

Journal of Networks

ISSN 1796-2056

Volume 9, Number 4, April 2014

Contents

REGULAR PAPERS

- Virtual Content-Centric Networking 807
Keiichiro Tsukamoto, Masato Ohtani, Yuki Koizumi, Hiroyuki Ohsaki, Kunio Hato, and Junichi Murayama
- Adaptive Distributed Load Balancing Routing Mechanism for LEO Satellite IP Networks 816
Xiao Ma
- A Noise-Correlated Cancellation Transmission Scheme for Cooperative MIMO Ad Hoc Networks 822
Wanni Liu, Long Zhang, and Yanping Li
- CluLoR: Clustered Localized Routing for FiWi Networks 828
Yousef Dashti and Martin Reisslein
- Water Quality Monitoring and Control for Aquaculture Based on Wireless Sensor Networks 840
Daudi S. Simbeye and Shi Feng Yang
- An NFC-based Scenic Service System 850
Jie Ma and Jinlong E
- Research on Delay and Packet Loss Control Mechanism in Wireless Mesh Networks 859
Qiuling Yang, Zhigang Jin, and Xiangdang Huang
- Compressed Wideband Spectrum Sensing with Partially Known Occupancy Status by Weighted l_1 Minimization 866
Zha Song, Huang Jijun, and Li Ning
- An IP-Traceback-based Packet Filtering Scheme for Eliminating DDoS Attacks 874
Yulong Wang and Rui Sun
- Multi-Object Optimization Based RV Selection Algorithm for VCN 882
Rong Chai, Bin Yang, Li Cai, Xizhe Yang, and Qianbin Chen
- The Effect of MAC Parameters on Energy Efficiency and Delay in Wireless Sensor Networks 889
Zhihua Li, Bin Lian, Zhongcheng Wei, Liang Xue, and Jijun Zhao
- A Multicast Routing Algorithm for Datagram Service in Delta LEO Satellite Constellation Networks 896
Yanpeng Ma, Xiaofeng Wang, Jinshu Su, Chunqing Wu, Wanrong Yu, and Baokang Zhao
- Bandwidth Consumption Efficiency Using Collective Rejoin in Hierarchical Peer-To-Peer 908
Sri Wahjuni, A.A.Putri Ratna, and Kalamullah Ramli
- A Method of Case Retrieval for Web-based Remote Customization Platform 914
Yuhuai Wang, Hong Jia, and Xiaojing Zhu
- Symbol Timing Estimation with Multi-h CPM Signals 921
Sheng Zhong, Chun Yang, and Jian Zhang
-

Vector-Based Sensitive Information Protecting Scheme in Automatic Trust Negotiation <i>Jianyun Lei and Yanhong Li</i>	927
An Improved Byzantine Fault-tolerant Program for WSNs <i>Yi Tian</i>	932
EESA Algorithm in Wireless Sensor Networks <i>Zhang Pei and Feng Lu</i>	941
Routing Algorithm Based on Delay Rate in Wireless Cognitive Radio Network <i>Gan Yan, Yuxiang Lv, Qiyin Wang, and Yishuang Geng</i>	948
Energy Hole Solution Algorithm in Wireless Sensor Network <i>Lu Yuting and Wang Weiyang</i>	956
Identification Method of Attack Path Based on Immune Intrusion Detection <i>Huang Wenhua and Yishuang Geng</i>	964
Online Order Priority Evaluation Based on Hybrid Harmony Search Algorithm of Optimized Support Vector Machines <i>Zhao Yuanyuan and Chen Qian</i>	972
Framework and Modeling Method for Heterogeneous Systems Information Integration Base on Semantic Gateway <i>Xianwang Li, Yuchuan Song, Ping Yan, and Xuehai Chen</i>	979
Satellite Formation based on SDDF Method <i>Wang Yu, Wu Zhi-qiang, and Zhu Xin-hua</i>	986
Heterogeneous Web Data Extraction Algorithm Based On Modified Hidden Conditional Random Fields <i>Cheng Cui</i>	993
Nearly Optimal Solution for Restricted Euclidean Bottleneck Steiner Tree Problem <i>Zimao Li and Wenying Xiao</i>	1000
Computer Crime Forensics Based on Improved Decision Tree Algorithm <i>Ying Wang, Xinguang Peng, and Jing Bian</i>	1005
Demand-oriented Traffic Measuring Method for Network Security Situation Assessment <i>Xu Zhenhua</i>	1012
Visual Simulation of Explosion Effects Based on Mathematical Model and Particle System <i>Gong Lin and Hu Dingjun</i>	1020
Reliable Transmission Protocol based on Network Coding in Delay Tolerant Mobile Sensor Network <i>Luo Kan, Wang Hua, and Shyi-Ching Liang</i>	1027
Routing Optimization Based on Taboo Search Algorithm for Logistic Distribution <i>Yang Hongxue and Xuan Lingling</i>	1033
Opportunistic Cooperative Reliable Transmission Protocol for Wireless Sensor Networks <i>Hua Guo, Yu Sheng-Wen, and Douglas Leith</i>	1040
An Improved Channel Estimation Method based on Jointly Preprocessing of Time-frequency Domain in TD-LTE System <i>Yang Jianning, Lin Kun, and Zhao Xie</i>	1047

Dynamic Routing Algorithm Based on the Channel Quality Control for Farmland Sensor Networks <i>Dongfeng Xu</i>	1055
DCSK Multi-Access Scheme for UHF RFID System <i>Keqiang Yue, Lingling Sun, Bin You, and Shengzhou Zhang</i>	1061
An Effective Scheme for Performance Improvement of P2P Live Streaming Systems <i>Xiaosong Wu, Xingshu Chen, and Haizhou Wang</i>	1067

Virtual Content-Centric Networking

Keiichiro Tsukamoto^a, Masato Ohtani^a, Yuki Koizumi^a, Hiroyuki Ohsaki^b,
Kunio Hato^c, Junichi Murayama^d

^a Graduate School of Information Science and Technology Osaka University, Japan
Email: {k-tukamt,m-ohtani,ykoizumi}@ist.osaka-u.ac.jp

^b Department of Informatics, School of Science and Technology, Kwansai Gakuin University, Japan
Email: ohsaki@kwansai.ac.jp

^c NTT Secure Platform Laboratories NTT Corporation, Japan
Email: hato.kunio@lab.ntt.co.jp

^d Department of Communication and Network Engineering,
School of Information and Telecommunication Engineering, Tokai University, Japan
Email: murayama@m.ieice.org

Abstract—Data-centric networking has recently been gaining attention. A representative design for data-centric networking is Content-Centric Networking (CCN), which routes packets based on content identifiers. CCN is basically designed to be open because ease of data reuse is one of the greatest advantages of data-centric networking. However, for real-world networking, completely open data-centric networking is not sufficient; it is necessary to allow for private communication within a group of users. In this paper, we propose Virtual Content-Centric Networking (VCCN), which realizes private communication within a group of users through CCN router virtualization. We present four building blocks of VCCN: extension of the content identifier, CCN router virtualization, packet transport between virtualized CCN routers, and Social Network Services cooperative user/group identification. We have implemented VCCN's basic features by extending the CCNx software and have conducted a preliminary performance evaluation of our VCCN implementation.

Index Terms—Content-Centric Networking, Router Virtualization, Group-Based Communication, Social Network Services Cooperative User/Group Identification

I. INTRODUCTION

Data-centric networking, which takes named data rather than hosts as being connected via the network as its central abstraction, has recently been gaining attention [1]–[4].

A representative design for data-centric networking is Content-Centric Networking (CCN) [5], [6], in which routers forward packets based on unique content identifiers. CCN adopts a *request-and-response* communication model. A request packet from a user, called an *Interest packet*, is routed between CCN routers according to the longest prefix matching the requesting content identifier. If the Interest packet is successfully delivered to the source, the content packet, called a *Data packet*, is sent back to the user by traversing the path of the Interest packet in reverse.

CCN routers cache forwarded content in a buffer memory called the *contents store (CS)* for later reuse. When a CCN router receives an Interest packet for cached content, it returns the cached content as a Data packet so that the amount of traffic transferred over the network can be reduced.

Because ease of data reuse is one of the greatest advantages of data-centric networking [6], CCN is basically designed to be open: any user requesting some content by specifying its identifier will receive it. CCN assumes that the primary means of controlling access to content is encryption in a layer higher than CCN [6], [7].

However, for real-world networking, a completely open data-centric network is not sufficient. For example, it is expected that security threats that abuse the global openness, such as spamming and phishing, will become more frequent on data-centric networks. However, advanced security measures to solve these problems may reduce the convenience of networks in many cases.

In this paper, we focus on private communication within a closed group of users where only specific users can access content. In such *group-based communication* the above security issues are minimized.

We propose Virtual Content-Centric Networking (VCCN), which realizes group-based communication on a content-centric network. In VCCN every user can freely and dynamically create and change groups, as users are identified personally rather than by the host on which they reside. This has the advantage of preserving the location-independence of CCN [6].

The fundamental idea of VCCN is to operate a CCN router as logically independent multiple VCCN router instances by virtualization. Group-based communication is realized by building VCCN networks, each of which is composed of multiple VCCN router instances. In VCCN, a user communicates through an edge router that identifies the user and the relevant group memberships.

The main contributions of this paper are the following. First, we present a general and practical network architecture (VCCN) for constructing virtual private networks on

Manuscript received June 26, 2013; revised September 13, 2013; accepted January 16, 2014.

This research is partly supported by Grant-in-Aid for Scientific Research (B) (25280030).

a content-centric network by CCN router virtualization. Second, we show that VCCN is scalable with respect to content request rate and the number of VCCN networks, through a preliminary performance evaluation of our VCCN implementation.

The organization of this paper is as follows. Section II contains a summary of related work. In Section III we give an overview of VCCN and its four building blocks. In Section IV, we describe our VCCN implementation and the results of a preliminary performance evaluation. In Section V, we discuss open research issues in VCCN network construction. Finally, in Section VI we give our conclusions and indicate future work.

II. RELATED WORKS

One attempt to realize group-based communication on data-centric networks is the Virtual Private Community (VPC) service [8], [9]. VPC is a CCN-based service architecture designed to share content among users who belong to the same network domain or to external domains. In VPC, a virtual private community is built hierarchically from three types of members: creator, owners, and members. If a user is invited by the creator or owner of a virtual private community, the user can join the community and share content with its members. VPC realizes the construction of virtual private community for a group of users, but controlling access to content among the users is done simply by content encryption in a layer higher than CCN.

The VCCN design proposed in this paper was inspired by the Virtual Data-Oriented Network Architecture (VDONA) [10]. VCCN is similar to VDONA in the sense that a name space is split into multiple subspaces for enabling group-based communication. VCCN is, however, significantly different from VDONA in terms of how a router is virtualized and how packet transport between virtualized routers is accomplished.

III. VCCN (VIRTUAL CONTENT-CENTRIC NETWORKING)

A. VCCN Overview

In VCCN, several VCCN router instances are created on a CCN router and a network is built by logically connecting VCCN router instances. An example of such a VCCN network is shown in Fig. 1. Users are allowed to send Interest packets to VCCN networks that they belong to, and they can receive Data packets only from those networks. An Interest packet is routed within the VCCN network by the logically connected VCCN router instances. If the Interest packet is successfully delivered, the corresponding Data packet is sent back to the user within the VCCN network by traversing the path of the Interest packet in reverse.

The four building blocks of VCCN are as follows:

- **Extension of the content identifier**, which enables a virtualized CCN router to identify the VCCN network to which every Interest/Data packet belongs.

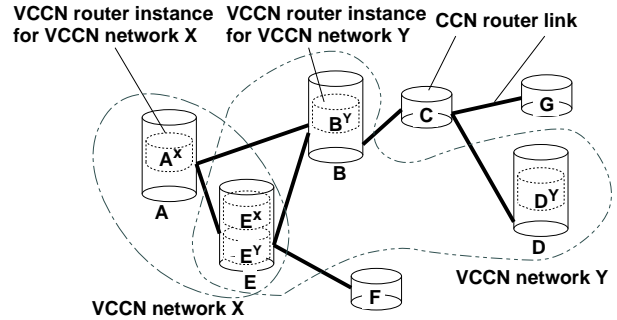


Figure 1. Example of a VCCN network built on a CCN network; two logically independent VCCN networks X and Y are built on the network of seven CCN routers, A through G .

- **CCN router virtualization**, which makes it possible to operate a single CCN router as multiple VCCN router instances.
- **Packet transport between virtualized CCN routers**, which enables packet delivery between virtualized CCN routers (i.e., CCN routers running VCCN router instances) which are not adjacent in the CCN network.
- **SNS cooperative user/group identification**, which enables virtualized CCN routers to identify the sender and the receiver of Interest and Data packets for realizing group-based communication.

The first three building blocks—extension of the content identifier, CCN router virtualization, and packet transport between virtualized CCN routers—realize traffic separation for VCCN networks. The last building block, SNS cooperative user/group identification, prevents injection of unauthorized traffic into a VCCN network by an outsider.

In the following, we describe these building blocks in more detail.

B. Extension of the Content Identifier

Content identifiers in CCN are extended to enable a virtualized CCN router to identify the VCCN network to which every Interest/Data packet belongs. Specifically, a VCCN identifier is embedded in a content identifier. Since content identifiers are variable-length bit strings, a VCCN identifier can be embedded in a content identifier in various ways.

