate cryptography including QUAD stream cipher on CPUs. They implemented 3 cases of QUAD instances, over $GF(2)$, $GF(2^4)$, and $GF(2^8)$. Arditti et al. [1] showed FPGA implementations of QUAD for 128, 160, 256 bits blocks over $GF(2)$. Chen et al. [6] presented throughputs of a GPU implementation of QUAD for 320 bits blocks over $GF(2)$. However the results show that GPU implementations are slower than ideal CPU implementaitons.

Most of these related works just implemented several QUAD instances. They did not evaluate computational costs of QUAD stream ciphers. Only Berbain et al. [3] showed the computational costs of QUAD with n unkonowns and m multivariate quadratics, which are $O(mn^2)$. We extended several implementation strategies for multivariate quadratic of Berbain et al. to GPU implementations and evaluated the computational cost of QUAD [12].

This is an extension work of our previous result [12]. We present extended GPU implementation results from $GF(2)$ case to $GF(2^p)$ cases, and comparisons with other works. Moreover, we refine the evaluations of computational costs of QUAD for general cases and optimized $GF(2)$ cases.

1.3. Motivation

Our goal is to implement efficient QUAD stream cipher. Since QUAD has a rigorous security proof as public key cryptography, we can use a fast and secure cipher when QUAD becomes fast like other stream ciphers.

1.4. Our Contribution

We provide two techniques to implemnt QUAD stream cipher. One is a parallel implementation for computing multivariate polynomials. The other is an optimization technique for implementing QUAD on GPUs.

In this paper, we discuss the computational time for generating keystreams of QUAD in more detail than [12]. Moreover, we report results of implementation of QUAD stream cipher over $GF(2)$, $GF(2^2)$, $GF(2^4)$, and $GF(2^8)$ on GPU.

2. CUDA Computing

2.1. GPGPU

Originally, Graphical Processing Units (GPUs) are process units for drawing the computer graphics. Recently, some online network games and simulators require very high level computer graphics. The GPU performance is growing to satisfy such requirements. Therefore, GPU has a large amount of power for computation.

GPGPU is a technique for any general process by using GPUs. In cryptography, it is used for some implementations. For example, Manavski proposed

an implementation of AES on GPU, which is 15 times faster than an implementation on CPU, in 2007 [7]. Moreover, Osvik et al. presented a result of an over 30 Gbps GPU implementations of AES, in 2010 [9]. On the other hand, the GPGPU technique is also used for cryptanalysis. Bonenberger et al. used a GPU to generating polynomials of the General Number Field Sieve [5].

Because GPUs are designed based on SIMD, it is better to handle several simple tasks simultaneously. On the other hand, the performance of a GPU core is not higher than CPU. Therefore, if we use GPU for sequencial processing, it is not effective. In the GPGPU techniques, how to parallelize algorithms is an important issue.

2.2. CUDA API

CUDA is a development environment for GPU, based on C language and provided by NVIDIA. Pregnancy tools for using GPU have existed before CUDA is proposed. However, such tools as OpenGL and DirectX need to output computer graphics while processing work. Therefore, these tools are not efficient. CUDA is efficient, because CUDA uses computational core of GPU directly.

In CUDA, hosts correspond to computers, and devices correspond to graphic cards. CUDA works by making the host control the device. Kernel is a function the host uses to control the device. Because only one kernel can work at a time, a program requires parallelizing processes in a kernel. A kernel handles some blocks in parallel. A block also handles some threads in parallel. Therefore a kernel can handle many threads simultaneously.

NVIDIA GeForce GTX 480 In this paper, we use a GPU, which is named GeForce GTX 480 by NVIDIA. It is a high-end GPU of GeForce 400 series released in March 2010. GTX 480 is constructed by a Fermi architecture which is a new architecture. GTX 480 uses 15 streaming multiprocessors(SMs), which are constructed by 32 cuda cores instead of by 8 cuda cores.

3. Multivariate Cryptography

3.1. Cryptography

Cryptography is a technique to prevent data from being leaked by adversaries. Mainly, we use it on network communication. Cryptography is categorized into two types, one is symmetric key cryptography and the other is asymmetric key cryptography.

Symmetric Key Cryptography Symmetric key cryptography uses the same keys or functions in encryption and decryption. It has two types, block cipher and stream cipher. Block cipher encrypts message block by block size. Stream cipher uses pseudorandom number generators as keystream generators. A message is encrypted with keystream in sequence.

Asymmetric Key Cryptography Asymmetric key cryptography has two types of keys. One is a public key, which is used for encryption. The other is a private key for decryption.

3.2. Multivariate Polynomial Systems

Multivariate Polynomials We use a finite field $GF(q)$. Let $X = (x_1, \ldots, x_n)$ be a n-tuple variable of $GF(q)$, we describe monomials as $\alpha^{(k)}_{s_1,\ldots,s_k} \prod_{i=1}^{k} x_{s_i}$, where $k \geq 0$, $1 \leq s_1 \leq \cdots \leq s_k \leq n$. $\alpha^{(k)}_{s_1,\ldots,s_k}$ is a coefficient of a k-dimensional monomial. Therefore, they consist of a coefficient and k variables. If a dimension of a monomial is 0, it is called a constant.

Multivariate polynomials contain a sum of monomials. Let $f^{(d)}(X)$ be a d-dimensional multivariate polynomial. It is denoted as Formula (1),

$$f^{(d)}(X) = \alpha^{(0)} + \sum_{k=1}^{d} \sum_{1 \leq s_1 \leq \cdots \leq s_k \leq n} \alpha^{k}_{s_1,\ldots,s_k} \prod_{i=1}^{k} x_{s_i}. \tag{1}$$

Especially when $k = 2$, polynomials are called quadratics. Let $Q(X)$ be a multivariate quadratics, and Formula (2) presents $Q(X)$ with n unknowns,

$$Q(X) = \sum_{1 \leq i \leq j \leq n} \alpha_{i,j} x_i x_j + \sum_{1 \leq i \leq n} \beta_i x_i + \gamma, \tag{2}$$

where $\alpha_{i,j} = \alpha^{(2)}_{i,j}$, $\beta_i = \alpha^{(1)}_{i}$ and $\gamma = \alpha^{(0)}$.

Multivariate Polynomials Systems and MP Problem A multivariate polynomial $f(X)$ can be considered as a multivariate function, which computes results with some given variables. A multivariate polynomial system is a group of such functions. The multivariate polynomial system $MP(X)$ which is constructed with n unknowns and m d-dimensional polynomials is given in Formula (3).

$$MP(X) = \{f^{(d)}_1(X), \ldots, f^{(d)}_m(X)\} \tag{3}$$

A multivariate quadratic system is a special case of the multivariate polynomial system, which uses quadratic functions $Q(X)$. The multivariate quadratic system $MQ(X)$ which is constructed with n unknowns and m quadratics is also given in Formula (4).

$$MQ(X) = \{Q_1(X), \ldots, Q_m(X)\} \tag{4}$$

We assume that $MP(X)$ is constructed with m d-dimensional polynomials. MP problem is to find $X = (x_1, \ldots, x_n)$ where $f^{(d)}_i(X) = 0$ for all $1 \leq j \leq m$. MP problem on a finite field is known as an NP-hard problem [11]. We can also define MQ problem for multivariate quadratic systems $MQ(X)$. It is also known as an NP-hard problem. The security of QUAD stream cipher depends on the MQ assumption.

3.3. QUAD Stream Cipher

QUAD is a stream cipher which is proposed by Berbain et al. [4]. However, it is a stream cipher, and the security of it is based on the MQ assumption.

Constructions QUAD uses a n-tuple internal state value $X = (x_1, \ldots, x_n)$ and a random multivariate quadratic system $S(x_1, \ldots, x_n)$ with m multivariate quadratic function $Q(X)$: $GF(q)^n \mapsto GF(q)$, such that

$$S(X) = \{Q_1(X), \cdots, Q_m(X)\}, \tag{5}$$

as a pseudorandom number generator. It is denoted by QUAD(q, n, r), where r is a number of output keystreams, and $r = m - n$. Usually, m is set to kn, where $k \geq 2$, and therefore $r = (k-1)n$.

Keystream Generation Let $m = kn$ and $S(X) = \{Q_1(X), \ldots, Q_{kn}(X)\}$ be divided two parts as $S_{it}(X) = \{Q_1(X), \ldots, Q_n(X)\}$ and $S_{out}(X) = \{Q_{n+1}(X), \ldots, Q_{kn}(X)\}$. The keystream generator of QUAD follows three steps, such that,

Computation Step
The generator computes values of system $S(X)$, where $X = (x_1, \ldots, x_n)$ is a current internal value.

Output Step
The generator outputs keystreams $S_{out}(X) = \{Q_{n+1}(X), \ldots, Q_{kn}(X)\}$ from values of $S(X)$.

Update Step
The current internal value $X = (x_1, \ldots, x_n)$ is updated to a next internal value with a n-tuple value $S_{it}(X) = \{Q_1(X), \ldots, Q_n(X)\}$ from $S(X)$.

The sketch illustrating the keystream generation algorithm is shown in Fig. 1. It indicates that the generator outputs keystreams by repeating the above three steps.

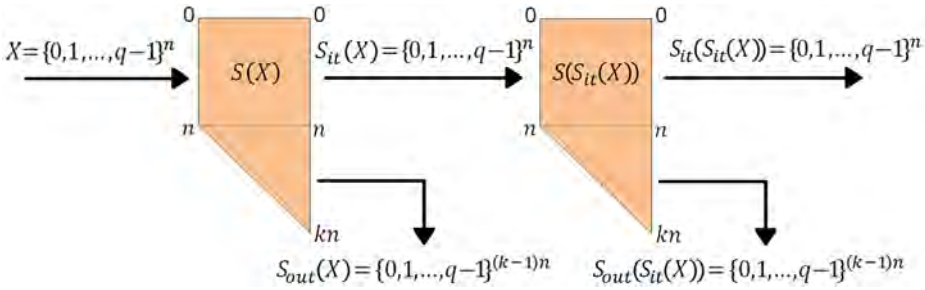


Fig. 1. Generating keystream of QUAD

The generated keystreams are considered to be a pseudorandom bit string and used to encrypt a plaintext with the bitwise XOR operationg.

Key and Initialization of Current State Berbain et al. [4] also provides a technique for initialization of the internal state $X = (x_1, \ldots, x_n)$. For QUAD(q, n, r), we use the key $K \in GF(q)^n$, the initialization vector $IV = \{0,1\}^{|IV|}$ and two carefully randomly chosen multivariate quadratic systems $S_0(X)$ and $S_1(X)$, mapping $GF(q)^n \mapsto GF(q)^n$ to initialize X.

The initialization of the internal state X follows two steps, such that,

Initially Set Step

We set the internal state value X to the key K.

Initially Update Step

We update X for $|IV|$ times. Let i be a number of iterating initially update and $IV_i = \{0,1\}$ be a value of i-th element of IV, and we change the value of X to

- $S_0(X)$, where $IV_i = 0$, and
- $S_1(X)$, where $IV_i = 1$.

Computational Cost of QUAD The computational cost of multivariate quadratics depends on computing quadratic terms. The summation of quadratic terms requires $n(n+1)/2$ multiplications and additions. Therefore the computational costs of one multivariate quadratic is $O(n^2)$. QUAD(q, n, r) requires to compute m multivariate quadratics. Since $m = kn$, the computational cost of generating key stream is $O(n^3)$.

Security Level of QUAD The security level of QUAD is based on the MQ assumption. Berbain et al. [4] prove that solving QUAD needs solving MQ problem. However, according to the analysis of QUAD using the XL-Wiedemann algorithm which was proposed by Yang et al. [14], QUAD$(256, 20, 20)$ has 45-bit security, QUAD$(16, 40, 40)$ has 71-bit security, and QUAD$(2, 160, 160)$ has less than 140-bit security. Actually, secure QUAD requires larger constructions such as QUAD$(2, 256, 256)$, QUAD$(2, 320, 320)$.

4. Strategy

4.1. Existing Methods of Berbain et al.

Berbain et al. [3] provided efficent implementation techniques of computing multivariate polynomial systems for multivariate cryptography. In this paper, we use these strategies from [3].

- Variables are used as vectors. For example, C language defines int as a 32-bit integer variable. Therefore, we can use int as a 32-vector of boolean.
- We precompute each quadratic term. Because in multivariate quadratic systems, we must compute the same $x_i x_j$ for every polynomials, we can make efficient implementation by precomputing quadratic terms.
- We compute only non-zero terms in $GF(q)$. Because the probability of $x_i = 0$ is $1/2$ the probability of $x_i x_j = 0$ is $3/4$. Therefore, we can reduce computational cost to $1/4$.

4.2. Parallelizing on the GPU

In the GPGPU, the most important point is the parallelization of algorithms. Because the performance of GPU cores is worse than that of CPU, serial implementations with GPU are expected to be slower than CPU implementations.

Since all the polynomials of a multivariate quadratic system are independent of each other, parallelization of system is easy. We discuss how to parallelize a multivariate quadratics of system.

Summation of quadratic terms can be considered as summation of every element of a triangular matrix as the left side of Fig. 2. We assume that other elements of matrix are zero. Therefore we can compute summation of quadratic terms as summation of regular matrix as the right side of Fig. 2. Then, we can compute the summation as $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_{i,j}x_ix_j$ in the following method.

1. We compute $S_k(x) = \sum_{i=1}^{n}\alpha_{k,i}x_kx_i$ for all k in parallel.
2. We compute $\sum_{k=1}^{n}S_k(x)$.

However computations increase trivial computations; we can make efficient implementations by parallelization with GPU.