An example of embedding a VCCN identifier in a content identifier is illustrated in Fig. 2. In this case, components of the content identifier are separated by slash delimiters. The first two components are used as the VCCN declaration and the VCCN identifier. Specifically, if the first component in a content identifier is `VCCN_ID`, a virtualized CCN router regards the packet as belonging to a VCCN network and treats the second component as a VCCN identifier. If the first component is not `VCCN_ID`, the content identifier is interpreted as a standard CCN content identifier. Such a simple extension of the content identifier enables the isolation of name spaces, one of which is assigned to every VCCN network.

/ VCCN_ID / groupX / x.com / videos / a.mpg / _v<timestamp> / _s3
VCCN declaration identifier **VCCN identifier** **standard CCN content identifier**

Figure 2. Example of an extended content identifier; the first two components are used as the VCCN declaration and the VCCN identifier.

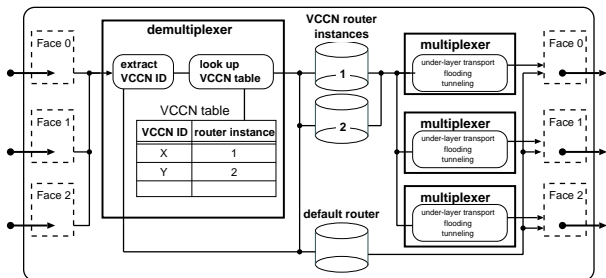


Figure 3. A virtualized CCN router; it is composed of a demultiplexer, VCCN router instances, and multiplexers.

C. CCN Router Virtualization

CCN router virtualization can be easily realized by switching three data structures used for packet routing in CCN: the forwarding information base (FIB), CS, and pending interest table (PIT) [6]. A CCN router can be equipped with multiple FIBs, CSs, and PITs and one of each of these tables is assigned to each VCCN network. The CCN router selects an appropriate set of FIB, CS, and PIT according to the VCCN identifier embedded in a content identifier.

A virtualized CCN router is composed of a demultiplexer, VCCN router instances, and multiplexers (see Fig. 3). We explain the operations of the demultiplexer, VCCN router instances, and multiplexers by describing the flow of packet processing.

An Interest/Data packet arriving at a face of a CCN router is first passed to the demultiplexer. The demultiplexer tries to extract a VCCN identifier embedded in the content identifier of the packet. If the VCCN identifier can be extracted, the demultiplexer checks whether a VCCN router instance corresponding to the VCCN identifier exists in the CCN router. If the VCCN router instance exists, the packet is passed to that instance. If the VCCN identifier cannot be extracted from the content identifier or the VCCN router instance does not exist, the packet is passed to the default router, which routes and forwards packets as an ordinary CCN router.

A VCCN table manages the correspondence between a VCCN identifier and a VCCN router instance. Each entry of a VCCN table is a pair of a VCCN identifier and an identifier of the corresponding VCCN router instance.

A VCCN router instance routes packets received from the demultiplexer using its own data structures (i.e., FIB, CS, and PIT), and it determines one or more faces through which to send the packet out. Note that the VCCN router instance uses the remainder of the content identifier (i.e., a content identifier in a VCCN network) rather than the entire content identifier. Finally, the CCN router emits the packet from one or more faces through

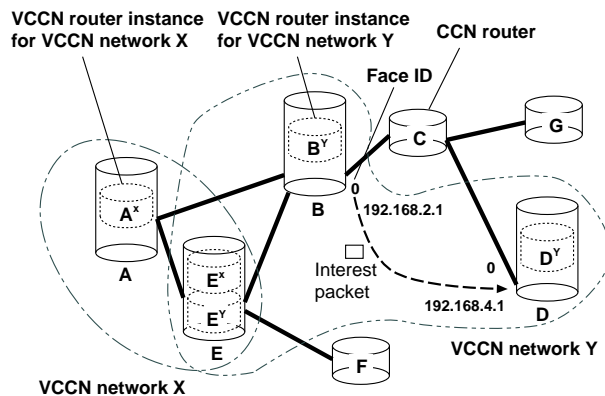


Figure 4. Packet transport in a lower layer; if a lower layer protocol supports communication between an arbitrary pair of nodes (e.g., IP, UDP, TCP, and broadcast communication), any pair of CCN routers can communicate using the lower layer protocol.

multiplexers, which are responsible for realizing packet transport between virtualized CCN routers.

D. Packet Transport between Virtualized CCN Routers

A multiplexer emits the packet received from a VCCN router instance through faces of the virtualized CCN router. The multiplexer enables packet transport between virtualized CCN routers, which are commonly not adjacent in the CCN network.

VCCN supports the following three types of packet transport between virtualized CCN routers.

- **Packet transport in a lower layer**
 The simplest and the most efficient approach is to use a protocol layer lower than CCN if that layer supports *any-to-any* communication (Fig. 4). CCN can operate on variety of lower layer protocols such as IP, UDP, TCP, broadcast communication, Ethernet, and P2P [11]. If a lower layer protocol supports communication between an arbitrary pair of nodes (e.g., IP, UDP, TCP, and broadcast communication), any pair of CCN routers can communicate using the lower layer protocol. Hence, packet transport between virtualized CCN routers can be easily realized.
- **Flooding in the CCN layer**
 If any-to-any communication is not supported in a lower layer protocol, then a simple approach is to flood the CCN layer (Fig. 5). In CCN, duplicate Interest packets are simply discarded. Hence, flooding can be realized simply by duplicating Interest packets and sending them through all faces of every CCN router. However, flooding is not efficient and might result in an excessive amount of traffic in a CCN network. So flooding should not be permitted, especially when VCCN networks are sparsely constructed.
- **Tunneling in CCN layer**
 A complicated but more efficient approach than flooding is to tunnel packets through intermediate CCN routers. Even when any-to-any communication is not supported in a lower layer protocol than CCN

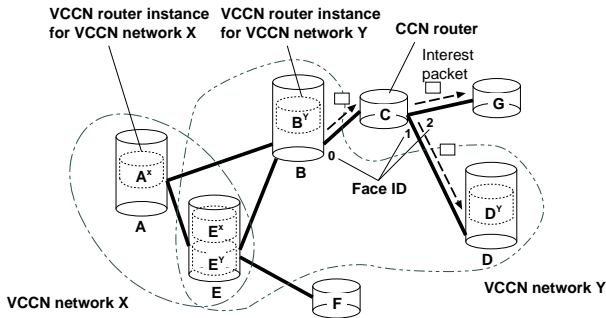


Figure 5. Flooding in the CCN layer; in CCN, flooding can be realized simply by duplicating Interest packets and sending them through all faces of every CCN router.

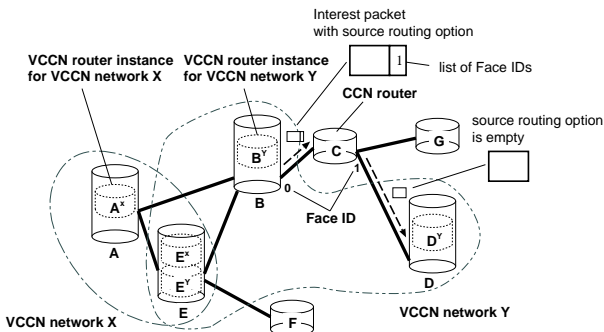


Figure 6. Tunneling in CCN layer; as in the source routing option of IP, the CCN router forwards the packet to the face listed at the head of the source route.

and inefficiency caused by the flooding in CCN layer is not acceptable, tunneling in CCN layer can transport packets between virtualized CCN routers (Fig. 6).

Since CCN is not a host-centric network architecture, tunneling in the CCN layer cannot be realized by a simple approach like IP-in-IP [12]. However, tunneling in the CCN layer is still realizable with source routing [13].

In CCN, a Data packet is sent back to the user by traversing the path of the Interest packet in reverse. Such path symmetry for Interest and Data packets is realized using the PIT as *bread crumbs* [6]. Hence, if a list of faces through which a packet should traverse is specified in any way, the locus of the packet can be controlled.

Based on this idea, Interest/Data packet headers are extended to store *source routing options* for realizing the tunneling in the CCN layer. Like the source routing option in IP [13], a CCN router forwards the packet to the face written at the head of the source route. Specifically, a multiplexer provides list of faces that the packet should traverse as a source routing option in the packet header. If a source routing option is specified in a packet, the demultiplexer in each CCN router pops the face from the head of the list, and transfers the packet through that face.

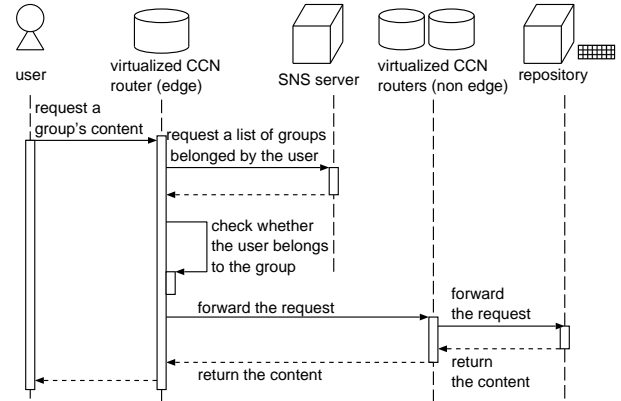


Figure 7. Sequence diagram for requesting content in VCCN.

E. SNS Cooperative User/Group Identification

Social Networking Services (SNSs) such as Facebook and Google+ have become increasingly popular in the last decade. In those SNSs, users can dynamically create and modify groups, each of which generally corresponds to a set of friends and colleagues.

In VCCN, to significantly simplify user/group management, virtualized CCN routers utilize user/group information registered in an SNS for authenticating senders and receivers of Interest and Data packets. That is, VCCN and SNS work cooperatively to realize group-based communication. We believe such a cross-layer cooperation between the network layer (i.e., VCCN) and the application layer (i.e., SNS) should dramatically ease the realization and management of user-aware communication services, such as group-based communication. Note that a similar idea has been proposed in SocialVPN [14].

In SNS cooperative user/group identification, virtualized CCN routers at the edge of a VCCN network identify whether a user is allowed to access that VCCN network. Access to a VCCN network is checked only at these edge routers; once an Interest packet has been forwarded, the downstream virtualized CCN routers do not care about the source of the Interest packet.

Basically, every router at the edge of a VCCN network is also a proxy for an SNS authentication service (see Fig. 7). When users want to access a VCCN network, they communicate with a router at the edge of the VCCN network and sends their identification information (e.g., username and password in an SNS). The router forwards the identification to the SNS server to check its validity and determine whether the user belongs to the group corresponding to the VCCN network. The user is allowed to send or receive packets only when both of these conditions are satisfied.

When a content is registered to a VCCN network, every router at the edge of the VCCN network is a proxy to an SNS authentication service, too (see Fig. 8). A repository communicates with a router at the edge of the VCCN network before content registration. The repository sends its identification information (e.g., repository name and password in an SNS). As with access to a VCCN network,

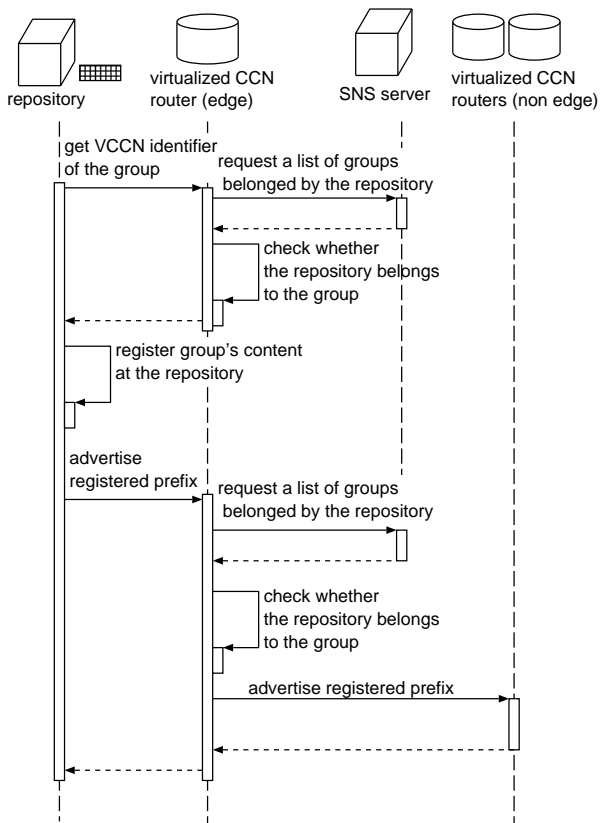


Figure 8. Sequence diagram when registering a content in VCCN.

the router forwards the identification to the SNS server, and checks to see whether the repository identification is valid and whether the repository belongs to the group corresponding the VCCN network. When both conditions are satisfied, the router returns the VCCN identifier of the group to the repository. The repository embeds the received VCCN identifier in a content identifier and registers the content. The repository then performs the *Register* operation [6] in order to advertise the prefix of the registered content to VCCN routers on the VCCN network. In a *Register* operation, the repository forwards its identification information and an Interest packet to advertise the prefix to the router at the edge of the VCCN network. The repository is allowed to advertise the registered prefixes to routers in the VCCN network only if both identification checks are successful again.

Although such cross-layer cooperation may debase the performance of a VCCN network, there are some remedies. For example, after the diffusion of CCN an edge router will be able to communicate with an SNS server, using CCN rather than through an application layer protocol (e.g., HTTP). Moreover, since a CCN router can cache content in its CS, the CCN router can reuse the group information that it has received from an SNS server. Hence, requests for the group information in the procedures of content request and registration can be skipped and these procedures will be simpler than what was estimated.

IV. IMPLEMENTATION AND EVALUATION

A. VCCN Implementation

We implemented VCCN’s basic features by extending the CCNx software [5], an open-source implementation of the CCN protocol. Our VCCN implementation is realized as wrapper programs for CCNx commands (e.g., *ccndstart*, *ccndstop*, *ccndc*, *ccndgetfile*, *ccndputfile*), and proxy software for SNS cooperative user/group identification. Our VCCN implementation allows users to initiate and terminate VCCN router instances, connect arbitrary VCCN router instances, and register and fetch content in a VCCN network.

Our VCCN implementation realizes traffic separation for VCCN networks in the following way. An edge router of a VCCN network embeds a user’s VCCN identifier in the content identifier immediately after the user requests some content through the wrapper programs. In our VCCN implementation, a CCN router is virtualized by logically splitting the FIB for each VCCN network: specifically, every FIB entry is tagged with a VCCN identifier. For simplicity, the CS and PIT are shared among all VCCN networks. Packet transport between virtualized CCN routers is realized with a lower layer protocol (UDP).

We prevent the injection of unauthorized traffic from a user using Facebook’s authentication mechanism. When a user/repository accesses a VCCN network, an edge router with the proxy software checks for the relevant authorization using the provided identification information and an access token. Specifically, the edge router uses the Graph API of Facebook [15] to perform user/group identification. The Graph API can acquire a user’s information from Facebook using an access token that is created at the time of the user’s login (Fig.9). In the implemented identification, when a user/repository accesses to a VCCN network, the user/repository passes an access token and a group name to the edge router of the VCCN network. The edge router makes identification by checking whether there is the specified group in the group list to which the user/repository belongs obtained through Graph API. If the user/repository belongs to the specified group, the edge router replaces the group name with the identifier managed by Facebook and embeds that group identifier in a content identifier. The edge router then looks up the FIB corresponding to the group and forwards the extended Interest packet to relay routers.

In our VCCN implementation, an outsider cannot request any content of a group through VCCN. In particular, our VCCN implementation can discard several types of illegal Interest packets: (1) an Interest packet that a user who does not belong to any group requests through VCCN; (2) an Interest packet that a user belonging to another group requests through VCCN; and (3) an Interest packet that a user belonging to the group requests through CCN. Fig. 10 shows the processes for discarding these three types of packet. In case (1), an edge router judges the user to be unauthorized and discards the Interest

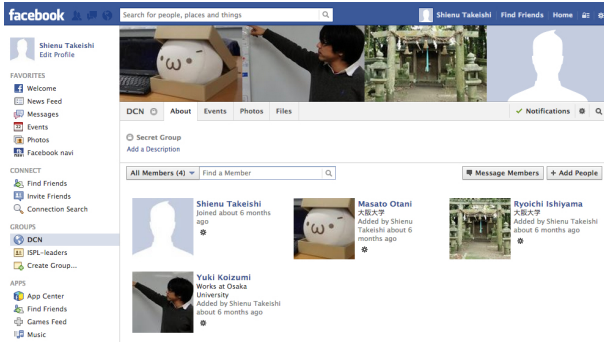


Figure 9. Example of creating a group; four members are registered with the data-centric networking group on Facebook and every registered member can take part in group-based communication.

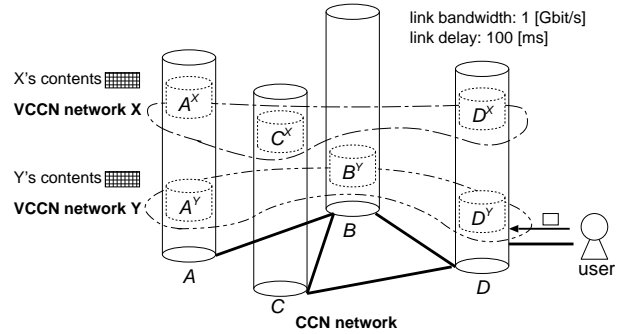


Figure 11. Network topology used in the CCNx/VCCN comparison; four CCN routers are connected and two VCCN networks, X and Y, are created.

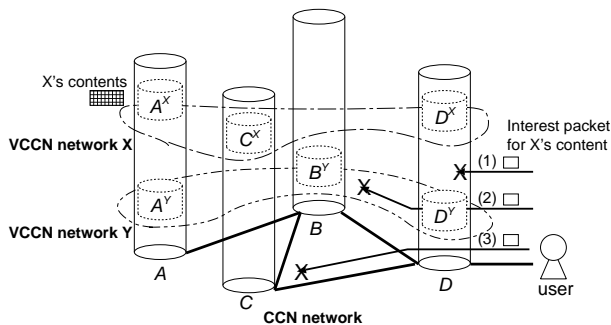


Figure 10. Processes for discarding three types of packet: (1) an Interest packet for X's content that a user who does not belong to any group requests through VCCN; (2) an Interest packet for X's content that a user belonging to Y requests through VCCN; and (3) an Interest packet for X's content that a user belonging to X requests through CCN.

packet during SNS cooperative user/group identification. In case (2), an edge router does not discard the Interest packet during SNS cooperative user/group identification. However, one of relay routers misses the longest-prefix matching of the Interest packet and discards it because it is transported in the VCCN network of a different group. In case (3), an edge router does not discard the Interest packet during SNS cooperative user/group identification. However, one of relay routers misses the longest-prefix matching of the Interest packet and discards it because it is not correctly extended based on a group identifier and is being transported in a global CCN network.

B. Performance Evaluation of the VCCN Implementation

We conducted preliminary performance evaluations of our VCCN implementation. In the first experiment, content delivery delays in our VCCN implementation and the original CCNx are compared. In the second experiment, we evaluate the scalability of virtualized CCN routers in a CCN network.

For the first experiment, we used the network topology shown in Fig. 11—four CCN routers are connected, and two VCCN networks X and Y are built.

In the CCNx setup, 100 items of size 10 [Kbyte] are stored in CCN router A, and CCN router D randomly requests one of those items 3,000 times. Note that the

hop count from the source (CCN router A) to the user (CCN router D) is always one.

In the VCCN setup, 50 items of size 10 [Kbyte] are stored in each of VCCN router instances A^X and A^Y . VCCN router instances D^X and D^Y randomly request one of those items in their VCCN network 3,000 times. Note that the average hop count from the source (CCN router A) to the user (CCN router D) is 1.5 (i.e., one hop in VCCN network Y and two hops in VCCN network X).

The communication delays of all links are identically set to 100 [ms] using network emulators. The size of the CS ($CCND_CAP$) is set to 100 in all CCN routers except CCN routers A and D, whose packet caching is disabled. We measured the content delivery delay disregarding the delays caused by identification processing.

Figure 12 shows the CDF (Cumulative Distribution Function) of content delivery delays in our VCCN implementation and in the original CCNx. Somewhat surprisingly, the content delivery delays in VCCN and CCNx are comparable even though VCCN has a larger hop count between the source and the consumer than CCNx: the average content delays were 2.79 [s] in VCCN and 2.53 [s] in CCNx. This similarity can be explained by the effect of content caching in CCN routers: CCNx utilizes the CS only in CCN router B, but VCCN utilizes the CSs in routers B and C. For instance, in our experiment, the average cache hit rate of CCN routers with VCCN was 51.8% whereas that without VCCN was 44.9%. VCCN router instances are dispersed in the network, so that VCCN can effectively utilize, at least in this experiment, the content stores in CCN routers.

It should be noted that efficiency of VCCN relies significantly on several factors, such as the CCN and VCCN network topologies, so we do not claim that VCCN is more efficient than CCN. Instead, we just addressed the question of whether the introduction of VCCN has a positive or negative impact on CCN performance. We are planning to conduct more detailed experiments.

Secondly, since it is expected that the performance of VCCN networks will be debased by CCN router virtualization, we evaluated the scalability of virtualized CCN routers in our VCCN implementation. We consider

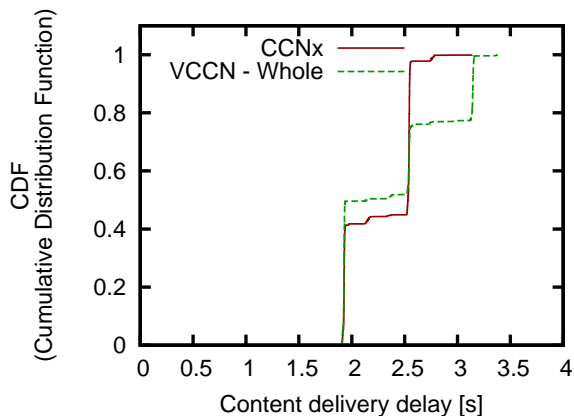


Figure 12. CDFs for content delivery delay when content is requested through a CCN network and the VCCN networks.

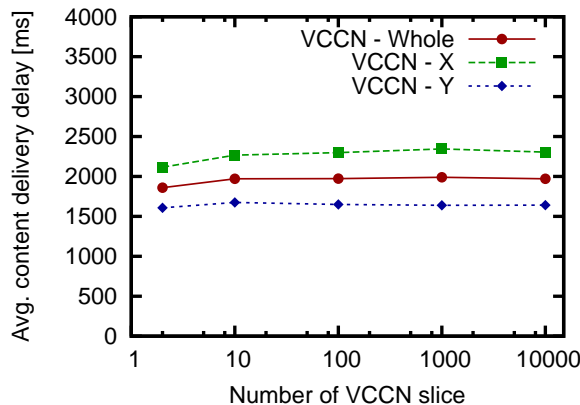


Figure 14. Average content delivery delays against the number of VCCN networks in a CCN network.

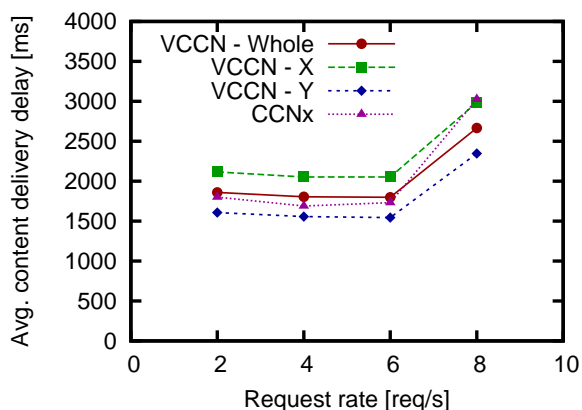


Figure 13. Average content delivery delays against content request rate.

two types of scalability: scalability with respect to request rate and scalability with respect to the number of VCCN networks. In the experiment, the result of setting the number of content items, which are stored in CCN router A or VCCN router instances A^X and A^Y , to 10,000 and setting $CCND_CAP$ to 1,000 were investigated.

Fig. 13 shows average content delivery delays against content request rate. This figure indicates that there is almost no change in average content delivery delay until the request rate reaches 8 [request/s]. Moreover, this figure indicates that CCN router virtualization does not affect the performance of a VCCN network because the average content delivery delays in our VCCN implementation and in CCNx both increase at a rate of 8 [request/s]. Therefore, a VCCN has the scalability of a virtualized CCN router with respect to content request rate.

Next we considered how average content delivery delays vary with the number of VCCN networks in a CCN network. When increasing the number of VCCN networks, the number of content items in the CCN network is fixed and equal numbers of topologies X and Y for the VCCN networks are constructed. For example, when the number of VCCN networks is 10, there are five X and five Y VCCN network topologies. In this experiment, the content request rate is set to 2 [request/s]. Fig. 14

indicates that the performance of virtualized CCN routers is not debased even if the number of VCCN networks is substantially increased. Moreover, the fact that the average content delivery delays are maintained does not depend on the network topology. Therefore, VCCN also has the scalability of a virtualized CCN router with respect to the number of VCCN networks that can be accommodated.

V. OPEN ISSUES

In this section, we discuss open research issues of VCCN network construction based on knowledge acquired by designing, implementing and evaluating VCCN networks.

A. CCN Router Resource Management

One important issue for virtualizing a CCN network is how resources (i.e., the FIB, CS, and PIT) of a CCN router are allocated to each VCCN router. Since a CCN router uses the three structures for routing a packet, the allocation of these resources affects the performance and robustness of a network.

In CCN network virtualization, we will have to focus on the trade-off between overall performance and fairness. Sharing the resources of a CCN router among groups/applications is better than allocating the resources to each group/application in order to maximize overall network performance [16], [17]. On the other hand, sharing the resources of a CCN router among groups may cause unfairness between groups. For instance, if the CS of a physical CCN router is shared between VCCN routers and the traffic of a certain group is especially large, the CS can be effectively occupied by the VCCN router of that group [16]. Then, while network performance for the group, whose VCCN router occupied the CS, may be very high, network performance for the other groups will be low. In a similar way, if the PIT is monopolized by a certain group, users of the other groups will not be able to communicate. This also means that, if a malicious user can gain access to any VCCN network, that user can obstruct another VCCN network by interest flooding [18].

To prevent resource occupation of a CCN network and improve overall network performance, we should design a method to allocate the resources of a physical CCN router to each group. Some related methods have been proposed [17], [19] and our research group is planning to investigate analytically the effect of CS allocation methods and content request patterns in VCCN networks on the average content delivery time of each separate VCCN network and the entire network.

B. VCCN Network Mapping

The virtual network mapping/embedding problem, which means mapping virtual routers and links to specific nodes and links in the substrate network, has been investigated in previous studies of virtualization [20]–[22]. Since this mapping is an NP-hard problem, heuristics for providing efficient performance were proposed in these studies.