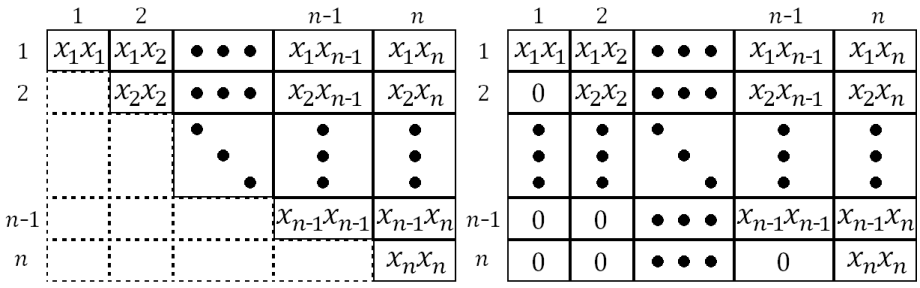


Fig. 2. Left: handling as a triangular matrix. Right: handling as a rectangle matrix with trivial 0 elements.

Next we reduce trivial computations of that way. We reshape a triangular matrix to a rectangular matrix as in Fig. 3, which presents the method of reshaping matrix. By this reshaping, we can reduce efficiently about 25% of the cost for computing an equation of a multivariate quadratic system.

4.3. Optimization on GPU architectures

On GPU implementations, we must consider characteristics of GPU. The strongest point of GPU is the computational power by processing cores, and a core is slower than a CPU. Therefore, non-active cores in a process affect results.

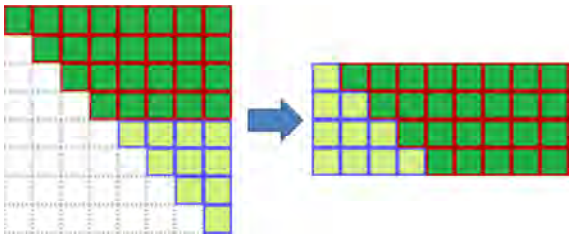


Fig. 3. Reshaping triangle to rectangle

Optimization of Matrix Calculation NVIDIA GeForce GTX 480 has 15 SMs, and every SM has 32 cuda cores. Since every SM handle 32 threads at a time, a process, which handles 32 threads, is not suitable for GPU implementations. Therefore, we should make sure that the number of threads in a process is divisible by 32. In the same way, we should make sure that an algorithm can be handled by 15 SMs in parallel. Finally, an algorithm should be paralleled as a multiple of 32×15.

An n-dimensional triangular matrix has $n(n+1)/2$ elements. Then a long side of a rectangle matrix, which is reshaped by an n-dimensional triangular matrix, has n or $n+1$ elements. However in $GF(2)$, a number of a long side's elements can be counted in a process, counting is a cost of computing a matrix. Therefore, we assume that $n = 30k$ where k is a natural number. In this way, a rectangle matrix is constructed by $15k \times (31k+1)$ in Fig. 4, and we can handle the rectangle matrix as $15k \times 32k$. Moreover, computing k-dimensional square submatrix from the matrix in parallel, we can reduce the $15k \times 32k$ matrix to a 15×32 matrix. Thus we can parallelize calculating of a matrix by 15 SMs and 32 cuda cores per SM.

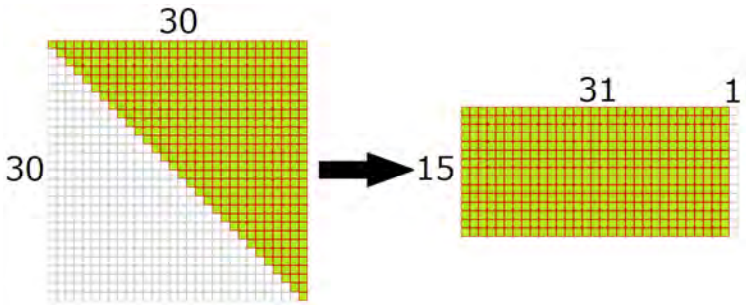


Fig. 4. Handling a $15k \times 32k$ matrix e.g. $k = 1$

Optimizing in Processes For realizing an efficient GPU implementation, we must design an algorithm as a chunk of similar small computings. Moreover, conditional branches are not suited to processing on GPU. Then we handle conditional branches as difference of kernels, and implement them by processing on CPU. In this case, we make kernels by difference of a number of non-zero terms. However, making all kernels every number of non-zero temrs is a heavy cost of implementaions. Therefore, we make kernels just every number of k. E.g. QUAD$(2, 320, 320)$, the maximum of k is 11, thus we have to make only 11 kernels.

4.4. Evaluation of GPU Implementation

Accelerating by strategies of Berbain et al. Originally, QUAD(q, n, n) requires $n(n+1)(n+2)$ additions and $2n^2(n+2)$ multiplications. Moreover, QUAD(q, n, n) requires $2n$ times computation of equations.

By the strategy of Berbain et al. [4], we can reduce multiplications of monomials. A monomial $x_i x_j$ is a common value for each polynomial. Then we need to calculate monomials only once. Since it takes $n(n+1)/2$ multiplications, we reduce multiplications from $2n^2(n+2)$ times to $n^2(n+3)+n(n+1)/2$ times.

Moreover, we can compute some polynomials at a time by vectorization of variables. In case of $q = 2^t$, i.e. using $GF(2^t)$, a 32-bit integer variable handles $32/t$ polynomials. Therefore QUAD(q, n, n) (=QUAD$(2^t, n, n)$) requires $\lceil t/16n \rceil (n+1)(n+2)$ additions and $\lceil t/16n \rceil n(n+3)/2 + n(n+1)/2$ multiplications.

Accelerating by parallelization In GPU implementations, we can parallelize some computational steps of evaluating multivariate polynomial systems.

By computing $x_i x_j$ in parallel for each i's, it takes n multiplications.

The computational cost of summations of row elements in a matrix requires $(n+1)(n+2)/2$ additions and multiplications for $64n/t$ polynomials. By parallelization using C cores in GPU, it takes $\lceil tn(n+2)/32C \rceil (n+1)$ additions and multiplications. Also, the computational cost of summations of column elements in a matrix can be reduced from $\lceil tn/16 \rceil (n+2)$ additions to $\lceil tn/16C \rceil (n+2)/2$ additions.

Actually, GTX 480 has 480 cores. Then it computes all the polynoimals at a time over $GF(2^t)$ field, where $tn/32 \leq 480$. Therefore, QUAD$(2^t, n, n)$ requires $\lceil (n+2)/P \rceil (n+1)+(n+2)/2$ additions and $\lceil (n+2)/P \rceil (n+1)+n$ multiplications for generating keys, where $P = 32C/tn$.

Acclerating for $GF(2)$ By the strategy of Berbain et al. [4], we can compute only non-zero terms in $GF(2)$. Because the probability of $x_i = 0$ is $1/2$ the probability of $x_i x_j = 0$ is $3/4$. Therefore, we can reduce computational cost to $1/4$.

Then we can compute an equation of QUAD$(2, n, n)$ by $\lceil (n+4)/4P \rceil (n+2)/2 + (n+4)/4$ additions and $\lceil (n+4)/4P \rceil (n+2)/2 + n/2$ multiplications, where $P = 32C/tn$.

Suppose the number of non-zero variables is $30k$. Then k-dimensional submatrix requires $k \times k$ additions. Using our strategy, we can compute the summations of QUAD$(2, n, n)$ by $(\lceil (15 \times 32nk)/16C \rceil + \lceil 15 \times 32n/16C \rceil)k + 15 + 32$ additions. Since the GTX 480 has 480 cores (i.e., $C = 480$), it requires $(\lceil nk/16 \rceil + \lceil n/16 \rceil)k + 47$ additions. For example, QUAD$(2, 320, 320)$ requires $20k^2 + 20k + 47$ additions.

5. Experiments

In this section, we present and discuss results of experiments. We used NVIDIA GeForce 480 GTX as a GPU, and also used Intel Core $i7\ 875K$ as a CPU. Moreover, the memory of implementation environment was 8GB.

5.1. Experimentations of Encryption

We implement QUAD$(2, n, n)$ on CPU and GPU; set $n = 32, 64, 128, 160, 256, 320, 512$, and measure the time of encrypting 5MB file. Also, we implement QUAD$(2^t, n, n)$ on GPU; set $t = 2, 4, 8$ and $n = 32, 64, 128, 160, 256, 320, 512$, and measure the time of encrypting messages 1000 times.

Moreover, we optimized GPU implementation of QUAD$(2, n, n)$ by our optimization strategies, and also measure the time of encrypting 5MB file.

5.2. Results

We present the results of QUAD$(2, n, n)$ implementations in Table 1, and QUAD$(2^t, n, n)$ in Table 2. Table 1 presents the time of encrypting 5MB files and throughputs of implementations of each QUADs. In Table 2 we compared our implementations of QUAD$(2^p, n, n)$, where $p = 1, 2, 4, 8$.

Table 1. Results of QUAD implementations.

QUAD(q,n,r)	Encryption time(s)		Throughputs(Mbps)	
	CPU	GPU	CPU	GPU
$(2, 32, 32)$	0.35	66.02	114.286	0.606
$(2, 64, 64)$	13.56	46.58	2.949	0.859
$(2, 128, 128)$	52.56	36.82	0.761	1.086
$(2, 160, 160)$	82.07	36.23	0.487	1.104
$(2, 256, 256)$	206.80	35.87	0.193	1.115
$(2, 320, 320)$	326.52	38.96	0.123	1.027
$(2, 512, 512)$	858.20	72.80	0.047	0.549

The image results of QUAD$(2, n, n)$ are shown in Fig. 5. Encryption time of CPU implementations follows square of n. However the computational cost of generating keystream is $O(n^3)$, number of generating keystream follows n.

On the other side, the results of GPU Implementations are not almost different. NVIDIA GeForce GTX 480 can use 15 blocks, and every block can use 32 threads, programs can handle 480 processes at the same time. QUAD$(2, n, n)$ requires $\lceil (n+4)/4P \rceil (n+2)/2 + (n+4)/4$ additions and $\lceil (n+4)/4P \rceil (n+2)/2 + n/2$ multiplications for generating keystreams, where $P = 32C/tn$. However when $(n+4)/4P \leq 1$ (i.e., $tn(n+4) = 128C$), we can generate keystreams of QUAD$(2, n, n)$ by only $(3n+8)/4$ additions and $n+1$ multiplications. On the GTX 480, we can set $C = 480$. Then we have that $n \leq 247$. Therefore the computational cost of QUAD$(2, n, n)$ is proportional to the number of unknowns n, and it generates keystreams in stable throughputs, when $n \leq 247$. Actually, we can see the decrease of the throughput of QUAD$(2, n, n)$ between $n = 256$ and 320.

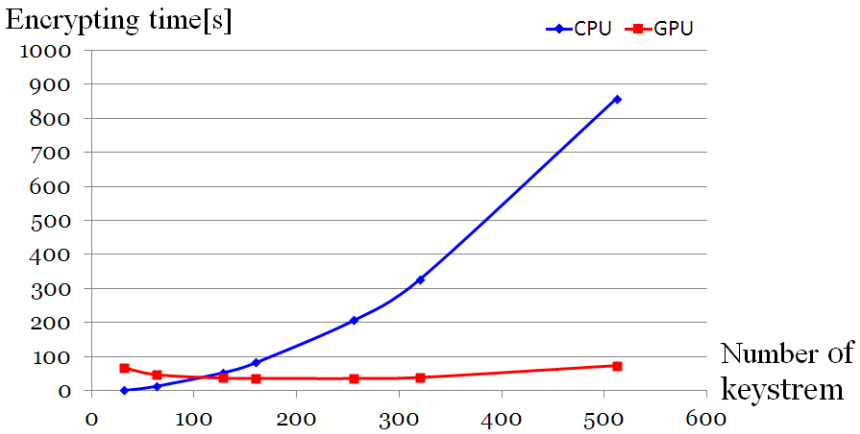


Fig. 5. Comparison of encryption time between CPU and GPU.

Table 2 and Fig. 6 also show the decrease of throughputs of QUAD$(2^t, n, n)$, where $t = 1, 2, 4, 8$ and $n = 32, 64, 128, 160, 256, 320, 512$. Especially, Fig. 6 shows that the higher the degree of $GF(2^t)$ is, the more drastic the decrease of the throughput of QUAD is. For example, when $n > 44$, the computational cost of QUAD over $GF(2^8)$ follows $O(n^3)$. Therefore the larger the number of unknowns n is, the slower the throughput of QUAD is.

Moveover, we provide the results of our optimized implementation with the results of a non-optimized implementation Table 3 shows that the throughputs of our optimized GPU implementations with the throughputs of our non-optimized GPU implementations, and ratio of GPU and CPU implementations. The results of GPU implementations show that the throughputs of optimized QUAD

Table 2. Implementation of QUAD$(2^t, n, n)$.

Unknowns n	Througputs(Mbps)			
	$GF(2)$	$GF(2^2)$	$GF(2^4)$	$GF(2^8)$
32	0.606	2.517	4.128	6.110
64	0.859	3.353	4.810	4.863
128	1.086	3.271	4.424	1.405
160	1.104	2.603	2.570	0.827
256	1.115	1.161	0.858	0.249
320	1.027	0.581	0.473	0.189
512	0.549	0.177	0.146	0.072

was improved compared with the non-optimized implementation. Our optimized implementations are 2.0 to 4.4 times faster than non-optimized implementation. We infer that the main cause of accelerated encryption time is due to the optimizations to handle $n = 30k$. Because computing n elements in serial is heavy to GPU, we can reduce serial computation by such handling.

Table 3. Throughputs of QUAD Implementations with Optimization and Non-Optimization.

QUAD(q,n,r)	Throughputs(Mbps)		Speed Up
	Non-optimized	Optimized	Rate(times)
$(2, 32, 32)$	0.606	1.234	x2.037
$(2, 64, 64)$	0.859	2.319	x2.699
$(2, 128, 128)$	1.086	4.132	x3.805
$(2, 160, 160)$	1.104	4.872	x4.413
$(2, 256, 256)$	1.115	4.115	x3.691
$(2, 320, 320)$	1.027	3.656	x3.560
$(2, 512, 512)$	0.549	1.494	x2.722

In Table 4, we compared related works Berbain et al. [3], Arditti et al. [1], and Chen et al. [6] for QUAD$(2, n, n)$, where $n = 128, 160, 256, 320, 512$, and compared throughputs of QUAD$(2^4, 40, 40)$ and QUAD$(2^{16}, 20, 20)$ with Berbain et al. [3]. According to Table 4 although our optimized implementation of QUAD$(2, 160, 160)$ is slower than Berbain et al. [3], our optimized implementation of QUAD$(2, 256, 256)$ and QUAD$(2, 320, 320)$ is faster than the GPU implementation of Arditti et al. [1] and Chen et al. [6]. [3] These results show that our optimized GPU implementation technique is suited to large QUAD constructions.

[3] Chen et al. also presented the CPU implmentations result of QUAD stream cipher. Although 6.10Mbps is the fastest known result, it was just a theoretic estimate [13] without a real implementation.

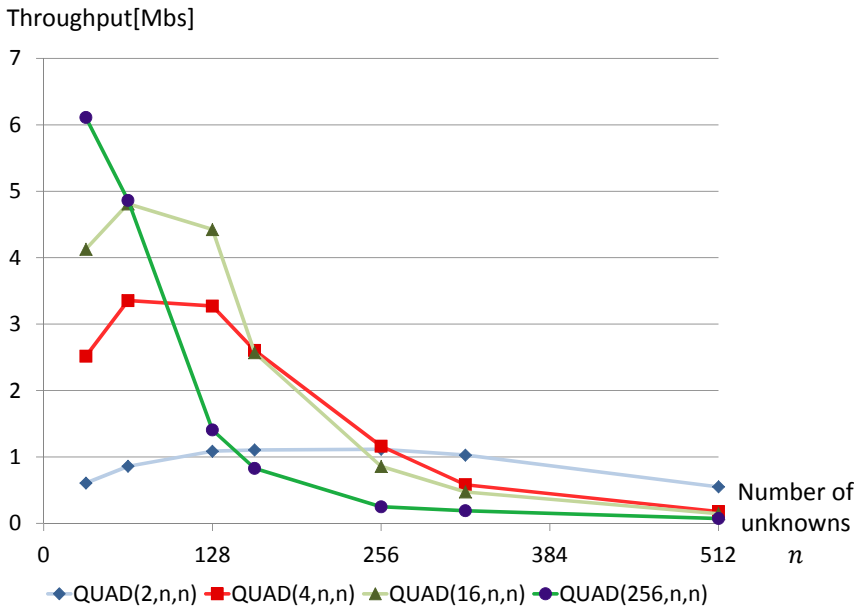


Fig. 6. Throughputs of QUAD implementations over $GF(2^t)$ $(t = 1, 2, 4, 8)$

On the other hand, Table 4 shows that both of our GPU implementation both for QUAD$(2^4, 40, 40)$ and QUAD$(2^8, 20, 20)$ are slower than Berbain et al. [3]. We infer that our implementation strategies are very specialized to $GF(2)$.

6. Conclusion

We presented and evaluated the GPU implementation techniques for QUAD stream cipher. Also we provided optimization techniques of QUAD to suit NVIDIA GeForce GTX 480.

Moreover, we carried out the experiments on the implementations of QUAD over $GF(2)$, $GF(2^2)$, $GF(2^4)$ and $GF(2^8)$. As a result, the larger the number of unknowns n is, the slower the throughput of QUAD is. However, when $tn(n+2) \leq 32C$, it is stable. The condition for stable throughputs depends on the number of cores C. Although the GTX 480 has only 480 cores, the GTX 680, which is the latest high-performance GPU, has 1536 cores. Therefore, the throughput of QUAD$(2, n, n)$ is stable if $n \leq 439$. We expect that future GPUs allow efficient implementation of QUAD$(2, 512, 512)$ and more heavy constructions of QUAD.

Table 4. Comparison QUAD Implementations with Related Works.

	Throughputs(Mbps)			
	Our Works	Berbain et al. [3]	Arditti et al. [1]	Chen et al. [6]
	GPU(Optimized)	CPU	FPGA	GPU
QUAD$(2, 128, 128)$	4.132	N.A.	4.1	N.A.
QUAD$(2, 160, 160)$	4.872	8.45	3.3	N.A.
QUAD$(2, 256, 256)$	4.115	N.A.	2.0	N.A.
QUAD$(2, 320, 320)$	3.656	N.A.	N.A.	2.6
QUAD$(2^4, 40, 40)$	4.320	23.59	N.A.	N.A.
QUAD$(2^8, 20, 20)$	3.895	42.15	N.A.	N.A.

Acknowledgments. This work is partially supported by Japan Science and Technology agency (JST), Strategic Japanese-Indian Cooperative Programme on Multidisciplinary Research Field, which combines Information and Communications Technology with Other Fields, entitled "Analysis of Cryptographic Algorithms and Evaluation on Enhancing Network Security Based on Mathematical Science." The authors are grateful to three anonymous referees of ComSIS-2013 for improving this article.

References

1. Arditti, D., Berbain, C., Billet, O., Gilbert, H.: Compact fpga implementations of quad. In: Proceedings of the 2nd ACM symposium on Information, computer and communications security. pp. 347–349. ACM (2007)
2. Bard, G.: Algebraic cryptanalysis. Springer (2009)
3. Berbain, C., Billet, O., Gilbert, H.: Efficient implementations of multivariate quadratic systems. In: Selected Areas in Cryptography. pp. 174–187. Springer (2007)
4. Berbain, C., Gilbert, H., Patarin, J.: Quad: A practical stream cipher with provable security. Advances in Cryptology-EUROCRYPT 2006 pp. 109–128 (2006)
5. Bonenberger, D., Krone, M.: Factorization of rsa-170. Tech. rep., Tech. rep., Ostfalia University of Applied Sciences (2010)
6. Chen, M.S., Chen, T.R., Cheng, C.M., Hsiao, C.H., R., N., Yan, B.Y.: What price a provably secure stream cipher? (2010)
7. Manavski, S.: Cuda compatible gpu as an efficient hardware accelerator for aes cryptography. In: Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on. pp. 65–68. Ieee (2007)
8. Miyaji, A., Rahman, M.S., Soshi, M.: Efficient and low-cost rfid authentication schemes. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 2(3), 4–25 (2011)
9. Osvik, D., Bos, J., Stefan, D., Canright, D.: Fast software aes encryption. In: Fast Software Encryption. pp. 75–93. Springer (2010)
10. Pakniat, N., Eslami, Z.: A proxy e-raffle protocol based on proxy signatures. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 2(3), 74–84 (2011)
11. Patarin, J., Goubin, L.: Asymmetric cryptography with s-boxes is it easier than expected to design efficient asymmetric cryptosystems? Information and Communications Security pp. 369–380 (1997)

12. Tanaka, S., Nishide, T., Sakurai, K.: Efficient implementation of evaluating multivariate quadratic system with gpus. In: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on. pp. 660–664. IEEE (2012)
13. Yang, B.Y.: (private communicate) (2011)
14. Yang, B., Chen, O., Bernstein, D., Chen, J.: Analysis of quad. In: Fast Software Encryption. pp. 290–308. Springer (2007)

Satoshi Tanaka received a B.E. degree from National Institution for Academic Degrees and University Evaluation, Japan in 2010, and an M.E. degree from Kyushu University, Japan in 2012. Currently he is a candidate for the Ph.D. in Kyushu University. His primary research is in the areas of cryptography and information security.

Takashi Nishide received a B.S. degree from the University of Tokyo in 1997, an M.S. degree from the University of Southern California in 2003, and a Dr.E. degree from the University of Electro-Communications in 2008. From 1997 to 2009, he had worked at Hitachi Software Engineering Co., Ltd. developing security products. Since 2009, he has been an assistant professor in Kyushu University. His primary research is in the areas of cryptography and information security.

Kouichi Sakurai is Professor of Department of Computer Science and Communication Engineering, Kyushu University, Japan since 2002. He received B.E., M.E., and D.E. of Mathematics, Applied Mathematics, and Computer Science from Kyushu University in 1982, 1986, and 1993, respectively. He is interested in cryptography and information security. He is a member of IPSJ, IEEE and ACM.

Received: October 06, 2012; Accepted: January 31, 2013.

DOI: 10.2298/CSIS121114041G

A Hybrid Approach to Secure Hierarchical Mobile IPv6 Networks

Tianhan Gao[1], Nan Guo[2,*], Kangbin Yim[3]

[1] Faculty of Software College, Northeastern University,
110819 Shenyang, China
gaoth@mail.neu.edu.cn
[2] Faculty of Information Science and Engineering College, Northeastern University,
110819 Shenyang, China
guonan@ise.neu.edu.cn
[3] Faculty of Information Security Engineering, Soonchunhyang University,
336745 Asan, Korea
Yim@sch.ac.kr

Abstract. Establishing secure access and communications in a hierarchical mobile IPv6 (HMIPv6) network, when a mobile node is roaming into a foreign network, is a challenging task and has so far received little attention. Existing solutions are mainly based on public key infrastructure (PKI) or identity-based cryptography (IBC). However, these solutions suffer from either efficiency or scalability problems. In this paper, we leverage the combination of PKI and certificate-based cryptography and propose a hierarchical security architecture for the HMIPv6 roaming service. Under this architecture, we present a mutual authentication protocol based on a novel cross-certificate and certificate-based signature scheme. Mutual authentication is achieved locally during the mobile node's handover. In addition, we propose a key establishment scheme and integrate it into the authentication protocol which can be utilized to set up a secure channel for subsequent communications after authentication. As far as we know, our approach is the first addressing the security of HMIPv6 networks using such a hybrid approach. In comparison with PKI-based and IBC-based schemes, our solution has better overall performance in terms of authenticated handover latency.

Keywords: hierarchical mobile IPv6, mutual authentication, identity-based cryptography, certificate-based cryptography, cross-certificate

1. Introduction

MIPv6 [1], developed by Internet Engineering Task Force (IETF), has been recognized as the best solution for linking different mobile networks. More

* Corresponding author. Tel.: +8624-83681822. E-mail: Guonan@ise.neu.edu.cn

specifically HMIPv6 extends MIPv6 [2] by introducing local mobility management. However, HMIPv6 does not specify nor endorse any particular security mechanisms which may thus result in a variety of threats such as redirection, denial of service (DoS), man in the middle attacks, and resource misuse [3, 4]. Consequently, how to secure HMIPv6 network is currently the focus of intense attention in the research community.

In order to securely deploy HMIPv6 services, mutual authentication between mobile nodes and access points in the visited networks is essential. Moreover, it is crucial that secure channels be dynamically set up with respect to key establishments among participants for subsequent communications after a successful authentication.

The general approach for achieving mutual authentication and secure channels is based on the use of a public key infrastructure (PKI) [5]. In this approach, mutual authentication between the mobile node and the access point is performed by verifying the other party's digital signature and public key certificate (PKC) issued by a certificate authority (CA). Communications can also be protected via public key cryptography. As a result, no shared keys or security associations are needed for the mobile node and the access point. They only need to have their own private and public key pair. However, the major drawback of a PKI solution is that if the mobile node and the access point belong to different trust domains that have different CAs, they have to piggyback and verify a long PKC chain which typically results in a heavy burden on each side and affects performance. Another obstacle that impedes PKI's employment in HMIPv6 networks is the overhead due to the transmission and storage of PKC. Frequent changes in network topology make the management of PKC even harder.

Some of the drawbacks of PKI have been addressed by identity-based cryptography (IBC) [13]. The use of IBC protocols greatly simplifies the key management procedures of conventional PKI and eliminates the need for PKC. Therefore, several schemes [8-11] have been proposed to integrate IBC into HMIPv6 network for authentication and key management services. In such schemes, the private key generator (PKG) introduced by IBC is used for distributing secret keys to all entities in a HMIPv6 network. Mutual authentication and secure communications are then directly implemented between mobile nodes and access points through IBC-based signature and encryption mechanisms without the help of PKI. However these schemes are based on the assumption that the PKG is trusted by all the participants, which makes them only suitable for small scale mobile networks. Moreover, the IBC protocols adopted by these schemes have also some intractable problems, such as the secret key escrow and distribution problems as well as the computational costs incurred by pairing-based operations.

In general, although PKI suffers from a heavy maintenance workload, it has been widely deployed in real world and can support authentication even for large scale, hierarchical groups. On the other hand, IBC supports an efficient key management but is only suitable for a closed organization where the PKG is completely trusted by every entity. Consequently, a promising approach is to concatenate these two techniques in order to gain the benefits

from both. This combination can support secure communications between group managers already in possession of certificates, as well as between individual users without certificates. Therefore a few approaches have been proposed that combine PKI and IBC [14-16]. Their focus is however on scalability and they do not address security in HMIPv6 networks. It is thus crucial to develop a hybrid PKI and IBC scheme for securing HMIPv6 networks.