In CCN network virtualization, existing virtual network mapping methods may not be applicable because these methods do not take data reuse into account. Mapping VCCN networks influences the effect of caching as well as performance and traffic. For example, in the experiment of Section IV, the content delivery delays in VCCN and CCNx are comparable due to a change of caching effect, despite the mapping increasing the average hop count from the user to the source. Furthermore, the efficiency of caching and network performance may be increased by increasing the number of relay VCCN routers in a VCCN network. It is desirable to study this problem, taking the effect of caching into account.

C. Reliability

Although VCCN is a general and practical network architecture, there are some improvements required in order for a VCCN to operate as a reliable network architecture in various environments.

One necessary improvement is the decentralized management of a VCCN declarator and VCCN identifiers. VCCN realizes traffic separation between VCCN networks and a substrate CCN network by checking if a VCCN declaration exists. Moreover, as in IP-VPN, VCCN uses label switching based on a VCCN identifier. Hence, in VCCN, it is necessary that an unauthorized user cannot specify a valid VCCN declaration and identifier. Our VCCN implementation solves this problem by defining a VCCN declaration `VCCN_ID` as a block phrase and getting Facebook to manage the VCCN identifiers of all groups. However, this solution places a lot of management load on Facebook. If the decentralized management of VCCN identifiers can be realized, VCCN will be more reliable network architecture. Moreover, if VCCN networks are constructed on a CCN network composed of multiple autonomous systems, the decentralized management of VCCN declarators and identifiers must be performed reliably between the autonomous systems.

Another requirement is a lightweight and robust authentication mechanism, since routers at the edge of a VCCN

network authenticate users and consequently experience a huge load. On the other hand, countermeasures against the attacks of a malicious user should be implemented. For instance, a malicious user may attempt a denial-of-service attack on a VCCN network by repeatedly accessing an edge router because of the load applied to the router in SNS cooperative user/group identification. This method may also be used to attack the authentication server itself. In regard to these attacks, we will need not only to divide authentication processes and routing processes between a control plane and a forwarding plane but also implement a quick and lightweight authentication mechanism in order to prevent a CCN network going down.

VI. CONCLUSIONS

In this paper, we have proposed VCCN, which realizes group-based communication through CCN router virtualization. The fundamental idea is to operate a CCN router as multiple instances of VCCN routers, which run logically independently. Group-based communication is realized by building VCCN networks, which are composed of multiple VCCN router instances.

We have implemented VCCN's basic features by extending the CCNx software and have conducted a preliminary performance evaluation of our implementation. The evaluation showed that virtualization has both positive and negative impacts on CCN performance and has the scalability of virtualized CCN routers with respect to request rate and the number of VCCN networks. We have also discussed open research issues in VCCN network construction based on knowledge acquired by designing, implementing and evaluating VCCN networks.

In the future we will consider who names and manages VCCN identifiers and how such tasks should be done. We will also consider where VCCN router instances should be created or removed when a group is changed.

REFERENCES

- [1] S. Shenker, "The data-centric revolution in networking," in *Proceedings of the 29th International Conference on Very Large Data Bases (VLDB 2003)*, Sept. 2003, p. 15.
- [2] C. Esteve, F. L. Verdi, and M. F. Magalhães, "Towards a new generation of information-oriented internetworking architectures," in *Proceedings of the first Workshop on Re-Architecting the Internet (ReArch 2008)*, Dec. 2008, pp. 1–6.
- [3] J. Choi, J. Han, E. Cho, T. Kwon, and Y. Choi, "A survey on content-oriented networking for efficient content delivery," *Communications Magazine, IEEE*, vol. 49, no. 3, pp. 121–127, Mar. 2011.
- [4] K. Cho, J. Choi, D. il Diko Ko, T. Kwon, and Y. Choi, "Content-oriented networking as a future internet infrastructure: Concepts, strengths, and application scenarios," in *Proceedings of the third International Conference on Future Internet Technologies (CFI 2008)*, June 2008.
- [5] "Project CCNx," <http://www.ccnx.org/>.
- [6] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the fifth International Conference on emerging Networking EXPERiments and Technologies (CoNEXT 2009)*, Dec. 2009, pp. 1–12.

- [7] L. Zhang *et al.*, “Named Data Networking (NDN) project,” <http://www.named-data.net/ndn-proj.pdf>, Palo Alto Research Center, Tech. Rep. NDN-0001, Oct. 2010.
- [8] D. Y. Kim, M. Wuk Jang, B.-J. Lee, and K. Kim, “Content-centric network-based virtual private community,” in *Proceedings of the 29th International Conference on Consumer Electronics (ICCE 2011)*, Jan. 2011, pp. 843–844.
- [9] D. Y. Kim and J. Lee, “CCN-based virtual private community for extended home media service,” *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 532–540, May 2011.
- [10] K. Kanamori, “A data-oriented network architecture for group-based communication,” Master’s thesis, Graduate School of Information Science and Technology, Osaka University, Feb. 2010.
- [11] W. A. Simpson, “The Point-to-Point Protocol (PPP),” *Request for Comments (RFC) 1661*, July 1994.
- [12] —, “IP in IP tunneling,” *Request for Comments (RFC) 1853*, Oct. 1995.
- [13] J. Postel, “Internet protocol,” *Request for Comments (RFC) 791*, Sept. 1981.
- [14] P. S. Juste, D. Wolinsky, P. O. Boykin, M. J. Covington, and R. J. Figueiredo, “SocialVPN: Enabling wide-area collaboration with integrated social and overlay networks,” *Computer Networks*, vol. 54, no. 12, pp. 1926–1938, Aug. 2010.
- [15] “Facebook Graph API,” <http://developers.facebook.com/docs/reference/api/>.
- [16] K. Ohsugi, K. Tsukamoto, and H. Ohsaki, “A study on the effect of ccn router virtualization on content delivery time (in Japanese),” *the 2012 IEICE Society Conference (B-7-4)*, p. 83, Sept. 2012.
- [17] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, “Evaluating per-application storage management in content-centric networks,” *Elsevier Computer Communications*, vol. 36, no. 7, pp. 750–757, Apr. 2013.
- [18] P. Gasti, G. Tsudik, E. Uzun, and L. Zhang, “DoS and DDoS in named-data networking,” <http://arxiv.org/abs/1208.0952>, Tech. Rep., 2012.
- [19] G. Carofiglio, V. Gehlen, and D. Perino, “Experimental evaluation of memory management in content-centric networking,” in *Proceedings of 10th IEEE International Conference on Communications (ICC '11)*, June 2011, pp. 1–6.
- [20] G. P. Alkmim, D. M. Batista, and N. L. da Fonseca, “Mapping virtual networks onto substrate networks,” *Internet Services and Applications*, vol. 4, no. 3, pp. 1–15, Jan. 2013.
- [21] J. He, R. Zhang-Shen, Y. Li, C. yen Lee, J. Rexford, and M. Chiang, “DaVinci: dynamically adaptive virtual networks for a customized internet,” in *Proceedings of the fifth International Conference on emerging Networking EXperiments and Technologies (CoNEXT 2008)*, Dec. 2008, p. 15.
- [22] J. Lu and J. Turner, “Efficient mapping of virtual networks onto a shared substrate,” WUCSE-2006-35, Tech. Rep., 2006.

Keiichiro Tsukamoto received Bachelor of Information Science and Master of Information Science degrees from Osaka University in 2009 and 2011, respectively. He has been a graduate student of the Graduate School of Information Science and Technology, Osaka University since April 2011. His research interests are in the area of Web content mining and Content-Centric Networking. He is a member of the IEICE.

Masato Ohtani received Bachelor of Information Science and Master of Information Science degrees from Osaka University

in 2011 and 2013, respectively.

Yuki Koizumi received the M.E. and D.E. degrees in Information Science from Osaka University, Japan, in 2006 and 2009, respectively. He is currently an Assistant Professor at the Graduate School of Information Science and Technology, Osaka University, Japan. His research interest includes traffic engineering in photonic networks and biologically inspired networking. He is a member of IEEE and IEICE.

Hiroyuki Ohsaki received the M. E. degree in the Information and Computer Sciences from Osaka University, Osaka, Japan, in 1995. He also received the Ph. D. degree from Osaka University, Osaka, Japan, in 1997. He is currently a professor at Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Japan. His research work is in the area of design, modeling, and control of large-scale communication networks. He is a member of IEEE, IEICE, and IPSJ.

Kunio Hato received B.E. and M.E. degrees from Tokyo Institute of Technology in 1997 and 1999, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 1999, he has been engaged in research and development of IP VPNs, Wide Area Ethernet, network security systems and intercloud computing systems. He is now a senior research engineer in the Secure Communication project of NTT Information Sharing Platform Laboratories. He is a member of IEICE.

Junichi Murayama received B.E. and M.E. degrees in electronics and communication engineering from Waseda University in 1989 and 1991, respectively. He also received Ph.D. degree in information science and technology from Osaka University in 2011. From 1991 to 2013, he had worked for Nippon Telegraph and Telephone Corporation (NTT). He is currently a professor at Department of Communication and Network Engineering, School of Information and Telecommunication Engineering, Tokai University, Japan. He has been engaged in research and development of ATM networks, IP VPNs, optical IP networks, network security systems and intercloud computing systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Institute of Electrical Engineers of Japan (IEEJ).

Adaptive Distributed Load Balancing Routing Mechanism for LEO Satellite IP Networks

Xiao Ma

School of Computer and Telecommunication Engineering, University of Science and Technology Beijing, Beijing, 100083, P. R. China

Email: mxustb@hotmail.com

Abstract—LEO (Low Earth Orbit) satellite constellation is an ideal scheme for the next generation wideband internet. The constellation is formed as a mesh-like network with inter satellite links (ISLs) which are equipped between the neighbor satellites for transmitting directly. As to the future wideband IP services, an efficient routing mechanism will play an important role in improving the performance and balancing the network traffic. An Adaptive Distributed Load Balancing Routing Mechanism (ADLB) is proposed in this paper to address the above-mentioned issues. This mechanism makes well-performed routing decision based on the current and historical status of each ISLs in each satellite node. With collecting historical information from network initiated, a proper mechanism is contained in ADLB for making required computing power and storage space in a reasonable range. The performance of ADLB is verified via a series of simulations which demonstrate that the scheme can provide better throughput and lower packet drop rate.

Index Terms—Load Balancing; ISL; Routing; Adaptive; Distributed Routing

I. INTRODUCTION

Nowadays, satellites are wildly used in long distance communication. GEO (Geostationary Earth Orbit) satellite has wide coverage area but the signal delay makes negative effect for real-time service. N GEO (Non-GEO) satellite has shorter signal propagation delay than GEO systems. The N GEO satellite meets the requirements for interactive multimedia communication, and will be the essential part of the NGI (Next Generation Internet) [1].

As the coverage area of N GEO is considerably less than that of GEO, more N GEO satellites are needed for full coverage. To provide sufficient coverage, a constellation is needed with multiple satellites. Inter satellite links (ISLs) are used in most posted LEO constellations for the satellite nodes connecting directly with the neighboring satellites thus a mesh-like network can be formed. How to connect the satellites with ISLs together to form a satellite communication network [1-4] becomes an active research area. Therefore, the LEO satellite networks will become the backbone of the network for the ground terminal accessing.

Many problems associated with the system, such as propagation delay and the unbalanced loading with ISLs,

necessitates the request for an efficient routing mechanism. Furthermore, effective routing is key to ensuring inter-satellite data transmission, distribution and other functions [5, 6].

There is a common defect in the traditional source mechanism. The information that access node collected may be outdated for routing which cannot reflect the actual states of the constellation because of the large propagation delay, thus there is a great impact to the algorithm performance.

Centralized routing technology is able to achieve better global traffic engineering, but fails to solve the problem of the overhead and the real-time transmission of traffic information. In addition, the scalability of the centralized routing technology is poor due to the limited capacity of the central node with an expanding network and an increasing computational complexity.

According to this, we propose a distributed routing mechanism named Adaptive Distributed Load Balancing Routing Mechanism for LEO satellite IP networks (ADLB), which has capability of self-adapting to the dynamic traffic changes. The algorithm chooses proper links as routing paths by a new normalization method to utilize the historical and current link information based on the current node queue length and packet drop rate. In this algorithm, all the transmitted packets are forwarded in the dynamic traffic network for self-load-balancing purpose. We determine the best ratio between the link historical and current information, and the simulation shows that the proposed mechanism not only decreases the packet drop rate but also increases the maximum throughput.

The structure of this paper is as follows: in Section II, we discuss and compare the existing load-balancing routing algorithms, then in Section III, we describe the design of the proposed routing algorithm in detail. A simulation model is proposed for the distributed algorithm in Section IV and the simulation results are presented and analyzed in Section V. Conclusion and summary are given in Section VI.

II. EXISTING LOAD-BALANCING ROUTING ALGORITHMS

The performance of routing mechanism depends on the frequency of obtaining the network status. The traditional source/centralized load balancing routing mechanism has

a common drawback that the routing information may be outdated thus does not reflect the actual status.