In this paper, we present an authentication protocol for HMIPv6 roaming service based on the combination of PKI and IBC. A novel signature scheme based on cross-certificate [24] and certificated-based signature [22, 23] is proposed as building block for our protocol. Mutual authentication is achieved locally within the access network. The proposed protocol presents a more efficient PKC management because of the cross-certificate mechanism. Also the secret key escrow and distribution problems inherited from IBC are addressed by the use of certificate-based cryptography. A key establishment scheme is also incorporated into our protocol to build a secure channel for subsequent communications. To further improve the efficiency of our protocol, we integrate the authentication operations into the HMIPv6 mobility management process. Performance analysis demonstrates that our proposed protocol outperforms existing ones in terms of handover latency during authentication.

The rest of this paper is organized as follows. Section 2 presents the HMIPv6 and certificate-based cryptographic primitives. We describe our proposed hybrid security architecture in Section 3 as well as the mutual authentication and key establishment scheme for HMIPv6 roaming service in Section 4. Performance analysis of our scheme is elaborated in Section 5. In section 6, we assess how our scheme satisfies the security requirements of HMIPv6 networks. Section 7 discusses the related work. Finally, we conclude the paper in Section 8.

2. Background

In this section, we provide an overview of the HMIPv6 protocol and certificate-based cryptography for readers to better understand our constructions.

2.1. HMIPv6 networks

To alleviate the latency and the amount of the signaling messages occurring during handover, HMIPv6 has been adopted by IETF as the hierarchical mobility management enhancement for MIPv6. A new entity, called mobile anchor point (MAP), is introduced, which is a mobility agent in charge of certain access routers (ARs). The MAP and these routers form an administrative MAP domain. According to HMIPv6, each mobile node (MN) is

addressable by two types of address on the visited link: the on-Link Care-of Address (LCoA), and the Regional Care-of Address (RCoA). The LCoA is configured based on the mobile node's interface, whereas the RCoA is an address on the MAP's subnet. As shown in Fig.1, a mobile node entering a MAP domain will receive a router advertisement (RA) with which it can configure its RCoA and LCoA. Thereafter, the mobile node sends a remote binding update (RBU) to its home agent (HA) in its home domain and its correspondent nodes whereby to bind its RCoA with its home address. In the meantime, the mobile node registers its LCoA with the MAP through a local binding update (LBU). The home agent intercepts the initial packets and tunnels them to the mobile node's RCoA. Function as a local home agent, the MAP will receive all the packets on behalf of the mobile node and will then encapsulate and forward them to the mobile node's current LCoA. The subsequent packets will directly hit the mobile node's RCoA by means of route optimization. If the mobile node moves within the MAP domain, only the LBU should be sent to the MAP in order to register its new LCoA. The RCoA remains unchanged as long as mobile node stays in the current MAP domain. As a consequence, the delays and signaling overhead induced by the RBU can be considerably reduced through such local mobility management strategy. With this salient feature, HMIPv6 is expected to become the fundamental support for next generation mobile networks.

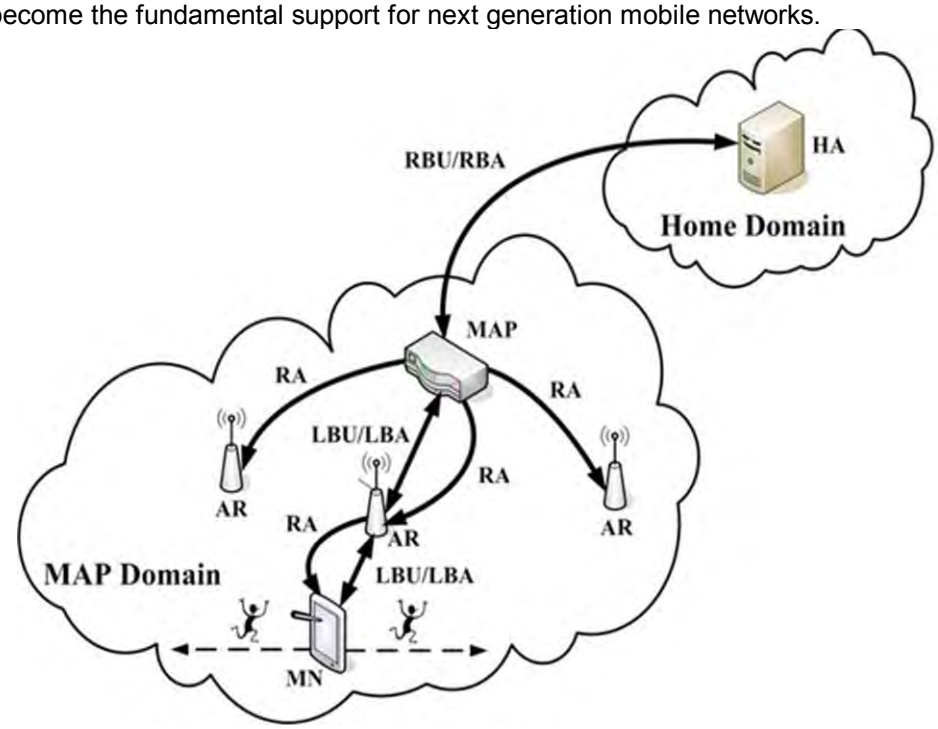


Fig 1 HMIPv6 network

2.2. Bilinear pairings

Let G be an additive group and G_T be a multiplicative group of the same prime order q. Let I_G and I_{GT} be the generator of G and G_T respectively. Assume that the discrete logarithm problem [21] is hard in both G and G_T. A mapping $\hat{e}: G \times G \to G_T$ which satisfies the following properties is called bilinear pairing:

(1) *Bilinear*: *For all* $P,Q \in G$ *and* $a,b \in Z_q^*$, $\hat{e}(aP,bQ) = \hat{e}(bP,aQ) = \hat{e}(P,Q)^{ab}$, *where* $Z_q^* = \{1,2,...q-1\}$;

(2) $Non-degenerate$: $\hat{e}(P,Q) \neq I_{GT}$;

(3) *Computable*: *For all* $P,Q \in G$, *there is an efficient approach to compute* $\hat{e}(P,Q) \in G_T$.

The Weil and Tate pairing [20] on supersingular elliptic curves can be modified to construct such bilinear pairing. Most literature IBC-based schemes employ these pairings as primitives [35].

2.3. Certificate-based Cryptography

In 1984, Shamir proposed the concept of identity-based cryptography (IBC) [13] which significantly reduced the system complexity and the cost for managing the public key compared with PKI. However, a major drawback of IBC is that the PKG can access all the communications among users, and thus can yield any user's secret key. Secret key escrow problem is inherent and in addition the secret keys must be sent over secure channels, making key distribution difficult.

To fill the gap between traditional PKI and IBC, the notion of certificate-based encryption (CBE) [21] was proposed by Gentry in 2003. Certificate-based Cryptography (CBC) combines PKI and IBC and consists of a CA and a set of users. Each user generates its own private and public key pair and requests a certificate from the CA. The CA uses the private key generation algorithm of the Boneh-Franklin IBE scheme [17] as well as the BLS scheme [20] to generate certificates for the users. Such approach provides an implicit certification by the fact that the signing key is composed of the certificate and the secret key generated by user. Moreover, it solves the inherent key escrow problem of IBC. Although the CA knows the certificate of user, it yet cannot forge the signature since it does not know the user's secret key.

Certificate-based signature (CBS) [22, 23], a fundamental branch of CBC, can provide high level of trust along with the shorter length and more efficient verification. It is especially useful in those environments where the computation power is very limited, or communication bandwidth is very expensive. Mobile networks are a good example of such environments. As the verification is efficient, the impact of verification on energy consumption is very low. In addition, the elimination of certificates from the verification process reduces the amount of information that needs to be transmitted thus

reducing the communication overhead. In the case of wireless mobile networks, communication bandwidth is a very expensive resource. The formal CBS scheme that we adopt in our work is specified as following algorithms.

CBS_Setup.

The CA takes as input a security parameter 1^{k1} and returns SK_C (the CA's master secret) as well as the public parameters params that include the CA's public key PK_C.

CBS_GenCert.

The user takes as input a security parameter 1^{k2} and returns a private key SK_U and a public key PK_U (the user's private and public key pair). The CA uses SK_C, params, i, PK_C and PK_U at the start of time period i to create $Cert_i'$ which is sent to the user. Then the user computes $Cert_i$ using params, i, $Cert_i'$ and (optionally) $Cert_{i-1}$ at the start of time period i.

CBS_Sign.

To sign a message *m* with params, $Cert_i$, SK_U in time period i. The signer computes the temporary signing key $SK = f(SK_U, Cert_i)$ where *f* is a public algorithm, and outputs a signature σ.

CBS_Verify.

To verify σ, the verifier takes σ, m, i, PK_C, PK_U as input and outputs a binary value 0 (invalid) or 1 (valid).

3. Network architecture and novel signature scheme

In this section we present the details of our approach. We first introduce our hierarchical security architecture for HMIPv6 networks which concatenates PKI and CBC. Then we propose a novel PKI-CBS-based signature scheme (PCS) under the proposed architecture in order to achieve mutual authentication for HMIPv6 networks.

3.1. Concatenated security architecture

As shown in Fig.2, our proposed architecture has three tiers. The top tier comprises the CAs and the repositories forming the trust infrastructure. The CAs are the trust authorities for the domain managers, while the repository stores the PKCs of CAs. Each CA can set up trust relationships with other CAs through cross-certificates as long as the underlying domains have roaming agreements. For example, consider Fig.2 and assume that the home agent (denoted by HA in Fig.2) has a roaming agreement with MAP1.Then CA1 can issue a PKC for CA2 and register it into the repository, and vice versa. Domain managers reside in the second tier. From the CA point of view, domain managers are PKI-aware users with PKCs issued by CAs. Nonetheless, from the domain perspective, domain managers are trust anchors of end-users (that is, mobile nodes and access routers) inside

domains which form the bottom tier of the architecture. We assume that all nodes within each domain support CBC operations. This implies that the domain managers also have identity-based key pairs and are able to issue certificates to end-users based on CBC. Moreover, as the signing and verifying operations in CBS depend on the same set of public parameters, the public parameters derived from different domains must be certified, which in our scheme is achieved by embedding the parameters into the domain manager's PKC.

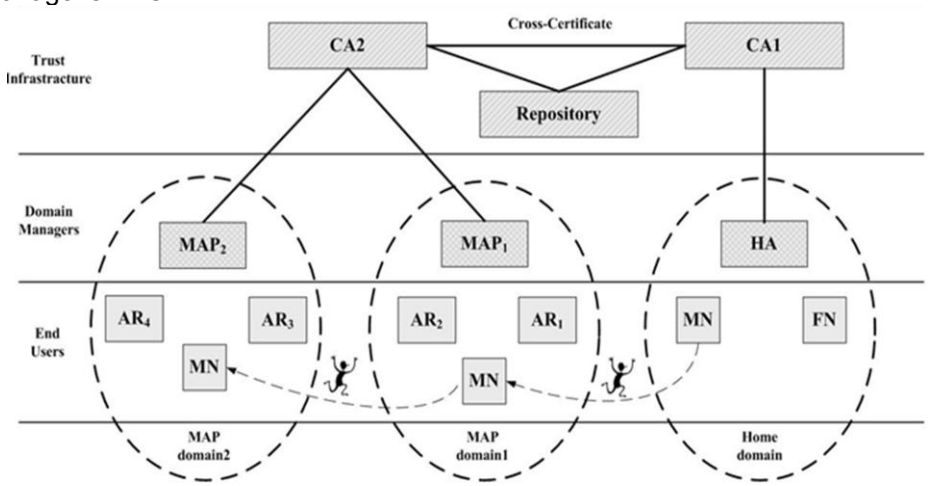


Fig 2 Concatenated security architecture

In short, the cross-domain trust of our architecture relies on cross-certificate between CAs at the trust infrastructure level which makes the architecture appropriate for large scale deployment, whereas the trust relationship inside each domain is achieved through CBC that is simple from the management point of view and suitable for bandwidth-limited wireless networks as well as computational constrained mobile nodes. We also assume that domain managers and their own end-users pre-share a symmetric key to build secure channels for subsequent communications. For the purpose of clarity, the notations and acronyms, used in the rest of the presentation, are listed in Tab.1.

Tab.1. Notations and acronyms

Notations	Meaning
DM	Domain manager, includes home agent (HA) and MAP
Domain_DM	Administrative domain managed by DM
User	End-user within Domain_DM, includes mobile node (MN) and access router (AR)
PKC_A	X.509 format PKC of entity A
Cert_User	CBC-based certificate of user issued by DM
ID_A	Identity information of entity A

PK_A	Public key of entity A
SK_A	Private key of entity A
$PARA_{DM}$	Public parameters of Domain_DM
$User_{INFO}$	Related information of user, includes ID_{User}, PK_{User} and PKC_DM
P_{User}	Hash value of $User_{INFO}$
$\{M\}_{\alpha\_Sign\_Signer}$	Signer signs message M with algorithm α
$\{\sigma\}_{\beta\_Verify\_Verifier}$	Verifier verifies signature σ with algorithm β
K_{A-B}	Shared key between entity A and entity B
SEK_{A-B}	Session key between entity A and entity B
TS	Timestamp
TP	Time period
A→B:[M]	Entity A sends message M to entity B through unsecure channel
A⇨B:[M]	Entity A sends message M to entity B through secure channel
M1,M2	Concatenation of two messages, M1 and M2

3.2. PKI-CBS-based signature scheme (PCS)

Roughly speaking, PCS is constructed by merging cross-certificates and CBS. The scheme consists of the following algorithms.

PCS_Setup.

DM initializes the following system parameters:

Additive group G_1 and multiplicative group G_2 of the prime order q, as well as a bilinear pairing $\hat{e}: G_1 \times G_1 \to G_2$;

Arbitrary $P \in G_1$, $SK_{DM} \in Z_q^*$ and $PK_{DM} = SK_{DM} \cdot P$;

Hash functions $H_1 : \{0,1\}^* \to G_1$, $H_2 : \{0,1\}^* \times G_1 \to Z_q^*$, $H_3 : \{0,1\}^* \to Z_q^*$, $H_4 : G_2 \to \{0,1\}^*$;

Time period TP_i.

DM publishes PK_{DM} and $PARA_{DM}$ = *(G_1, G_2, $\hat{e}$, P, TP_i, H_1, H_2, H_3, H_4)*, where *H_3, H_4* are used for the mutual authentication protocols (described in the following sections).

PCS_Cross-certificate.

CA first generates a public and private key pair (PK_{CA} and SK_{CA}). If two DMs (DM_i and DM_j) have roaming agreement, their CAs (CA_i and CA_j) issue a PKC to each other as below:

CA_i exchanges PKC with CA_j;

CA_i issues PKC_CA_j which includes PK_{CAj} and registers it to repository;

CA_j issues PKC_CA_i which includes PK_{CAi} and registers it to repository.

PCS_PKI-cert.

CA checks DM's identity (ID_{DM}) and issues PKC_DM to DM which includes ID_{DM}, PK_{DM} and $PARA_{DM}$.

PCS_CBC-cert.

User chooses the secret key SK_{User} and computes $PK_{User}=SK_{User} \cdot P$. DM checks User's identity (ID_{User}) and issues $User_{INFO}$= (ID_{User}, PK_{User}, PKC_DM) as well as Cert_User= $SK_{DM} \cdot P_{User}$ to User, where $P_{User} = H_1(TP_i, User_{INFO})$. Afterwards, User computes its signing key, SK_{sign_User}= Cert_User + $SK_{User} \cdot P_{User}$.

To deal with the certificate revocation problem, the time period TP_i is added into Cert_User to avoid the use of the current certification status.

PCS_Sign.

To sign message *m* with *Sign* algorithm, signer A in Domain_DM_i selects a random *r* and outputs a signature σ = (U, V), where U= $r \cdot P_A$, h= H_2 (m, U), V=$(r+h) \cdot SK_{sign\_A}$. Signer A then sends σ, A_{INFO} to verifier.

PCS_Verify.

Verifier B in Domain_DM_j uses following algorithm to verify σ.

If B is DM then

B requests PKC_CA_i from repository;

B verifies PKC_CA_i with PK_{CAj};

B verifies PKC_DM_i in A_{INFO} with PK_{CAi} in PKC_CA_i;

If B is User then

B asks its DM to verify PKC_DM_i in A_{INFO};

B picks PK_{DMi} and $PARA_{DMi}$ from PKC_DM_i;

With parameters in $PARA_{DMi}$, B checks whether $\hat{e}(PK_{DMi} + PK_A, U + h \cdot P_A) == \hat{e}(P, V)$, where $h = H_2(m, U)$, if the equation holds, outputs 'Valid', otherwise outputs 'Invalid'.

4. The proposed scheme

We now present a key establishment and mutual authentication scheme based on the concatenated architecture and PCS. We further integrate mutual authentication into the mobility management procedure to improve authentication and handover efficiency.

We consider the scenario in Fig.2 as roaming scenario. Before MN starts roaming, each entity should run *PCS.Setup* to configure the relative parameters. Afterwards, MN leaves the home domain and accesses the AR1 of MAP domain1, then handovers from AR1 to AR2 within the same MAP domain. Finally, MN roams to MAP domain2.

4.1. Key establishment scheme (KES)

In order to secure the communications during authentication procedure, a key establishment scheme (KES) is necessary to build security channel between

MN and MAP or AR. As such, we propose a novel KES, in this section, which can be integrated into the later proposed mutual authentication protocol.

To establish a common shared key, two messages need to be exchanged between MN and MAP as shown in Fig.3. MN first sends a message to MAP that includes MN_{INFO} (message K1 in Fig.3). Upon receiving this message, MAP picks $PARA_{HA}$ from PKC_HA in MN_{INFO} and selects a time period TP_j. Afterwards, MAP computes $P'_{MN}=H_1(TP_j, MN_{INFO})$, as well as $PK'_{MAP}=SK_{MAP}\cdot P$ using the parameters in $PARA_{HA}$ and sends TP_j, PK'_{MAP}, MAP_{INFO} back to MN (message K2 in Fig.3). Upon receiving this message, MN picks $PARA_{MAP}$ from PKC_MAP in MAP_{INFO} and checks whether $\hat{e}(PK'_{MAP}, P') == \hat{e}(PK_{MAP}, P)$ holds to verify the validity of PK'_{MAP}. If the validity verification is successful, MN computes $P'_{MN}=H_1(TP_j, MN_{INFO})$.

(K1) MN→MAP: $[MN_{INFO}]$,
(K2) MAP→MN: $[TP_j, PK'_{MAP}, MAP_{INFO}]$

Fig.3. Key establishment scheme

With all these parameters, MN computes $K_{MN\text{-}MAP} = \hat{e}(SK_{MN}\cdot P'_{MN}, PK'_{MAP})$, MAP computes $K_{MAP\text{-}MN} = \hat{e}(SK_{MAP}\cdot P'_{MN}, PK_{MN})$. It can be easily proved that $K_{MN\text{-}MAP} = \hat{e}(SK_{MN}\cdot P'_{MN}, PK'_{MAP}) = \hat{e}(SK_{MN}\cdot P'_{MN}, SK_{MAP}\cdot P) = \hat{e}(SK_{MN}\cdot P, SK_{MAP}\cdot P'_{MN}) = \hat{e}(SK_{MAP}\cdot P'_{MN}, PK_{MN}) = K_{MAP\text{-}MN}$.

It should be noted that, for the security and convenience in the exchange of the time period TP_j, DM can use the time period TP_i chosen during *PCS.Setup*, instead of TP_j.

4.2. Mutual authentication protocol with KES (PCS-K-HMIPv6)

We incorporate the previous KES into our proposed mutual authentication protocol (PCS-K-HMIPv6), and presents the details of inter-domain as well as intra-domain authentication procedures in the following subsections.

4.2.1. Inter-domain authentication of PCS-K-HMIPv6

In our roaming scenario, inter-domain authentication occurs when MN first enters MAP domain1 and accesses AR1. Fig.4 shows the messages that are exchanged as part of the authentication procedure of PCS-K-HMIPv6.

(C1) AR1→MN: [RA, $AR1_{INFO}$]
(C2) MN→AR1: [LBU, RBU, CVR(PKC_MAP), MN_{INFO}, TS_1, σ_1, σ_2]
(C3) AR1⇨MAP: [LBU, CVR(PKC_HA), RBU, CVR(PKC_MAP), MN_{INFO}, TS_1, σ_2]
(C4a) MAP→ Repository: [CVR(PKC_CA1)]
(C4b) MAP →HA: [RBU, CVR(PKC_MAP), TS_1, σ_2]
(C5) HA→ Repository: [CVR(PKC_CA2)]
(C6) Repository→ HA: [CVA(PKC_CA2)]
(C7a) HA→MAP: [RBA, CVA(PKC_MAP), TS_2, σ_3]
(C7b) Repository→ MAP: [CVA(PKC_CA1)]
(C8) MAP ⇨ AR1: [LBA, CVA(PKC_HA), RBA, CVA(PKC_MAP), TS_2, σ_3, TP_j,
PK'_{MAP}, SEK_{MN-MAP}].
(C9) AR1→MN: [LBA, RBA, CVA(PKC_MAP), TS_2, σ_3, TS_3, σ_4, TP_j, PK□'$_{MAP}$]

Fig.4. Inter-domain authentication of PCS-K-HMIPv6

AR1 periodically broadcasts a message (message C1 in Fig.4) to its coverage area through router advertisement (RA) which carries $AR1_{INFO}$. Upon receiving this message, MN starts the mobility registration procedure. In order to protect registration signaling, MN signs LBU with *PCS.Sign* and outputs σ_1={LBU, TS_1}$_{PCS\_Sign\_MN}$, where TS_1 is the current timestamp. MN also signs RBU by using HMAC [34] and outputs:

σ_2={RBU,CVR(PKC_MAP),TS_1}$_{HMAC\_Sign\_MN}$
=H_3(RBU,CVR(PKC_MAP),TS_1,K_{MN-HA})

where CVR (certificate verification request) is a new message introduced by PCS-K-HMIPv6, to request a valid PKC from DM. Without the ability of verifying PKC_MAP in $AR1_{INFO}$, MN should send the CVR to its DM (HA) to request a valid PKC_MAP. MN combines registration signaling (together with signature and timestamp), CVR and MN_{INFO} into one message (message C2 in Fig.4) and sends it to AR1. AR1 checks the freshness of TS_1 to protect against replay attacks and forwards the message (message C3 in Fig.4) to MAP through a secure channel. As AR1 is not DM, this message also includes a CVR to MAP to verify the PKC_HA. After receiving this message, MAP requests the PKC_CA1 (message C4a in Fig.4) from the repository in order to verify PKC_HA. In the meantime, MAP forwards RBU, CVR to HA (message C4b in Fig.4). Upon receiving this message, HA executes the following steps:

(1) It verifies σ_2 with HMAC, {σ_2}$_{HMAC\_Verify\_HA}$. If the signature is verified, HA updates its binding cache.

(2) It requests PKC_CA2 (message C5 in Fig.4) from the repository the public key (PK_{CA2}) in order to verify PKC_MAP. The repository then returns HA PKC_CA2 (message C6 in Fig.4) through a certificate verification acknowledgement message (CVA) which is the response to the CVR.

(3) It verifies PKC_CA2 with PK_{CA1}, and then verifies PKC_MAP with PK_{CA2}.

(4) It returns RBA, CVA (message C7a in Fig.4) to MN together with the HMAC signature, where $\sigma_3 = H_3$ (RBA, CVA, TS_2, K_{MN-HA}).

As a reply to message C4a, the repository sends PKC_CA1 to MAP (message C7b in Fig.4). In order to establish a common key between MN and MAP, upon receiving message C7a from HA, MAP executes the following steps:

(1) It verifies PKC_CA1 with PK_{CA2}, and then verifies PKC_HA with PK_{CA1}.

(2) It executes protocol KES using $PARA_{HA}$ in PKC_HA in order to generate K_{MAP-MN}.

(3) It computes the session key $SEK_{MN-MAP} = H_3(TS_1, H_4(K_{MAP-MN}))$.

(4) It records the relationship of TP_j, MN_{INFO} and K_{MAP-MN}.

(5) It inserts LBA, CVA, TP_j, PK'_{MAP} and SEK_{MN-MAP} into a message (message C8 in Fig.4), and then sends this message to AR1 through a secure channel.

Upon receiving such message, AR1 executes the following steps:

(1) It signs LBA with HMAC instead of *PCS.Sign* using SEK_{MN-MAP} in (C8) and outputs $\sigma_4 = H_3$ (LBA, TS_3, SEK_{MN-MAP}).

(2) It sends a message (message C9 in Fig.4) to MN that includes σ_4 and other information from message C8.

(3) It uses a valid PK_{HA} and $PARA_{HA}$ in PKC_HA to verify σ_1 with *PCS.Verify,* $\{\sigma_1\}_{PCS\_Verify\_AR1}$.

After receiving the message from AR1, MN first checks the freshness of TS_3. It then executes KES using TP_j, PK'_{MAP} in (C9) to generate K_{MN-MAP}. MN computes the session key $SEK_{MN-MAP} = H_3(TS_1, H_4(K_{MN-MAP}))$ and uses this key to verify σ_4 with HMAC, $\{\sigma_4\}_{HMAC\_Verify\_MN}$. If the verification is successful, the mutual authentication between MN and AR1 is completed.

It should be noted that the implementation of timestamp is a critical factor. We suggest using 'Mobility Message Replay Protection Option' in [25] to carry timestamp and utilize NTP [26] for time synchronization among the participants.

4.2.2. Intra-domain authentication of PCS-K-HMIPv6

Fig.5 shows the messages that are exchanged as part of the intra-domain authentication process when MN moves from AR1 to AR2 within the same MAP domain.