In the source/centralized routing algorithm in LEO satellite networks, the source node relies on the topology regularity of the LEO satellite networks in routing process. If a link or node failure occurs, the source node could not detect link (or node) failure and congests. DRA [7] (Datagram Routing Algorithm) is an earlier load-balancing routing mechanism which chooses a shortest transmission path for each packet using the predictability of the polar orbit LEO satellite network topology. But when link congestion occurs, the routing performance will decline rapidly because there are no exchanges in network status and no control information between the satellites. CEMR [8] (Compact Explicit Multi-path Routing) achieves load-balancing depending on multi-path mechanism and queuing delay predicting, but neither the congestion status nor the queuing delay of the next hop can be informed by this mechanism as well as the packet drop rate.

This situation has been partially improved in the distributed load-balancing routing mechanism, that each satellite node independently calculates the routing table and forwards the packet which has a good adaptability to flow changes. ELB (Explicit Load Balancing) [9, 10] is a protocol to ensure satellite network load balancing purposes intelligently by exchanging congestion status information. Once the satellite turns into the busy state, it sends the BSA (Busy State Advertisement) to the neighbor satellite to request an adjustment to its packet transmitting ratio before congestion and packet drop occur. The neighbor satellite reduces its data rate and finds another alternative path excluding the busy node to ensure better flow distribution, avoiding the congestion and packet losses. But the feedback of network congestion situation is entirely dependent on the heavy load satellites while the large number of busy status notification should increase the network burden even more. LAOR [11] (Location-assisted On-demand Routing) is the modified version of AODV. It is adjusted to receive the flood routing request using the grid-like LEO network topology. The recent path should be failure because of the ISLs high propagation delay and the dynamic characteristics of LEO satellite network, whereas the algorithm doesn't propose a solutions. In addition, each satellite must configure a LAOR queue to store packets that are waiting for allocating paths temporarily. PAR (Priority-based Adaptive Routing) and enhanced PAR [12] are designed based on the priority of adaptive routing mechanisms relying on the priority of the ISL historical usage and buffer information. PAR tends to choose the low usage link for packet transmission at per hop. PAR doesn't exchange information with its neighbor node thus the overhead has less impact on the network. However, only the past several time intervals is involved in the historical information. As the same question, the metric parameters also contain duplicates and conflicts..

In recent year, some people gain research on multi-layer satellite network. The proposed algorithms regard the GEO/MEO layer nodes as relay or backbone nodes,

relatively the LEO satellites are often used as only access nodes or routing in a certain area. A congestion prediction mechanism are proposed for those GEO/LEO hybrid satellite networks [13], which has ability on traffic load balancing and QoS guarantee for improving the performance of real-time and non-real-time transmitting. In the mechanism, the network efficiency is enhanced and the QoS requirement of terminal is satisfied. In the next research, a cross-layer distribution traffic load balancing mechanism is proposed for achieving optimized result [14]. A distributed flow model is adopted. This model is based on network capacity estimation and the congestion probability theory analysis of each layer. Because of the parameter setting optimization, a better throughput and lower drop rate are gained. Some researchers focus on the security routing mechanism [15]. A cluster-based protocol is designed for multi-layers satellite network using ID-based sign scheme.

Taking these remarks into account, the objective of this research is to develop a distributed load-balancing routing mechanism aiming to achieve lower overhead and give full consideration of historical information from the network initiated with the minimum calculation and storage cost.

Currently only at the theoretical research stage, the fast-moving satellites in multi-layers satellite network require high stability of communication servo system with high technical risks and cost. But the LEO single-layer global coverage satellite network was put into commercial operation. The network structure is confirmed feasible. The researching is positive over this network structure. CEMR and PAR are representative distributed load balancing routing mechanism in recent years, so that the simulation results of ADLB should be compared with those of CEMR and PAR.

III. OPERATION OVERVIEW

The proposed mechanism relies on the ISL's state determined by the queue length and drop ratio for the routing decision. We use a formula to express the relationship between these two parameters, and the value of this formula is called *status parameter* λ_n . In the formula, the queue length is defined as the buffer occupancy during a time interval of the satellite, the drop ratio reflects the relationship between packets transmitted and dropped.

ADLB not only considers the current interval state λ_n but also its historical statistics to reflect the continued usage of the link. This enables us to obtain the key value for routing, called *Routing Factor* (F_n). When a newly-arrived (via ISL or the sat-to-ground link) packet needs to be forwarded, the satellite detects the minimum F_n from the four available ISLs sources for the forwarding path.

The value of F_n will continue to grow with time so that it is difficult for the satellite to maintain a sufficient calculation and storage capacity. Therefore, a normalization process is proposed to scale F_n in an acceptable range.

A. System Model

In this study, a polar constellation with ISLs is considered for its advantages such as simplified constellation management and relaxed challenges in global coverage design. Most LEO global coverage projects are designed to equip ISLs to directly form a complete network with the satellites in the constellation; otherwise, the communication between satellites must depend on ground station relay. Each satellite is assigned four ISLs in order to communicate with its four neighboring node; two of them are located in the same orbit, whereas the other two in its adjacent orbits. Therefore, inter-plane ISLs connect the satellites in the same orbit. Similarly the intra-plane ISLs connect the ones in the left hand and right hand orbits. One satellite equips four independent buffers for the ISLs in the four directions to accommodate their respective queues.

The following study concentrates on routing within the network based on the satellites. The network can be modeled as a graph $G(V,E)$, comprising of a set of nodes V and a set of edges E . Set V indicates the satellite node in the network, and set E models the ISLs connecting the adjacent satellites. Each $v_i \in V (i = 0, 1, \dots, p)$ represents a satellite node in the network as each satellite is assigned a unique number from 0 to p . However, with the two possible transmission directions between v_i and v_j , e_{ij} describes the link from v_i to v_j while e_{ji} denotes the opposite direction ($e \in E$).

B. Setting of Status Parameter

In ADLB, the key of status optimization is to reflect the buffer usage and the packet drop ratio in the current interval under the actual operating conditions of this ISL. The satellite performance differs for different networks in their buffer size, ISL and ground terminal data rate. As such, system compatibility and universal applicability demand for a parameter manifesting the ISL's relative performance, rather than the absolute value. According to this, we use the status parameter as λ_n , where n is the n th interval.

In light of the above design merit, the queue length is defined as the ratio of the buffer occupancy l_q and the whole buffer length B in the current interval. The drop ratio is defined as the ratio of the dropped packets P_d and the transmitted packets P_{all} . Assume that we only consider the packet loss due to link congestion, regardless of other reasons, the status parameter can be expressed as follows:

$$\lambda_n = \frac{l_q}{B} + \frac{P_d}{P_{all}} \quad (1)$$

Notice that $\frac{l_q}{B} \in [0, 1]$ and $\frac{P_d}{P_{all}} \in [0, 1]$, so $\lambda_n \in [0, 2]$.

If the link is idle in the past interval, $\lambda_n = 0$. Under normal condition, λ_n hardly reaches the boundary (0 or 2).

$\lambda_n = 0$ when the network are running in the current interval if no packet is dropped and no buffer is occupied.

$\lambda_n = 2$ is an extreme condition when all the packets are dropped and the buffer is completely full.

C. Setting of Routing Factor

The main objective of the Routing Factor parameter F_n is to guide the satellite node to make a proper routing decision which indicates the up-to-date ISL performance. When a packet arrives at an intermediate node, the satellite will check the routing table for an appropriate ISL to the next hop. The routing decision, apart from solely depending on the current status, resorts to the usage history. This is because a currently good λ_n does not necessarily infer a satisfactory historical record.

It should be emphasized that the historical status is less important than the current interval status λ_n . This can be compensated for by a *decay function* a^n which weights preferentially against earlier values. a^n as an increment function which, when multiplied with λ_n , approaches zero as n decreases. This defines $a^n \in [0, 1]$.

Accordingly, assuming the index of the current interval as n , the process for getting F_n is

$$F_n = \sum_{m=0}^n a^{n-m} \lambda_m \quad (2)$$

From Eq.2, it is evident that calculating F_n and storing historical λ_m become increasingly memory-consuming as n grows. This forces one to compromise between accuracy and speed when the summation in eq.2 needs to be truncated.

It can be found that equation (2) is actually an iterative process

$$\begin{aligned} F_n &= \lambda_n + a \sum_{m=0}^{n-1} a^{n-1-m} \lambda_m \\ &= \lambda_n + aF_{n-1} \end{aligned} \quad (3)$$

from which F_n is derived from λ_n and aF_{n-1} directly. In detail, calculating F_n involves 3 steps: 1) calculate λ_n ; 2) multiply a with F_{n-1} and 3) add results in 1) and 2). Thus, the requirements of satellite computing and storage capacity will be greatly reduced.

As time goes on, F_n gradually increases and may finally reach the storage limit. In order to solve this we propagate the normalizing procedure. This procedure limits F_n in scale to $0 \leq F_n \leq 1$ for a certain range.

Variable $\frac{1}{1+a}$ is multiplied to (2):

$$F_n = \sum_{m=0}^n \frac{a^{n-m}}{(1+a)^{n-m+1}} \lambda_m \quad (4)$$

Let $\frac{a}{1+a} = \alpha$, compare with (3), we can acquire the normalized F_n in (5):

$$F_n = (1-\alpha)\lambda_n + \alpha F_n \quad (5)$$

Notice that α satisfies $\begin{cases} \alpha < 1-\alpha \\ 0 < \alpha < 1 \end{cases}$ yielding $\alpha \in \left(0, \frac{1}{2}\right)$.

D. Direction Estimation Process

Satellite nodes are capable to arrange the optimized link for routing requests after the above procedure. However, a possible situation may exist when the destination node is in the opposite direction to the current node. The direction estimation process will exclude those links from the alternative links in order to complete the routing decision in the oriented direction of the destination node.

This process divides the network into 4 areas regarding the current node as the center. The position of the destination node which should be excluded from the routing regions has two cases: 1. in the area, 2. at the boundary of two areas. Only the ISLs in the area or on the boundary can be selected as alternative paths otherwise all the others will not be considered.

The process is suitable for grid networks, and also for other symmetric ISLs LEO networks with few modifications.

This method not only avoids loops but also eliminates the possibility that a selected link with optimized routing factor is not toward the destination node. The opposite directions induce more hops to the destination node. As a result, those links are deemed disadvantageous compared with the “forwards links” in the area because of the associated delay.

```

Procedure Direction_Estimation( $S_i, S_d$ )
  Given: The coordinate of node  $S_i: (x_i, y_i), S_d: (x_d, y_d)$ ,
         the alternative ISLs  $E = \{e_{ih}, e_{ij}, e_{ik}, e_{il}\}$ 
  Find: The proper directions towards the destination  $E$ .
  Get Area(direction).
  If  $S_i$  in Area(up_left) then  $E = \{e_{ih}, e_{ij}\}$ .
  else if  $S_i$  in Area(low_left) then  $E = \{e_{ij}, e_{ik}\}$ .
  else if  $S_i$  in Area(low_right) then  $E = \{e_{ik}, e_{il}\}$ .
  else if  $S_i$  in Area(up_right) then  $E = \{e_{il}, e_{ih}\}$ .
  Return  $E$ .
    
```

Figure 1. Drop rate at different α with different transmitting bit rates

Suppose a routing decision event is launched in node S_i . Its adjacent nodes arranged in accordance with the upper, right, lower, left in clockwise order are $\{S_h, S_j, S_k, S_l\}$. The destination node is S_d .

Definition 1: Area (direction) is the routing region of, while

$$direction = \begin{cases} up_left, & x_d - x_i > 0, y_d - y_i > 0 \\ low_left, & x_d - x_i > 0, y_d - y_i < 0 \\ low_right, & x_d - x_i < 0, y_d - y_i > 0 \\ up_right, & x_d - x_i < 0, y_d - y_i < 0 \end{cases} \quad (6)$$

Then we propose the direction estimation procedure after definition 1.

Fig. 1 describes the process of obtaining the alternative links for routing decision when the routing request occurs in the current node.

IV. PERFORMANCE EVALUATION

A. Simulation Setup

In this section, we evaluate the performance of the proposed scheme. The Iridium-like constellation is studied which is formed by 66 satellites evenly distributed in six orbits. In our simulation platform, it is considered that there are no ISLs between the counter-traveling orbits as a seam, so that these two orbits are at the left and right boundaries of the mesh network topology. The rest of the simulation parameters are presented in Table 1. During the experiment, all links are set to error-free in order to better expose the performance difference among various mechanisms. In the simulation process, the average packet size is set to 1KB, and the ISLs delays are set to 20ms. The buffer of each ISL is set to 200kb (storage capacity of 200 packets). The rest of the simulation parameters are shown in Table I.

TABLE I. SIMULATION PARAMETERS

Number of orbits	6
Number of satellites per plane	11
Satellite altitude	780km
Cross-seam ISLs	No
Number of ISLs	2 intra-plane + 2 inter-plane
ISL bandwidth	25Mb/s
ISL buffer size	200 Packets

In the simulation, the network traffic is generated by 400 earth stations which are distributed in the world-wide continents following [10] in Table II, and all numbers represent the percentage (%). The earth stations provide On-Off flows connecting to the satellites using Poisson arrival process. Furthermore, the duty cycle is set to 400ms evenly split between burst time and idle time. The earth stations send data at the same rate from 0.8Mbps to 1.5Mbps in the same simulation process.

TABLE II. DISTRIBUTION OF EARTH STATIONS

Dest. \ Src.	NA	SA	EU	AF	AS	OA
NA	60	10	15	2	10	3
SA	35	40	12	2	8	3
EU	40	5	40	2	10	3
AF	40	2	30	20	5	3
AS	20	2	10	2	50	6
OA	40	2	10	2	12	34

B. Simulation Results

In the performance evaluation, we first experimented on the parameter α which indicates the combined effects of the current and the historical link status. Then additional routing algorithms are selected for comparison.