When accessing AR2, MN receives a message (message W1 in Fig.5) from AR2 which carries $AR2_{INFO}$. For the sake of intra-domain handover, only the LBU should be sent to MAP according to HMIPv6. MN signs the LBU with *PCS.Sign* and outputs $\sigma_5 = \{LBU, TS_4\}_{PCS\_Sign\_MN}$. MN sends a message (message W2 in Fig.5) to AR2 that includes the LBU, the current timestamp (TS_4), MN_{INFO}, σ_5. AR2 first checks the freshness of TS_4 to protect from replay attacks; then it sends a CVR to MAP to request valid PKC_HA (message W3 in Fig.5). To achieve an efficient KES with MN, upon receiving message W3 from AR2, MAP checks the freshness of time period TP_j which was recorded during inter-domain authentication. If the time period is fresh,

MAP computes the new session key $SEK'_{MN\text{-}MAP} = H_3(TS_4, H_4(K_{MAP\text{-}MN}))$. Otherwise, MAP must re-execute a KES protocol with MN. MAP then sends a message to AR2 together with $SEK'_{MN\text{-}MAP}$ through a secure channel (message W4 in Fig.5). AR2 signs LBA with HMAC using $SEK'_{MN\text{-}MAP}$ and outputs $\sigma_6 = H_3(LBA, TS_5, SEK'_{MN\text{-}MAP})$. AR2 sends a message (message W5 in Fig.5) to MN that includes LBA, σ_6, and the current timestamp (TS_5). After receiving this message, MN first checks the freshness of TS_5 and also computes $SEK'_{MN\text{-}MAP} = H_3(TS_4, H_4(K_{MAP\text{-}MN}))$. Then MN verifies σ_6 with HMAC using $SEK'_{MN\text{-}MAP}$, $\{\sigma_6\}_{HMAC\_Verify\_MN}$. If the verification is successful, the mutual authentication between MN and AR2 is completed.

(W1) AR2→MN: [RA, $AR2_{INFO}$]
(W2) MN→AR2: [LBU, TS_4, MN_{INFO}, σ_5]
(W3) AR2 ⇨ MAP: [LBU, CVR(PKC_HA), MN_{INFO}]
(W4) MAP ⇨ AR2: [LBA, CVA(PKC_HA), $SEK'_{MN\text{-}MAP}$]
(W5) AR2→MN: [LBA, TS_5, σ_6]

Fig.5. Intra-domain authentication of PCS-K-HMIPv6

When MN roams to MAP domain2, the same operations are executed as the ones executed in the inter-domain authentication. It should be noted that, after the mutual authentication, MN and MAP/AR can set up secure channel for their subsequent communications using the shared SEK generated as part of the PCS-K-HMIPv6 protocol.

4.3. Compatibility of the scheme

Recently another novel local mobility management protocol, proxy mobile IPv6 (PMIPv6 [36]), is proposed by IETF and receives comprehensive attentions in research community. PMIPv6 is intended for providing network-based mobility management support to a MN without requiring MN's participation in any IP mobility-related signaling. Two functional entities are introduced in PMIPv6: local mobility anchor (LMA) and mobile access gateway (MAG). LMA is the home agent for the MN in the home network. MAG, located at the visiting network, is responsible for managing the mobility-related signaling by the deputy of the MN that is attached to its managed ARs. In spite of the increasing focus on the efficiency and deployment issues of PMIPv6, few security concerns have been conducted [37].

Fortunately, our proposed concatenated security architecture and mutual authentication protocol can be well adapted to PMIPv6 to address the security problem. Similiar as HA and MAP, LMA and MAG may also act as domain managers in our security architecture. They are in charge of issuing certificates to the managed MNs and ARs respectively through *PCS*. MN and

accessing AR are thus able to generate the correspoding signing keys and further perform the mutual authentication as well as key establishment operations according to the scheme described in section 4.1 and 4.2. However, some revisions are still necessary for the compatibility to PMIPv6 since both topology and signalings are quite different between PMIPv6 and HMIPv6, which will be left for the further research work.

5. Performance analysis

We evaluate the authenticated handover latency of MN for the following protocols: PKI-HMIPv6 [6], 2-IBS-HMIPv6 [10], and PCS-K-HMIPv6. The authenticated handover latency refers to the interval from the time when MN enters a new MAP domain or different ARs in the same MAP domain to the time when the mutual authentication and mobility registration are completed.

5.1. Analytical model

From the definition of authenticated handover latency (T_{ah}) we can see that the latency is incurred during the mutual authentication and mobility management procedure. T_{ah} consists of transport latency (T_t), authentication cost (T_c), and node processing time (T_p).

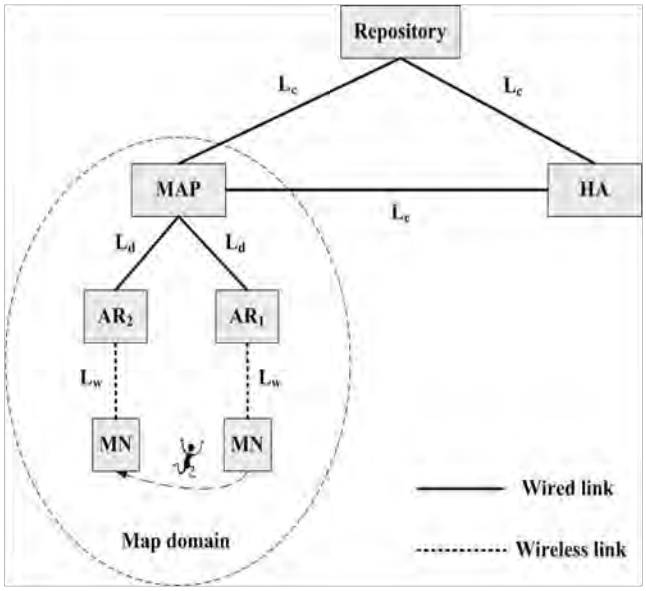


Fig.6. System model for transport latency analysis

$$T_{ah}=T_t+ T_c +T_p \quad (1)$$

We adopt the system model shown in Fig.6 to analyze T_t first. The transport latency can be categorized into three types: wireless link latency (L_w), intra-domain wired link latency (L_d), and inter-domain wired link latency (L_c). In most cases we have that $L_c > L_w > L_d$. L_w and L_d are fixed when the link type is determined. L_c is a variant with respect to the changeable distance between two administrative domains. We can treat L_c as multi-hop of $L_{d:}$

$$L_c=h\times L_d+(h-1) \times T_p \quad (2)$$

where h is the number of hops between two administrative domains, and T_p is the processing time of intermediate routers which is also fixed as long as the node type is determined. Consequently, we have that:

$$T_t = L_w+(h+1)\times L_d+(h-1) \times T_p \quad (3)$$

T_c is another variable which is primarily determined by the adopted authentication algorithm. Without loss of generality, we assume the classic RSA signature [27] is adopted for the verification of PKCs in PKI-HMIPv6 and PCS-K-HMIPv6. Compared with that, the computational cost of identity-based or certificate-based signature schemes in 2-IBS-HMIPv6 and PCS-K-HMIPv6 is higher. The involved operations consist of scale multiplication (SM), point addition (PA), bilinear pairing (BP), multiplication in group (MG), map to point function (MTP), and hash function (Hash).

We report the cost analysis of these operations in Tab.2. Let t_x denotes the computational cost of operation x. According to [28,29], t_{PA}, t_{MG}, t_{Hash} and t_{RSAv} are negligible compared with t_{BP} ,t_{MTP}, t_{SM} and t_{RSAs}. Note that t_{RSAs} and t_{RSAv} denote the computational cost of RSA sign and RSA verification, respectively.

Tab.2. Computational cost of the operations in the different schemes

	SM	PA	BP	MG	MTP	Hash
$2\text{-}IBS_{1\text{-}s}$	1	1	N/A	N/A	N/A	1
$2\text{-}IBS_{1\text{-}v}$	N/A	N/A	2	1	N/A	1
$2\text{-}IBS_{2\text{-}s}$	1	1	N/A	N/A	N/A	1
$2\text{-}IBS_{2\text{-}v}$	N/A	N/A	3	2	N/A	1
PCS_s	2	N/A	N/A	N/A	N/A	1
PCS_v	1	2	2	N/A	N/A	1
KA_MAP	2	N/A	1	N/A	1	N/A
KA_MN	1	N/A	2	N/A	1	N/A

Note that:

$2\text{-}IBS_{1\text{-}s/v}$: It denotes the signature and verification algorithm used by first tier PKG in 2-IBS-HMIPv6;

$2\text{-}IBS_{2\text{-}s/v}$: It denotes the signature and verification algorithm used by second tier users in 2-IBS-HMIPv6;

$PCS_{s/v}$: It denotes the signature and verification algorithm in PCS;

KA_MAP: It denotes the key agreement operations at the MAP side;

KA_MN: It denotes the key agreement operations at the MN side.

From expressions (1), (2), (3) we can conclude that:

$$T_{ah}=aL_w+bL_d+cL_c+T_p+T_c=aL_w+(b+c\times h)L_d+(c\times h-c+1)T_p+T_c \quad (4)$$

where a, b, c are the number of messages in each type of link. We define three types of authenticated handover latency: inter-domain authenticated handover latency, intra-domain authenticated handover latency and total authenticated handover latency. Each of these is evaluated in the following sections.

5.2. Inter-domain authenticated handover latency analysis

The inter-domain authenticated handover latency (T_{ah_IRD}) refers to the interval from the time MN receives the first RA in the access MAP domain to the end time of the remote mobility registration.

In PKI-HMIPv6, mutual authentication and mobility registration are executed separately. Both remote and local registration will occur after the successful mutual authentication, and the negotiation of security association between MN and AR is mandated to set up IPSec channel for mobility registration messages. T_{ah_IRD} of PKI-HMIPv6 can be evaluated as follows:

$$\begin{aligned} T_{ah\_IRD}\left(PKI-HMIPv6\right) &= 5L_w+4L_d+4L_c+14T_p+t_{RSAs}+3t_{RSAv} \\ &= 5L_w+\left(4h+4\right)L_d+\left(4h+10\right)T_p+t_{RSAs} \end{aligned} \quad (5)$$

In 2-IBS-HMIPv6, mutual authentication is integrated into the mobility registration procedure. A round trip message delivery between MN and HA is thus required to achieve both authentication and registration. Therefore we can evaluate T_{ah_IRD} of 2-IBS-HMIPv6 as follows:

$$\begin{aligned} T_{ah\_IRD}\left(2-IBS-HMIPv6\right) &= 2L_w+2L_d+2L_c+7T_p+t_{2-IBS1-v}+2t_{2-IBS2-s}+2t_{2-IBS2-v} \\ &= 2L_w+2t_{SM}+\left(2h+2\right)L_d+(2h+5)T_p+8t_{BP} \end{aligned} \quad (6)$$

PCS-K-HMIPv6 also incorporates mutual authentication with mobility registration procedure and there are additional queries of PKC between the domain managers (HA, MAP) and the repository. In addition, PCS-K-HMIPv6 has a key establishment between MN and MAP. We can evaluate T_{ah_IRD} of PCS-K-HMIPv6 as below:

$$\begin{aligned} &T_{ah\_IRD}\left(PCS-K-HMIPv6\right)=2L_w+2L_d+4L_c+9T_p+t_{PCSs}+t_{PCSv}+ \\ &t_{KA\_MAP}+t_{KA\_MN}=2L_w+\left(4h+2\right)L_d+\left(4h+5\right)T_p+5t_{BP}+6t_{SM}+2t_{MTP} \end{aligned} \quad (7)$$

5.3. Intra-domain authenticated handover latency analysis

The intra-domain authenticated handover latency (T_{ah_IAD}) refers to the interval between the time of the MN handover to another AR within the same MAP domain and the end time of the local mobility registration.

In terms of local handover, only the local mobility registration should be undertaken and no PKC verification and key establishment are needed since

these have been executed during the inter-domain handover. Authenticated handover latencies of the schemes are given by expressions (8), (9), and(10) .

$$T_{ah\_IAD}\left(PKI-HMIPv6\right)=5L_w+4L_d+10T_p+t_{RSAs} \tag{8}$$

$$\begin{aligned} T_{ah\_IAD}\left(2-IBS-HMIPv6\right) &= 2L_w+2L_d+5T_p+2t_{2-IBS2-s}+2t_{2-IBS2-v} \\ &= 2L_w+2L_d+5T_p+6t_{BP}+2t_{SM} \end{aligned} \tag{9}$$

$$\begin{aligned} T_{ah\_IAD}\left(PCS-K-HMIPv6\right) &= 2L_w+2L_d+5T_p+t_{PCSs}+t_{PCSv} \\ &= 2L_w+2L_d+5T_p+2t_{BP}+3t_{SM} \end{aligned} \tag{10}$$

5.4. Total authenticated handover latency analysis

HMIPv6 is designed for a scenario where MN handovers frequently within a domain far away from its home domain. Accordingly the total authenticated handover latency (T_{ah_TOT}), which is the sum of T_{ah_IRD} and all T_{ah_IAD}, must be taken into consideration. This sum is computed as:

$$T_{ah\_TOT}= T_{ah\_IRD} +\rho\ T_{ah\_IAD} \tag{11}$$

where ρ is the handover frequency of MN within the MAP domain.

Based on expressions (5)-(11), we have:

$$\begin{aligned} T_{ah\_TOT}\left(PKI-HMIPv6\right) &= (5\rho+5)L_w+(4\rho+4h+4)L_d \\ &\quad +(10\rho+4h+10)T_p+(\rho+1)t_{RSAs} \end{aligned} \tag{12}$$

$$\begin{aligned} T_{ah\_TOT}\left(2-IBS-HMIPv6\right) &= (2\rho+2)L_w+(2\rho+2h+2)L_d \\ &\quad +(5\rho+2h+5)T_p+(6\rho+8)t_{BP}+(2\rho+2)t_{SM} \end{aligned} \tag{13}$$

$$\begin{aligned} T_{ah\_TOT}\left(PCS-K-HMIPv6\right) &= (2\rho+2)L_w+(2\rho+4h+2)L_d \\ &\quad +(5\rho+4h+5)T_p+(2\rho+5)t_{BP}+(3\rho+6)t_{SM}+2t_{MTP} \end{aligned} \tag{14}$$

5.5. Numerical results and discussions

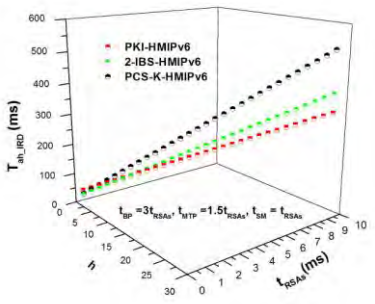
This section presents the performance differences of the above schemes through numerical results and discussions.