1) *The Effects of Parameter α* : As described above, the parameter α plays an important role in the routing process in determining the routing decisions. We evaluated the packet drop rate under different data rates

of the earth stations by varying the value of α from 0.1 to 0.9. The result is shown in Fig. 2 representing bit rates from 0.8Mbps to 1.5Mbps.

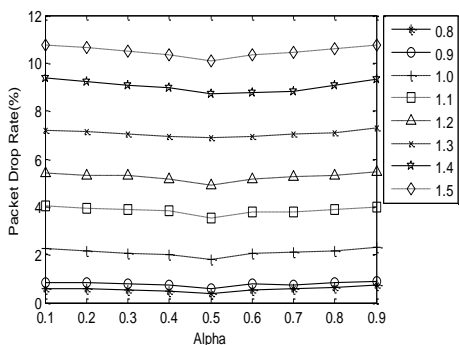


Figure 2. Drop rate at different α with different transmitting bit rates

To highlight the effect of α , we use the *Comparison Drop Rate (CDR)* to describe the change of drop rate normalized by the maximum rate (due to the different α) as (7) and the result is presented in Fig. 3.

$$CDR = \frac{r_{max} - r_i}{r_{max}} \quad (7)$$

where r_i is drop rate under current α and bit rate, Parameter r_{max} is the maximum drop rate during various α .

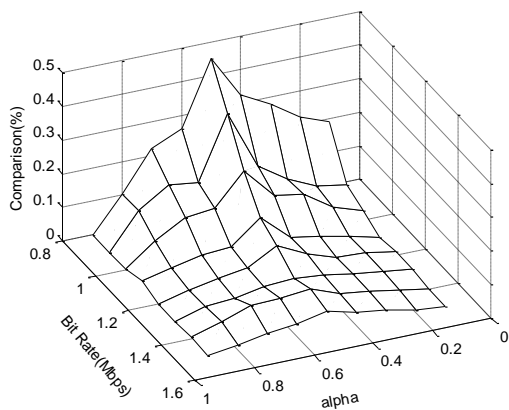


Figure 3. Comparison Drop Rate with different transmitting bit rates

Fig. 3 represents the CDR experienced by different α and bit rate. The drop rate reaches the minimum value when $\alpha = 0.5$ and maximizes at the two ends under the same bit rate. Note also that the CDR decreases slowly with increasing bit rate while α held constant. This is caused by the rising bit rate which leads to a heavier traffic load reducing the reliability of the routing decision.

According to the above analysis, while $\alpha = 0.5$, the packet drop rate under all bit rates reaches the lowest bounds, which can be selected as the well-performed values. This also indicates that the best ratio of current status and historical status is 0.5 of all tested α .

2) *Packet Drop Rate and Total Throughput*: For further performance evaluation, we use three other

routing algorithms [including Dijkstra's Shortest Path (Dijkstra), CEMR and PAR] as comparison. The result is shown in Fig. 4.

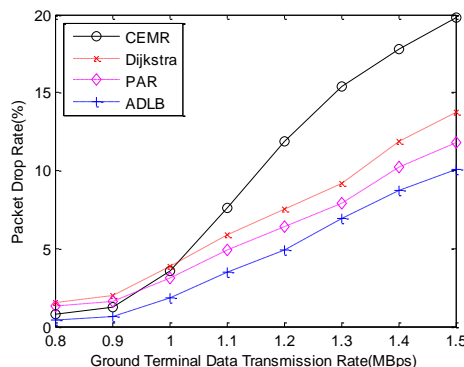


Figure 4. Packet drop rate with different transmitting bit rates

Fig. 4 clearly indicates that ADLB shows the best performance over the compared algorithms. This is because 1) ADLB calculates the transmission paths automatically from historical information which reduces the complexity of the algorithm and 2) the algorithm is able to respond quickly to the network topology change.

Compared to PAR, ADLB gets better performance because the ISLs historical information is fully considered. The Dijkstra finds only the shortest propagation delay paths without considering the load balancing. This results in more packet drops because of the overwhelmed buffers with increasing terminal transmission bit rate. It's also observed that the CEMR has the worst performance because the status messages are hardly exchanged under heavy traffic.

Fig. 5 shows the total throughput of the four routing mechanisms. In this test, ADLB also outperforms the other schemes in providing the highest throughput. The total throughput is found to be related to the packet drop rate with the same tendency.

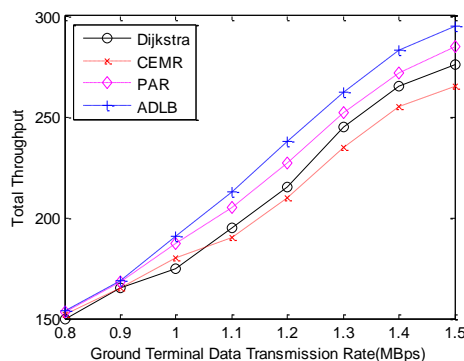


Figure 5. Total Throughput with different transmitting bit rates

V. CONCLUSIONS

In this paper, we propose a distributed routing mechanism which relies on the present and historical link status. A routing factor is applied to balance the two effects. The routing decision is determined in each satellite as well as the routing region is limited in the set of minimum hops. This method ensures the transmission

between the source node and destination node with the same number of hops.

To achieve higher performance, we present a series of simulations to determine the optimized parameters under different link loads. The simulation results with the best proportion between the current and the historical status demonstrate that the proposed algorithm guarantees decreased packet drop rate and increased maximum throughput.

The algorithm proposed is not only applicable to the Iridium-like constellation, but is also suitable for the other constellation with various ISLs of each satellite. By the same token, it can also be used in walker constellation of dynamic length ISL as the algorithm is based on hops. Our future work involves the application of this approach to the rosette constellation, and the corresponding performance evaluation.

REFERENCE

- [1] S. R. Pratt, R. A. Raines, C. E. Fossa JR., Temple, A. Michael, An Operational and Performance Overview of the Iridium Low Earth Orbit Satellite System. *IEEE Communications Surveys*, 1999, 2(2) pp. 1-10.
- [2] Y. C. Hubbel, A comparison of the IRIDIUM and AMPS systems. *IEEE Network Magazine*, 1997, 11(2) pp. 52-59.
- [3] M. Werner, A. Jahn, E. Lutz, A. Bottcher, Analysis of system parameters for LEO/ICO-satellite communication networks. *IEEE Journal on Selected Areas in Communications*, 1995, 13(2) pp. 371-381.
- [4] P. W. Lemme, S. M. Glenister, A. W. Mille, Iridium Aeronautical Satellite Communications, *IEEE Aerospace and Electronic Systems Magazine*, 1999, 14(11) pp. 11-16.
- [5] L. Wood, A. Clerget, I. Andrikopoulos, G. Pavlou, W. Dabbous, IP routing issues in satellite constellation networks, *International Journal of Satellite Communications (Special Issue on IP)*, 2001, 19(1) pp. 69-92.
- [6] L. Wood L, Internet working with satellite constellations. *University of Surrey, Guildford, United Kingdom*, 2001.
- [7] E. Ekici, I. F. Akyildiz, M. D. Bender, A Distributed Routing Algorithm for Datagram Traffic in LEO Satellite Networks, *IEEE/ACM Transaction on Networking*, 2001, 9(2) pp. 137-147.
- [8] J. Bai, X. Lu, Z. Lu, W. Peng, Compact explicit multi-path routing for LEO satellite networks, *Proceedings of IEEE International Workshop on High Performance Switching and Routing (HPSR2005)*, May 12-14, 2005, Hong Kong, China: IEEE, 2005, pp. 386-390.
- [9] T. Taleb, D. Mashimo, A. Jamalipour, ELB: An Explicit Load Balancing Routing Protocol for Multi-Hop N GEO Satellite Constellations. *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM'06), San Francisco (Globecom)*, Nov. 27-Dec. 1, 2006, California, USA, IEEE: 2006, 1-5.
- [10] T. Taleb, D. Mashimo, A. Jamalipour, Explicit Load Balancing Technique for N GEO Satellite IP Networks with On-Board Processing Capabilities. *IEEE/ACM Transactions on Networking*, 2009, 17(1), pp. 281-293.
- [11] E. Papapetrou, S. Karapantazis, F. N. Pavlidou, Distributed on-demand routing for LEO satellitesystems. *Computer Networks*, 2007, 51(15) pp. 4356-4376.
- [12] Ö. Korçak, F. Alagöz, A. J. amalipour, Priority-Based Adaptive Shortest Path Routing for IP over Satellite Networks, *International Journal of Communication Systems*, 2007, 20(3) pp. 313-333.
- [13] Z. Yu, H. Zhou, Z. Wu. A trust-based secure routing protocol for multi-layered satellite networks, *Information Science and Technology (ICIST)*, March 23-25, 2012, Hubei, China: IEEE, 2012, pp. 313-317.
- [14] H. Nishiyama, D. Kudoh, N. Kato, N. Kadowaki, Load Balancing and QoS Provisioning Based on Congestion Prediction for GEO/LEO Hybrid Satellite Networks, *Proceedings of the IEEE*, 2011, 99(11) pp. 1998-2007.
- [15] H. Nishiyama, Y. Tada, N. Kato, N. Yoshimura, M. Toyoshima, N. Kadowaki, Toward Optimized Traffic Distribution for Efficient Network Capacity Utilization in Two-Layered Satellite Networks, *IEEE Transactions on Vehicular Technology*, 2013, 62(3) pp. 1303-1313.

Xiao Ma, Ph.D. candidate with the Department of Communication Engineering, School of Computer and Telecommunication Engineering, University of Science and Technology Beijing. His research interests include space communications, satellite routing protocol. Email: mxustb@hotmail.com.

A Noise-Correlated Cancellation Transmission Scheme for Cooperative MIMO Ad Hoc Networks

Wanni Liu

Department of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China
Email: 18234085119@163.com

Long Zhang

Department of Mathematics, Taiyuan University of Technology, Taiyuan 030024, China
Email: 534468526@qq.com

Yanping Li*

Department of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China
*Corresponding author, Email: liyanping@tyut.edu.cn

Abstract—A new transmission scheme based on noise-correlated cancellation (NCC) is proposed, which absorbs the advantages of phase-inversion symmetric method and cooperative MIMO technology and makes full use of the correlation of noise in the adjacent channels to reduce channel noise. This paper firstly presents the implementation process of NCC transmission scheme in detail. Further, through theoretical analysis, it is showed that the signal-to-noise ratio gain which the proposed NCC transmission scheme gets is at least 4 times greater than the signal-to-noise ratio gain which the traditional cooperative MIMO transmission scheme gets. Finally, simulation experiment results also verify that the proposed NCC transmission scheme can make the channel capacity per bandwidth of cooperative MIMO Ad Hoc networks improve significantly and bit error rate (BER) of the network reduce greatly, which will help to expand application scopes of cooperative MIMO Ad Hoc networks.

Index Terms—Mobile Ad Hoc Networks, Cooperative Multi-Input Multi-Output, Phase-Inversion Symmetric Method, Anti-Noise Performance

I. INTRODUCTION

With the continuous development of computer and communication technology, there will be a new situation that people can communicate with each other anytime. Therefore, mobile Ad Hoc networks which could be built without relying on base stations have been extensively studied by scholars. A mobile Ad Hoc network (MANET) is a multi-hop and temporary peer-to-peer network whose nodes act as terminals and routers. The control and management of a MANET are distributed to the terminals, so the terminals involved work together to communicate and exchange information (such as voice, image, video, data, etc) via a wireless link [1, 2, 3]. Since wireless links are relatively weak, bandwidth is relatively limited and there is no support of base stations, signal transmission in

mobile Ad Hoc networks is easy to be interfered by channel noises. Therefore, mobile Ad Hoc networks have relatively poor channel transmission performance which is not conducive to its large-scale practical application.

Cover and Gamal put forward to the term of “Cooperative Communication” in 1979 [4] whose main idea is that neighboring mobile users share each other’s antenna to cooperatively transmit signals in the multi-users’ communication environment, thus generating a virtual environment similar to multi-antenna transmission environment and obtaining spatial diversity gain and improving the transmission performance of the network. The research of Sendonaris [5] showed that the multi-node cooperative transmission mechanism could greatly increase the network capacity and effectively resist fading effects of the wireless channels.

Ref. [6-9] introduced cooperative multi-input multi-output (MIMO) technique into Ad Hoc networks, which showed that the virtual cooperative transmission networks with omni-directional single antenna mobile users could overcome the limits of traditional MIMO transmission networks, sending-receiving diversity gain and array gain got could significantly increase the channel capacity, reduce the end-to-end transmission delay, the energy consumption of data transmission and output bit error rate (BER).

This paper presents a new transmission scheme based on noise-correlated cancellation (NCC) for cooperative MIMO Ad Hoc networks, which absorbs the advantages of phase-inversion symmetric method [10] and cooperative MIMO technology. The scheme not only makes full use of the correlation of noise in the adjacent channels in order to achieve the purpose of resisting channel noise interference but also combines with cooperative MIMO technology in order to achieve the

purpose of enhancing the whole network transmission performance.

II. NETWORK MODEL

As shown in Fig. 1, a certain number of mobile terminals adjacent in position are considered as a cluster. Each cluster has a head responsible for the management of this cluster. However, just as Feng et al. proposed in [6], each cluster is called as a virtual node (VN), the cluster head as a kernel node (KN) and the remaining mobile terminals as team nodes (TNs). And all the actual links corresponding between two VNs are set into a virtual link (VL). In addition, if TNs in two VNs can inter-exchange, they are known as adjacent VNs which can form a virtual backbone network via the VL. Now assuming that any mobile terminals in a virtual backbone network can send and receive signals synchronously, all channel noises are additive white Gaussian noises (AWGN).