Based on the comprehensive analysis of the experimental results in [29-33], t_{RSAv} can be omitted as it is negligible compared with t_{RSAs}. We also get following conclusions:

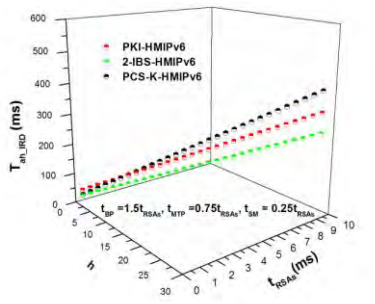
$$t_{BP}=1.5\sim3\ t_{RSAs},\ t_{MTP}=0.75\sim1.5\ t_{RSAs},\ t_{SM}=0.25\sim1\ t_{RSAs} \tag{15}$$

In order to analyze the performance differences, we select two groups of performance parameters: { $t_{BP}=3\ t_{RSAs}$, $t_{MTP}=1.5\ t_{RSAs}$, $t_{SM}=1\ t_{RSAs}$ } and { $t_{BP}=1.5t_{RSAs}$, $t_{MTP}=0.75t_{RSAs}$, $t_{SM}=0.25\ t_{RSAs}$ } for our analysis, where the two groups indicate the worst and best performance of the authentication operations respectively under the constrains of expression (15). Moreover,

we set L_w=4ms， L_d=2ms， T_p=0.5ms, which we called fixed parameters according to [10].



(a) T_{ah_IRD} vs. h and t_{RSAs} with $\{t_{BP}=3t_{RSAs}, t_{MTP}=1.5t_{RSAs}, t_{SM}=1\ t_{RSAs}\}$



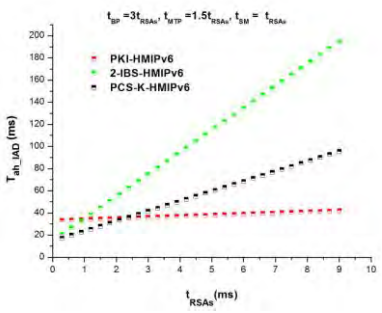
(b) T_{ah_IRD} vs. h and t_{RSAs} with $\{t_{BP}=1.5t_{RSAs}, t_{MTP}=0.75t_{RSAs}, t_{SM}=0.25t_{RSAs}\}$

Fig.7. Numerical results for inter-domain authenticated handover latency

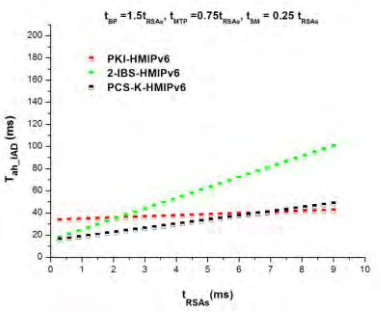
Fig.7-9 plot the results of T_{ah_IRD}, T_{ah_IAD}, and T_{ah_TOT} for each scheme in light of expressions (5)-(14) based on different groups of performance parameters.

As shown in Fig.7, although the authentication and mobility registration are separated，PKI-HMIPv6 only requires few RSA signatures and verifications to achieve mutual authentication. Therefore T_{ah_IRD} of PKI-HMIPv6 is lower than the other schemes which involve more expensive authentication operations such as BP, MTP or SM as shown in Fig.7 (a). PCS-K-HMIPv6 has the highest T_{ah_IRD} since it requires not only authentication operations but also KES operations during inter-domain handovers. However, with the performance enhancement for the authentication operations (t_{BP} *=1.5*t_{RSAs}, t_{MTP} *= 0.75*t_{RSAs}, t_{SM} *= 0.25* t_{RSAs}) (see Fig.7 (b)), the T_{ah_IRD} of each scheme drops obviously except for PKI-HMIPv6.

As there are no interactions among the MAP domain, the home domain, and the repository during the intra-domain handover, the parameter *h* has no impact. T_{ah_IAD} mainly depends on the performance of the authentication operations. As a consequence, 2-IBS-HMIPv6 has the highest T_{ah_IAD} among the three schemes because of more heavy BP computations. Our PCS algorithm mitigates such heavy operations in both signature and verification processes compared with the scheme in [10]. Thus T_{ah_IAD} of PCS-K-HMIPv6 is lower than 2-IBS-HMIPv6. As shown in Fig.8 (b), with the performance enhancement to the authentication operations (t_{BP} *=1.5*t_{RSAs}, t_{MTP} *= 0.75*t_{RSAs}, t_{SM} *= 0.25* t_{RSAs}), T_{ah_IAD} of PCS-K-HMIPv6 is even lower than PKI-HMIPv6 when t_{RSAs}< 6.6ms.



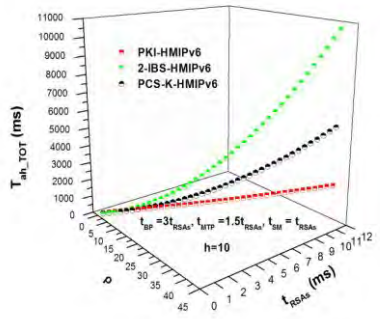
(a) T_{ah_IAD} vs. t_{RSAs} with { t_{BP} =3t_{RSAs}, t_{MTP} = 1.5 t_{RSAs}, t_{SM} = 1 t_{RSAs} }



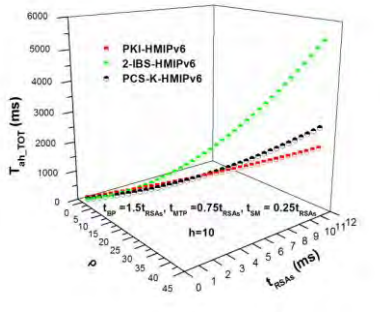
(b) T_{ah_IRD} vs. t_{RSAs} with {t_{BP} =1.5t_{RSAs}, t_{MTP}=0.75 t_{RSAs}, t_{SM}=0.25 t_{RSAs} }

Fig.8. Numerical results for intra-domain authenticated handover latency

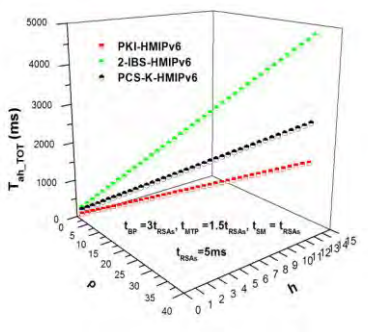
T_{ah_TOT} is important as it reflects the overall performance of each scheme. We first set h=10 to observe how T_{ah_TOT} is affected by ρ and tRSAs. From Fig.9 (a) and (b), we can see that 2-IBS-HMIPv6 performs worst. The reason is that 2-IBS-HMIPv6 requires more expensive authentication operations during both inter-domain and intra-domain handovers. In contrast, although PCS-K-HMIPv6 requires similar authentication and KES operations during inter-domain handover, these operations are eliminated or their costs are greatly mitigated in terms of intra-domain handovers. As shown in Fig.9 (b), T_{ah_TOT} of PCS-K-HMIPv6 is lower than PKI-HMIPv6 when t_{RSAs}<6.3ms. On the other hand, we set t_{RSAs}=5ms to see how T_{ah_TOT} is affected by ρ and h. A similar result is obtained. As shown in Fig.9 (d), T_{ah_TOT} of PCS-K-HMIPv6 is lower than the other two schemes when ρ>6.



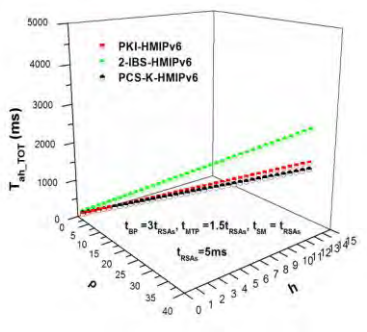
(a) T_{ah_TOT} vs. ρ and t_{RSAs} with { t_{BP} =3t_{RSAs}, t_{MTP} = 1.5 t_{RSAs}, t_{SM} = 1 t_{RSAs} }



(b) Tah_TOT vs. ρ and tRSAs with { tBP =1.5tRSAs, tMTP = 0.75 tRSAs, tSM = 0.25 tRSAs }



(c) T_{ah_TOT} vs. ρ and h with { t_{BP} =3t_{RSAs}, t_{MTP} = 1.5 t_{RSAs}, t_{SM} =1 t_{RSAs} }



(d) T_{ah_TOT} vs. ρ and h with { t_{BP} =1.5t_{RSAs}, t_{MTP} =0.75 t_{RSAs}, t_{SM} =0.25t_{RSAs} }

Fig.9. Numerical results for total authenticated handover latency

To summarize, PCS-K-HMIPv6 has better overall performance when the MN frequently handovers (higher ρ) in remote MAP domains with efficient authentication operations (lower t_{RSAs}).

6. Security analysis

In this section, we analyze the security of our proposed scheme with respect to key, signature, as well as mobility registration procedure**.**

6.1. Key security

There are three types of key in our proposed schemes: long-term shared key or self-generated key, mid-term singing key or agreed key, as well as short-term session key. Security of all these keys is critical.

(1) We assume that the long-term shared keys (e.g. $K_{MN\text{-}HA}$, $K_{AR\text{-}MAP}$) are pre-shared between two parties and the long-term self-generated keys (e.g. SK_{DM}, SK_{User}) are securely kept by their owners.

(2) User's mid-term signing key is generated as SK_{sign_User}=Cert_User+SK_{User} · P_{User}, where Cert_User is openly issued by DM. However, SK_{User} is randomly generated and securely kept by User. Hence no one but User can generate SK_{sign_User}. In addition, our scheme does not have the key escrow problem of IBC since DM cannot create SK_{sign_User} either. The mid-term agreed key (e.g. $K_{MN\text{-}MAP}$) is produced through the KA scheme. Although some information (PK'_{MAP}) will be exchanged openly between participants, the adversary has no means for getting SK_{MN} or SK_{MAP}, so it is unable to compute $K_{MN\text{-}MAP}$ by $\hat{e}$ (SK_{MN}·P'_{MN}, PK'_{MAP}) or $\hat{e}$ (SK_{MAP}·P'_{MN},

PK_{MN}). Moreover, in order to avoid malicious modifications, PK'_{MAP} is also verified by MN by checking whether $\hat{e}(PK'_{MAP}, P') == \hat{e}(PK_{MAP}, P)$.

(3) The short-term session key (e.g. $SEK_{MN\text{-}MAP}$) is derived from the agreed key ($K_{MN\text{-}MAP}$) and a valid timestamp by $SEK_{MN\text{-}MAP}=H_3(TS_1, H_4(K_{MN\text{-}MAP}))$. The security of $K_{MN\text{-}MAP}$ ensures that only MN and MAP can create this session key and the timestamp guarantees the freshness of the session key when MN handovers within the MAP domain.

6.2. Signature security

Our proposed scheme provides secure mutual authentication between MN and the MAP domain being visited based on PCS. Consider the following impersonation and modification attack scenarios:

(1) The adversary forges a valid signature to impersonate as legitimate User. As *PCS.sign* and *PCS.verify* are based on CBS which has been proved to be secure [23] under the condition of CDHP (computational Diffie-Hellman problem) difficulty in random oracle model [35], the only way by which an adversary can forge the signature is via stealing the signing key of legitimate User. However, as we discussed in section 6.1, User's signing key is secure to against such attack.

(2) The adversary collects a used signature to launch a replay attack. In our mutual authentication scheme, all the signatures are equipped with timestamps. Hence replay attacks can be easily detected by verifying the freshness of timestamps.

(3) The adversary modifies the public parameters so as to compromise the verification procedure. According to *PCS.verify,* the verifier must possess some public parameters, such as PK_{DM} and $PARA_{DM}$, in order to verify a signature. If these parameters are modified by the adversary, the verification will fail. To prevent this attack, we store the public parameters in DM's PKC (PKC_DM). The verifier should first get a valid PKC_DM from the repository, and then pick up the right parameters from PKC_DM to properly verify the signature.

6.3. Mobility registration security

Our mutual authentication protocol can provide protection for registration messages. As HMIPv6 has a local registration (LBU/LBA) and a remote registration (RBU/RBA), MN and AR sign LBU and LBA with PCS respectively during the mutual authentication procedure, which guarantees the security of the local registration. In order to protect the remote registration, MN signs the RBU with $K_{MN\text{-}HA}$ using HMAC. After receiving the RBU, HA verifies the signature with the same shared key. In addition, a timestamp is used to prevent replay attacks aiming at the RBU. The same

operations are carried out by HA on RBA messages. Hence the whole remote registration is secure.

7. Related work

PKI-based security schemes: PKI can be used to prevent different kinds of attacks and is suitable for large scale, hierarchical networks. To deploy PKI in HMIPv6 networks, Mizuno et al. [6] proposed a novel PKI-based security architecture. Mutual authentication is supported through IKE and cross-certificates [24] between mobile nodes and the mobile anchor point (MAP). The approach suffers from the problems that IKE has in dynamic mobile networks. In addition the MAP becomes a bottleneck of the system since it should handle authentications for all the accessing mobile nodes. The certificate-based binding update protocol [7] is another PKI-based solution for HMIPv6 networks which provides the functions of secure mobility registration, user authentication, and session key management. However, the goal of this scheme is to protect the communications between the mobile nodes and correspondent nodes[1]. Such scheme does not address the security issues that arise when mobile nodes move to different networks. Although PKI has certain advantages for large scale and explicit authentication, the complicated public key management as well as verification cost of PKC limits the applicability of these PKI-based schemes.