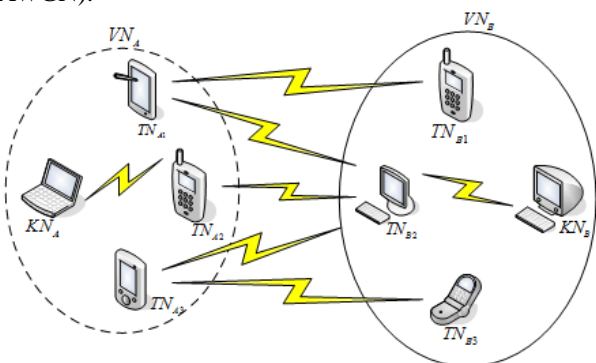


Figure 1. Network model

III. NCC TRANSMISSION SCHEME

A virtual link which connects two adjacent virtual nodes VN_A and VN_B is built in the network model shown in Fig. 1. Then an optimal cooperative terminals selection algorithm is used to select a set of 2×2 cooperative terminals for forming a cooperative MIMO wireless network, and a routing selection algorithm is used to search for the optimal path from the source terminal to the destination terminal. Now assuming that a source terminal A_s in VN_A sends information to a destination terminal B_r in VN_B and A_1, A_2, B_1, B_2 are selected as optimal cooperative terminals. The implementation process is shown in Fig. 2.

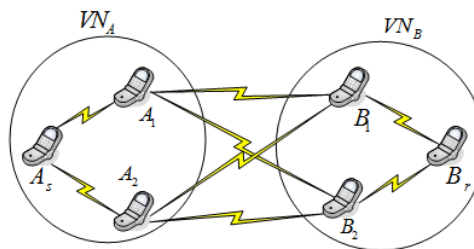


Figure 2. NCC transmission scheme

The source terminal A_s synchronously sends the original data to one cooperative sending terminal A_1 and the data which is reversed phase to another cooperative sending terminal A_2 . After modulating the received signals respectively, A_1 and A_2 send them to the cooperative receiving terminals B_1 and B_2 , thus B_1 and B_2 receive a modulated original data and a modulated inverted one respectively. Next in B_1 and B_2 the signals are subtracted in a sub-tractor after passing through a band-pass filter and a demodulator. Finally, the data is sent to the destination terminal B_r after combining processing in the receiving end.

In the implementation process, the cooperative sending terminals A_1 and A_2 respectively send the data to two cooperative receiving terminals B_1 and B_2 via two independent channels. Here we must claim that the distance between two cooperative sending/receiving antennas must be $\leq \lambda/2$ (here λ is radio wavelength), even in order to use the spatial noise correlation, the distance between the two antennas is as adjacent as possible.

IV. MATHEMATICAL MODEL AND THEORETICAL ANALYSIS OF NCC

The mathematical model of NCC transmission scheme is shown in Fig. 3. The source terminal A_s sends the signal $s(t)$, and two signals $s_1(t)$ and $s_2(t)$ are respectively received by two cooperative sending terminals A_1 and A_2 ($s(t) = s_1(t) = -s_2(t)$), then sent after modulated by two modulators who use different carriers frequencies (f_1, f_2).

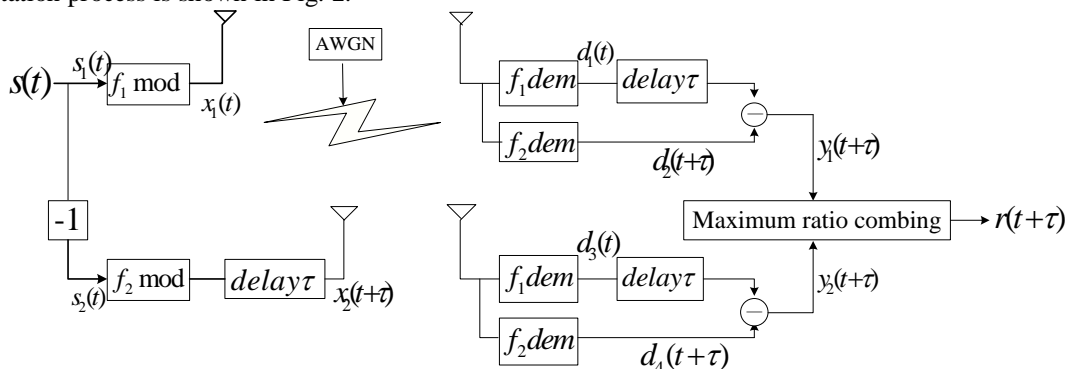


Figure 3. The mathematical model of NCC transmission scheme

Two cooperative receiving terminals B_1 and B_2 respectively receive and demodulate an original signal and a signal after reversed phase.

The modulated signals in B_1 are shown as

$$d_1(t) = h_{11}s_1(t) + n_{11}(t). \quad (1)$$

$$d_2(t + \tau) = h_{12}s_2(t + \tau) + n_{12}(t + \tau). \quad (2)$$

The modulated signals in B_2 are shown as

$$d_3(t) = h_{21}s_1(t) + n_{21}(t). \quad (3)$$

$$d_4(t + \tau) = h_{22}s_2(t + \tau) + n_{22}(t + \tau). \quad (4)$$

where $H = \begin{bmatrix} h_{11}(x) & h_{12}(x) \\ h_{21}(x) & h_{22}(x) \end{bmatrix}$ is a channel matrix,

$h_{ji}(x)$ is the channel fading coefficient from the sending antenna i to the receiving antenna j ($i=1,2; j=1,2$),

$N_{ji} = \begin{bmatrix} n_{11}(t) & n_{12}(t) \\ n_{21}(t) & n_{22}(t) \end{bmatrix}$ is additive white Gaussian noise(AWGN) in the independent channel of the sending antenna i to the receiving antenna j at time t , τ is delay time.

Two demodulated signals in the cooperative receiving terminal B_1 are sent into a sub-tractor, so are two demodulated signals in the cooperative receiving terminal B_2 . The two output signals are:

$$y_1(t + \tau) = d_1(t + \tau) - d_2(t + \tau) \quad (5)$$

$$= h_{11}s_1(t + \tau) - h_{12}s_2(t + \tau) + n_{11}(t + \tau) - n_{12}(t + \tau).$$

$$y_2(t + \tau) = d_3(t + \tau) - d_4(t + \tau) \quad (6)$$

$$= h_{21}s_1(t + \tau) - h_{22}s_2(t + \tau) + n_{21}(t + \tau) - n_{22}(t + \tau).$$

$$s(t) = s_1(t) = -s_2(t), \text{ so } s(t + \tau) = s_1(t + \tau) = -s_2(t + \tau).$$

Thus $y_1(t + \tau)$ and $y_2(t + \tau)$ can be simplified as:

$$y_1(t + \tau) = (h_{11} + h_{12})s(t + \tau) + n_{11}(t + \tau) - n_{12}(t + \tau). \quad (7)$$

$$y_2(t + \tau) = (h_{21} + h_{22})s(t + \tau) + n_{21}(t + \tau) - n_{22}(t + \tau). \quad (8)$$

As we have already stated that the distance between two cooperative sending/receiving antennas must be $\leq \lambda/2$ (here λ is radio wavelength), so can set $h_1 = h_{11} = h_{21}$, $h_2 = h_{12} = h_{22}$ in order to facilitate the analysis. Thus:

$$y_1(t + \tau) = (h_1 + h_2)s(t + \tau) + n_{11}(t + \tau) - n_{12}(t + \tau). \quad (9)$$

$$y_2(t + \tau) = (h_1 + h_2)s(t + \tau) + n_{21}(t + \tau) - n_{22}(t + \tau). \quad (10)$$

Assuming that the channel state information (CSI) in the receiving end has been known, the final output signal gained through using receiving combining technology is

$$r(t + \tau) = \alpha_1(t + \tau)y_1(t + \tau) + \alpha_2(t + \tau)y_2(t + \tau) \quad (11)$$

$$= [\alpha_1(t + \tau) + \alpha_2(t + \tau)](h_1 + h_2)s(t + \tau) + \alpha_1(t + \tau)[n_{11}(t + \tau) - n_{12}(t + \tau)] + \alpha_2(t + \tau)[n_{21}(t + \tau) - n_{22}(t + \tau)].$$

Ref. [11] showed that the performance of maximal ratio combining (MRC) technique is the best in three typical receiving combining techniques (maximal ratio combining, equal gain combining and selection combining), so MRC is used here. Set $\alpha_1(t + \tau) = \beta h_1$ and $\alpha_2(t + \tau) = \beta h_2$, then the final output total signal can be written as:

$$r(t + \tau) = \beta(h_1 + h_2)^2 s(t + \tau) + \beta h_1 [n_{11}(t + \tau) - n_{12}(t + \tau)] + \beta h_2 [n_{21}(t + \tau) - n_{22}(t + \tau)] = s_o(t + \tau) + n_o(t + \tau). \quad (12)$$

where $s_o(t + \tau)$ is the final output signal including the whole sending information, and $n_o(t + \tau)$ is the final output noise signal.

Assuming that S_o represents the power of $s_o(t + \tau)$, and N_o represents the power of $n_o(t + \tau)$. Now discussing the output signal-to-noise ratio separately in two different environments:

(1) First, we consider an absolutely ideal case. Of course it can not be really achieved. Assuming that the noise in the different independent channels of the different sending antennas i to the same receiving antenna j is the same absolutely ($n_{11}(t + \tau) = n_{12}(t + \tau)$, $n_{21}(t + \tau) = n_{22}(t + \tau)$), thus getting $r(t + \tau) = \beta(h_1 + h_2)^2 s(t + \tau) = s_o(t + \tau)$. However $n_o(t + \tau) = 0$, so $S_o/N_o = E[|s_o(t + \tau)|^2] / E[|n_o(t + \tau)|^2] \rightarrow +\infty$. It means that the output noise is fully offset so that the signal is transmitted without interference.

(2) The noise in real adjacent channels is correlated strongly. Ref. [12] has proved that the correlation coefficient ρ of two adjacent channel noises is up to 0.8 or more through actual measurement when the physical distance of two adjacent channels is less than three meters. Thereby, the output noise of the different independent channels of the different sending antennas i to the same receiving antenna j is extremely similar. Thus the output noise of the sub-tractor in the cooperative receiving terminal B_1 is $n_1(t + \tau) = n_{11}(t + \tau) - n_{12}(t + \tau)$, $n_1(t + \tau) < n_{12}(t + \tau) < n_{11}(t + \tau)$ and that of the sub-tractor in the cooperative receiving terminal B_2 is $n_2(t + \tau) = n_{21}(t + \tau) - n_{22}(t + \tau)$, $n_2(t + \tau) < n_{22}(t + \tau) < n_{21}(t + \tau)$.

So $r(t + \tau) = \beta(h_1 + h_2)^2 s(t + \tau) + \beta h_1 n_1(t + \tau) + \beta h_2 n_2(t + \tau) = s_o(t + \tau) + n_o(t + \tau)$. Thus,

the total output signal-to-noise ratio (SNR) can be written as:

$$\begin{aligned}
 S_o/N_o &= E\left[|s_o(t+\tau)|^2\right]/E\left[|n_o(t+\tau)|^2\right] \\
 &= E\left[|\beta(h_1+h_2)s(t+\tau)|^2\right]/E\left[|\beta h_1 n_1(t+\tau)+\beta h_2 n_2(t+\tau)|^2\right] \quad (13) \\
 &= |\beta(h_1+h_2)|^2 E\left[|s(t+\tau)|^2\right]/|\beta|^2 E\left[|h_1 n_1(t+\tau)+h_2 n_2(t+\tau)|^2\right].
 \end{aligned}$$

Set $h = h_1 = h_2$, then:

$$\begin{aligned}
 S_o/N_o &= |4\beta h^2|^2 E\left[|s(t+\tau)|^2\right]/|\beta h|^2 E\left[|n_1(t+\tau)+n_2(t+\tau)|^2\right] \\
 &= 16|h|^2 S/E\left[|n_1(t+\tau)|^2+|n_2(t+\tau)|^2+2|n_1(t+\tau)||n_2(t+\tau)|\right] \\
 &= (16|h|^2 S)/\{N_1+N_2+2E\left[|n_1(t+\tau)||n_2(t+\tau)|\right]\} \\
 &= (16|h|^2 S)/(N_1+N_2+2\rho\sqrt{N_1N_2}). \quad (14)
 \end{aligned}$$

where S represents the total power $E\left[|s(t+\tau)|^2\right]$ of signal in the sending end, N_1 represents the power $E\left[|n_1(t+\tau)|^2\right]$ of the output noise $n_1(t+\tau)$ of the sub-tractor in B_1 , N_2 represents the power $E\left[|n_2(t+\tau)|^2\right]$ of the output noise $n_2(t+\tau)$ of the sub-tractor in B_2 , and $\rho = E\left[|n_1(t+\tau)n_2(t+\tau)|\right]/\sqrt{N_1N_2}$ represents the correlation coefficient of the noise received by the two cooperative receiving terminals.