CGA-based and IBC-based security schemes: Cryptographically generated addresses (CGA [38]) is a security technique whereby the interface of IP address is generated by hashing a public key and some other parameters associated with node while not allocated by PKI. As such, [39] is a security extension to HMIPv6 based on CGA, which allows the MN to establish a security association with the selected MAP for authentication and other security operations. However CGAs themselves are not certified by any trusted authority, then the association between public key and MN cannot be verified. Therefore, a malicious node is able to generate its own public – private key pair and enter the visiting network as a free rider. In addition, the special construction of CGA renders it cannot be used in other address assignment mechanisms. Besides, several schemes [8-11] introduced IBC into HMIPv6 networks. Zhu et al. [8] developed an IBC-based security architecture to achieve authentication and non-interactive key establishment between access routers and mobile nodes. However such scheme concentrates on the security of wireless mesh networks. Kandikattu and Jacob [9] designed a secure framework with F-HMIPv6 [12] and a novel mobility management scheme. Access authentication and secure route optimization are implemented under the proposed framework by means of

[1] Correspondent nodes, defined in MIPv6 protocol, are the nodes with which a mobile node is communicating. The correspondent nodes may be either mobile or stationary.

IBC. Tian et al. [10] proposed a hierarchical identity-based signature scheme for mutual authentication in HMIPv6 networks. In such scheme, the authentication and mobility management procedures are integrated in order to improve efficiency. Wu et al. [11] further took reputation issues into consideration. However, the special format of IP address suggested in [8] and the low authentication efficiency of [10] and [11] constrain the appeal of these IBC-based solutions. Moreover, IBC is only suitable for small area networks where trust relationships can be easily established.

PKI and IBC hybrid architectures: A hybrid scheme combining PKI and identity-based encryption (IBE) was proposed by Chen et al. [14]. They suggested that the combination of the two mechanisms, PKI-based keys for trust authorities and IBC-based keys for users, has many advantages including scalability. Later, Price and Mitchell [15] dwelt into interoperation issues between conventional PKI and IBE infrastructures. Recently, Lee [16] proposed a unified public key infrastructure combining PKI and IBC. A new authority KGCA dedicated to the role of both PKG and CA was proposed for issuing certificates and partial private keys to the users. However, KGCA is critical for performance as it has to perform all the tasks of the PKG and CA. In general, none of these hybrid schemes have been applied to HMIPv6 networks.

8. Conclusions

In this paper, we have proposed an approach that incorporates PKI and CBC in a hierarchical security architecture and a novel mutual authentication and key establishment scheme for HMIPv6 networks. The motivation for our work is that none of the hybrid schemes previously proposed satisfy the security requirements of such networks. The proposed concatenated architecture harnesses the merits of both PKI and CBC, while addressing their limitations. Our mutual authentication protocol is based on a designated signature scheme (PCS), which ensures inter-domain trust by cross-certificate and intra-domain trust by CBS. In addition, a key establishment scheme has been defined to set up secure channels after authentication. The authentication scheme is integrated into the mobility management procedure in order to improve performance.

For the future research work, we plan to do the further simulations and implementations on our mutual authentication protocol. Moreover, the proposed hierarchical architecture and hybrid approach are expected to be explored for PMIPv6 security.

Acknowledgements. This work was supported in part by the Natural Science Foundation of Liaoning Province under Grant No. 201202069 and the Fundamental Research Funds for the Central Universities under Grant No. N120417003 and Grant No. N120404010.

References

1. H. Soliman, C. Castelluccia, K. ElMalki, L. Bellier. Hierarchical Mobile IPv6 (HMIPv6) Mobility Management. RFC5380. (2008)
2. Johnson D, Perkins C. Mobility Support in IPv6. RFC3775. (2004)
3. Hyun-Sun Kang ,Chang-Seop Park. Authenticated Fast Handover Scheme in the Hierarchical Mobile IPv6. Information Security Applications, LNCS, 211-224. (2007)
4. Miyoung Kim, Youngsong Mun, Jaehoon Nah, Seungwon Sohn. An Authentication Scheme using AAA in Hierarchical MIPv6. draft-mun-mip6-authhmip-mobileipv6-00.txt. (2005)
5. C. Adams, S. Farrell, T. Kause, T. Mononen. Internet X.509 Public Key Infrastructure Certificate Management Protocol (CMP). RFC4210. (2005)
6. S.Mizuno, J.koga, H.Ohwada, K.Suzuki, Y.Takagi. PKI Support in Hierarchical Mobile IPv6. draft-mizuno-mobileip-hmipv6-pki-00.txt. (2003)
7. Feng Bao, Robert Deng, Ying Qiu, Jianying Zhou. Certificate-based Binding Update Protocol (CBU). draft-qiu-mip6-certificated-binding-update-03.txt. (2005)
8. Ramanarayana Kandikattu, Lillykutty Jacob. A Secure IPv6-based Urban Wireless Mesh Network (SUMNv6). Computer Communications, 31(15): 3707-3718. (2008)
9. Xiaoyan Zhu, Yuguang Fang and Yumin Wang. How to Secure Multi-domain Wireless Mesh Networks. Wireless Networks, 16(5): 1215-1222. (2010)
10. Ye Tian, Yujun Zhang, Hanwen Zhang, Zhongcheng Li. Identity-based hierarchical access authentication in mobile IPv6 network. Proceedings of ICC '06, 1953 – 1958. (2006)
11. Zhi Zhang, Guohua Cui. A Secure Hierarchical Identify Authentication Scheme Combining Trust Mechanism in Mobile IPv6 Networks. Journal of Networks, 4(5):343-350. (2009)
12. HeeYoung Jung, Hesham Soliman, Seok Joo Koh, Jae Yong Lee. Fast Handover for Hierarchical MIPv6 (F-HMIPv6). draft-jung-mobopts-fhmipv6-00.txt. (2005)

A. Shamir. Identity-based cryptosystems and signature schemes. In Advances in Cryptology - Crypto '84, Springer-Verlag LNCS 196, 1984:47-53. (1984)

13. L. Chen, Keith Harrison, Andrew Moss, David Soldera, Nigel P. Smart. Certification of Public Keys within an Identity Based System. Proceedings of Information Security Conference/Information Security Workshop - ISC(ISW), 322-333. (2002)
14. Geraint Price, Chris J. Mitchell. Interoperation Between a Conventional PKI and an ID-Based Infrastructure. Proceedings of European Public Key Infrastructure Workshop - EUROPKI, 73-85. (2005)
15. Byoungcheon Lee. Unified Public Key Infrastructure Supporting Both Certificate-Based and ID-Based Cryptography. Proceedings of Availability, Reliability and Security - IEEEARES, 54-61. (2010)
16. Boneh D, Franklin M. Identity-based encryption from the weil pairing . SIAM Journal of Computing, 32(3): 586-615. (2003)
17. Antoine Joux. The Weil and Tate Pairings as Building Blocks for Public Key Cryptosystems Survey. Proceedings of the 5th International Symposium on Algorithmic Number Theory (ANTS-V), LNCS 2369, 11-18. (2002)
18. Joseph H. Silverman. The Arithmetic of Elliptic Curves. Springer, ISBN 0387094938. (2009)
19. Dan Boneh, Ben Lynn, Hovav Shacham. Short Signatures from the Weil Pairing. Proceedings of ASIACRYPT - ASIACRYPT, 514-532. (2001)

20. Craig Gentry. Certificate-Based Encryption and the Certificate Revocation Problem. Proceedings of Theory and Application of Cryptographic Techniques - EUROCRYPT, 272-293. (2003)
21. Bo Gyeong Kang, Je Hong Park, Sang Geun Hahn. A Certificate-Based Signature Scheme. Proceedings of The Cryptographer's Track at RSA Conference - CT-RSA, 99-111. (2004)
22. Wei Wu, Yi Mu, Willy Susilo, Xinyi Huang. Certificate-based Signatures Revisited. Journal of Universal Computer Science, 15(8):1659-1684, (2009).
23. Jim Turnbull. Cross-certification and PKI policy networking. http://hca.nat.gov.tw/download/012.pdf. (2000)
A. Patel, K. Leung, M. Khalil, H. Akhtar, K. Chowdhury. Authentication Protocol for Mobile IPv6. RFC4285. (2006)
24. D. Mills, U. Delaware, J. Martin, Ed, J. Burbank, W. Kasch. Network Time Protocol Version 4: Protocol and Algorithms Specification. RFC5905. (2010)
25. Jean-françois Misarsky. How (not) to Design RSA Signature Schemes. Proceedings of Public Key Cryptography - PKC, 14-28. (1998)
26. Sandip Vijay, Subhash C. Sharma. Threshold Signature Cryptography Scheme in Wireless Ad-Hoc Computing. Contemporary Computing, 40(7):327-335. (2009)
27. Mohamed Abid, Songbo Song, Hassnaa Moustafa, Hossam Afifi. Integrating identity-based cryptography in IMS service authentication. International Journal of Network Security Its Applications, 1-13. (2010)
28. Paulo S. L. M. Barreto, Ben Lynn, Michael Scott. Efficient Implementation of Pairing-Based Cryptosystems. Journal of Cryptology, 17(4):321-334. (2004)
29. Paulo S. L. M. Barreto, Ben Ly International Cryptology Conference on Advances in Cryptology, LNCS 2442, 354-368. (2002)
30. Elisavet Konstantinou. Efficient cluster-based group key agreement protocols for wireless ad hoc networks. Journal of Network and Computer Applications, 34(1):384-393. (2011)
31. Xiong, X., Wong, D.S., Deng, X. TinyPairing: A Fast and Lightweight Pairing-Based Cryptographic Library for Wireless Sensor Networks. Proceedings of WCNC'2010, 1-6. (2010)
32. H. Krawczyk, M. Bellare, R. Canetti. HMAC: Keyed-Hashing for Message Authentication. RFC2104. (1997)
33. R. Dutta, R. Barua, P. Sarkar. Pairing-based cryptographic protocols: A survey. Cryptology, ePrint Archive, Report 2004/064. (2004)
34. S. Gundavelli, Ed. K. Leung, V. Devarapalli, Wichorus, K. Chowdhury, B. Patil. Proxy Mobile IPv6, RFC5213. (2008)
35. Joong-Hee Lee, Jong-Hyouk Lee, Tai-Myoung Chung. Ticket-based Authentication Mechanism for Proxy Mobile IPv6 Environment, Proceedings of the Third International Conference on Systems and Networks Communications, 304-309. (2008)
36. Aura T. Cryptographically Generated Addresses (CGA), RFC 3972. (2005).
37. Haddad W, Krishnan S, Soliman H. Using cryptographically generated addresses (CGA) to secure HMIPv6 protocol (HMIPv6sec), draft-haddad-mipshop-hmipv6-security-06.(2006)

Tianhan Gao Tianhan Gao received the BE in Computer Science & Technology, the ME and the PhD in Computer Application Technology, from Northeastern University, China, in 1999, 2001, 2006, respectively. He joined Northeastern University in April 2006 as a lecturer of Software College. He obtained an early promotion to an associate professor in January 2010. He has been a visiting scholar at department of Computer Science, Purdue, from February 2011 to February 2012. He is the author or co-author of more than 30 research publications. His primary research interests are next generation network security, MIPv6/HMIPv6 security, wireless mesh network security, Internet security, as well as security and privacy in ubiquitous computing.

Nan Guo Nan Guo received the BE in Computer Science & Technology, the ME and the PhD in Computer Application Technology, from Northeastern University, China, in 1999, 2001, 2005, respectively. She joined Northeastern University in September 2005. She has been an associate professor since 2008. She has been a visiting scholar at department of Computer Science, Purdue, from August 2010 to August 2011. She is the author or co-author of more than 20 research publications. Her primary research interests are security and privacy in service computing and digital identity management.

Kangbin Yim Kangbin Yim received his B.S., M.S., and Ph.D. from Ajou University, Suwon, Korea in 1992, 1994 and 2001, respectively. He is currently an associate professor in the Department of Information Security Engineering, Soonchunhyang University. He has served as an executive board member of Korea Institute of Information Security and Cryptology, Korean Society for Internet Information and The Institute of Electronics Engineers of Korea. He also has served as a committee chair of the international conferences and workshops and the guest editor of the journals such as JIT, MIS, JISIS and JoWUA. His research interests include vulnerability assessment, code obfuscation, malware analysis, leakage protection, secure hardware, and systems security. Related to these topics, he has worked on more than fifty research projects and published more than a hundred research papers.

Received: November 14, 2012; Accepted: April 03, 2013

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

004

COMPUTER Science and Information Systems : the International journal / Editor-in-Chief Mirjana Ivanović. – Vol. 10, No 2 (2013) - . – Novi Sad (Trg D. Obradovića 3): ComSIS Consortium, 2012 - (Belgrade : Sigra star). –30 cm

Polugodišnje. – Tekst na engleskom jeziku

ISSN 1820-0214 = Computer Science and Information Systems
COBISS.SR-ID 112261644

Cover design: V. Štavljanin
Printed by: Sigra star, Belgrade

ComSIS Computer Science and Information Systems
Volume 10, Number 2, Special Issue, April 2013

Contents

ISSN: 1820-0214