In order to compare with traditional cooperative MIMO system, considering $h=1$ which means each channel has no attenuation and $N=N_1=N_2$ which means the noise in the output ends of the two sub-tractors is the same. Get $N < N_i$ (N_i is the noise power in the channel i) from $n_i(t+\tau) < n_{i2}(t+\tau) < n_{i1}(t+\tau)$ and $n_2(t+\tau) < n_{22}(t+\tau) < n_{21}(t+\tau)$. Thus the output signal-to-noise ratio (SNR) can be further transformed into:

$$S_o/N_o = (16S)/[2(1+\rho)N] > \frac{8}{(1+\rho)} \cdot \frac{S}{N_i}. \quad (15)$$

Finally, the SNR gain of NCC transmission scheme is:

$$G_{NCC} = (S_o/N_o)/(S/N_i) > \frac{8}{1+\rho}. \quad (16)$$

The SNR gain of the traditional cooperative MIMO system is [12]:

$$G_{CMIMO} = 2/(1+\rho). \quad (17)$$

Therefore,

$$G_{NCC} / G_{CMIMO} > 4. \quad (18)$$

Equation (18) shows that the SNR gain which the proposed NCC transmission scheme gets is at least 4 times greater than the SNR gain which the traditional cooperative MIMO transmission scheme gets.

V. SIMULATION AND ANALYSIS OF NCC

A. Analysis of the SNR Gain of NCC

According to Equation (16) and Equation (17), using MATLAB software for simulating NCC transmission scheme. Compared with the SNR gain of traditional cooperative MIMO (CMIMO) system, the simulation result is shown in Fig. 4.

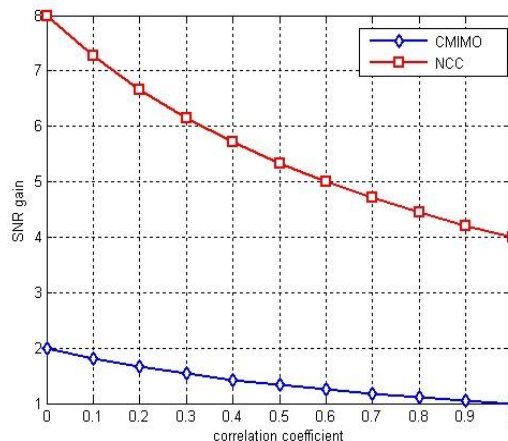


Figure 4. The comparison of SNR gains of NCC and CMIMO

In Fig. 4, the x-axis represents the correlation coefficient ρ of noise in the two adjacent channels which is increasing with decreasing of the spacing between the mobile terminal antennas. Through actually measuring, the correlation coefficient ρ of noise in adjacent channels can be up to 0.8 or more if the spacing between the mobile terminal antennas is less than three meters [12]. The y-axis represents the SNR gain of different systems. The simulation result shows that the SNR gain of cooperative MIMO system based on NCC transmission scheme is much better than that of traditional cooperative MIMO system which can contribute to boosting the anti-noise performance and signal transmission quality of the whole cooperative MIMO Ad Hoc networks.

Further, due to the increase of S/N , the channel capacity $C = B \log_2(1+(S/N))$ [13] would be improved in the case of the same channel bandwidth.

B. Analysis of Channel Capacity of NCC

C. E. Shannon gave the well-known Shannon formula [13]:

$$C = B \log_2(1+S/N). \quad (19)$$

where B is the channel bandwidth, S is the signal power and N is the noise power.

In order to compare with the traditional $C = lb \det(I_2 + \frac{4A}{2}Q)$ cooperative MIMO system, set $SNR = S/N$ and $SNR = 10 \log_{10} A$, then can get $C = \log_2 \det(I_2 + \frac{A}{2}Q)$ which is used in MATLAB simulation. Thus, the channel capacity per bandwidth of

Single-Input Single-Output (SISO) system can be changed into:

$$C = \log_2(1 + A). \tag{20}$$

The channel capacity per bandwidth of traditional cooperative MIMO system is the same as that of distributed MIMO system, so it is [14, 15]:

$$C = \log_2 \det(I_2 + \frac{A}{2} Q). \tag{21}$$

Because the SNR gain of NCC transmission scheme is at least 4 times greater than that of traditional cooperative MIMO system when $\rho > 0.8$, the output SNR of NCC transmission scheme also is at least 4 times greater than that of traditional cooperative MIMO system after normalizing the input SNRs of two systems. Therefore the channel capacity per bandwidth of 2×2 cooperative MIMO system based on NCC transmission scheme is:

$$C = lb \det(I_2 + \frac{4A}{2} Q). \tag{22}$$

According to Equation (20), Equation (21) and Equation (22), the channel capacities of SISO system, 2×2 traditional cooperative MIMO system and 2×2 cooperative MIMO system based on NCC transmission scheme are simulated and analyzed respectively. The result is shown in Fig. 5.

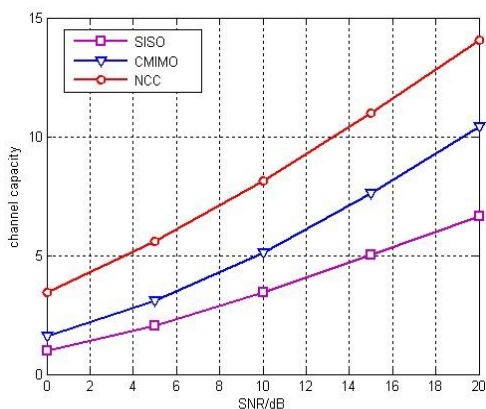


Figure 5. Relationship between the channel capacity and SNR of three systems

Fig. 5 shows that 2×2 cooperative MIMO system based on NCC transmission scheme has relatively the largest channel capacity. That is because NCC transmission scheme takes full advantages of the correlation of noise in adjacent channels to partly offset noise interference so that the SNR gain of the whole system is improved and the channel capacity of the system also is increased in the case of the same channel bandwidth.

The characteristics of cooperative MIMO Ad Hoc networks decides its limited channel bandwidth, therefore it is not suitable to increase channel capacity by increasing system bandwidth while the way of increasing SNR is relatively well. Above all, the NCC transmission scheme is feasible for cooperative MIMO Ad Hoc networks.

C. Analysis of BER of NCC

Using Monte Carlo method for simulating the real signal transmission processes of the 2×2 cooperative MIMO system based on NCC transmission scheme, the 2×2 cooperative MIMO system based on space-time block coding (STBC), the 2×1 cooperative MISO system based on Alamouti coding and SISO system respectively. Function randn() is used to generate a Gaussian random channel whose mean score is 0 and variance is 1. In the channel, there is AWGN whose mean score is zero and power spectral density is $N_0 / 2$. Besides, BPSK mapping method is used in simulation, normalizing all the receiving signals. Finally the maximum likelihood ratio criterion is used to recover the original sending signal at the receiving end of each system. Fig. 6 presents that the relationship between BER and the output SNR of each system.

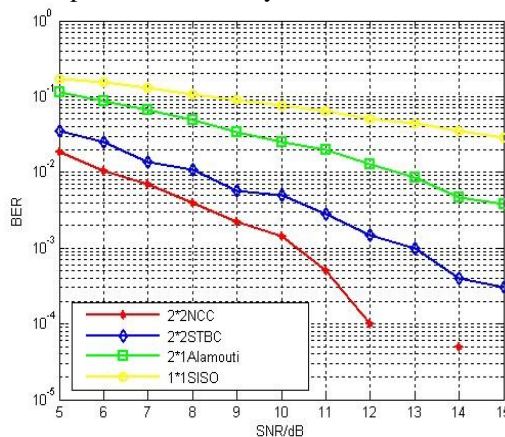


Figure 6. Relationship between BER and output SNR of NCC, STBC, Alamouti and SISO

Fig. 6 shows that the BERs of four systems (NCC, STBC, Alamouti and SISO) are all declining with the increase of the output SNR of the corresponding system. The reason is that the transmission performance will be improved with the increasing of the output SNR, thus the probability of transmitting error code will be decreased.

Vertical comparison of the four curves in Fig. 6 shows that the BER of the 2×2 cooperative MIMO system based on NCC transmission scheme is the lowest, and that of SISO transmission system is the relatively highest. The signal transmission qualities of the 2×2 cooperative MIMO system based on NCC transmission scheme and the 2×2 cooperative MIMO system based on space-time block coding (STBC) all are better than that of the 2×1 cooperative MISO system based on Alamouti coding and SISO system. That is because the Alamouti coding system only adopts transmitting diversity technique but the STBC coding system adopts transmitting diversity and receiving combining techniques, and the NCC system also adopts noise-correlated cancellation method except for transmitting diversity and receiving combining techniques.

VI. CONCLUSIONS

The traditional cooperative MIMO system claims the spacing between the mobile terminal antennas should be large enough in order to avoid the correlation of the noise in different channels, restricting its application in cooperative MIMO Ad Hoc networks. This paper proposes a transmission mechanism based on noise-correlated cancellation method (NCC) in which the correlation between noise in the adjacent channels is utilized as a favorable resource so that the minimum distance of the mobile terminal antennas is not limited. In addition, theoretical analysis and system simulation results show that NCC transmission scheme enables cooperative Ad Hoc networks to obtain greater output SNR, improve the channel capacity per bandwidth, and reduce the BER of the system, thus overcoming its shortcomings of self-limitation and big noise, and improving the transmission performance of cooperative MIMO Ad Hoc networks.

ACKNOWLEDGEMENT

This work was supported in part by a grant from the National Natural Science Foundation Project of China (No. 61271249).

REFERENCES

- [1] L. X. Chen, X. Zeng and Y. Cao, *Mobile Ad Hoc networks--Self-organizing Packet Radio Network Technology*, 2nd ed. CA: Beijing, 2012. (in Chinese)
- [2] M. Natkaniec, K. K. Szott, S. Szott and G. Bianchi, "A Survey of Medium Access Mechanisms for Providing QoS in Ad-Hoc Networks," *IEEE Communications Surveys & Tutorials*, PP (99) (2012) 1-29. doi: 10.1109/SURV.2012.060912.00004
- [3] S. Sesay, Z. K. Yang and J. H. He, "A Survey on Mobile Ad Hoc Wireless Network," *Information Technology Journal*, 3 (2) (2004) 168-175.
- [4] T. M. Cover and A. A. EL Gamal, "Capacity Theorems for the Relay Channel," *IEEE Trans. on Information Theory*, 25 (5) (1979) 572-584. doi: 10.1109/TIT.1979.1056084
- [5] A. Sendonaris, E. Erkip and B. Aazhang, "User cooperation diversity, part II: implementation aspects and performance analysis," *IEEE Trans. on Commun.* 51 (11) (2003) 1939-1948. doi: 10.1109/TCOMM.2003.819238
- [6] W. J. Feng, W. Zhao and D. Wang, "Cooperative MIMO transmission scheme for clustered Ad Hoc networks," *Journal on Communications*, 33 (3) (2012) 1-9. (in Chinese)
- [7] S. Chu, X. Wang and Y. Y. Yang, "Exploiting Cooperative Relay for High Performance Communications in MIMO Ad Hoc Networks," *IEEE Transactions on Computers*, 62 (4) (2013) 716-729. doi: 10.1109/TC.2012.23
- [8] J. Park and S. Lee, "Distributed MIMO Ad-hoc Networks: Link Scheduling, Power Allocation, and Cooperative Beamforming," *IEEE Transactions on Vehicular Technology*, 61 (6) (2012) 2586-2598. doi: 10.1109/TVT.2012.2198505
- [9] J. Hwang, T. Kim, J. So and H. Lim, "A receiver-centric multi-channel MAC protocol for wireless networks," *Elsevier Computer Communications*, 36 (4) (2013) 431-444. doi: http://dx.doi.org/10.1016/j.comcom.2012.11.006

- [10] Y. Z. Xiao, F. C. Ma, B. J. Xiao and X. F. Han, "Phase-Inversion Symmetric Method Principle and Application," presented at the 2011 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), July 26-30, 2011. doi: 10.1109/CSQRWC.2011.6037103
- [11] T. Luo, *Multi-antenna Wireless Communication Principles and Applications*, CA: Beijing, 2005.
- [12] B. J. Xiao and L. L. Hao, "Space Diversity Technology Based on Phase-Inversion Symmetric Method," *Journal of Taiyuan University of Technology*, 41(3) (2010) 245-247.
- [13] C. E. Shannon, "A Mathematical Theory of Communication," *the Bell System Technical Journal*, 27 (1948) 379-423, 623-656.
- [14] H. C. Chen, D. Z. Wu and X. Q. Gao, *MATLAB and its application in Electronic Information Courses*, 3rd ed. CA: Beijing, 2006. (in Chinese)
- [15] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, 6 (1998) 311-335.



Wanni Liu was born in Yuncheng City, Shanxi Province, China in 1989 and received the B.S. degree from Department of Physics and Electronic Engineering, Mudanjiang Normal University, Mudanjiang City, Heilongjiang Province, China in 2011. She is currently working toward the M.S. degree with the Institute of Information Engineering, Taiyuan University of Technology, Shanxi Province, China.

Her research interests include cooperative communications and mobile Ad Hoc networks.



Long Zhang was born in Taiyuan City, Shanxi Province, China in 1988 and received the B.S. degree from Department of Mathematics, Taiyuan University of Technology, Taiyuan City, Shanxi Province, China in 2011. He is currently working toward the M.S. degree with the Institute of Mathematics, Taiyuan University of Technology, Shanxi Province, China.

Her research interests include the application of mathematics in engineering.



Yanping Li was born in Taiyuan City, Shanxi Province, China in 1963 and received the B.S., M.S, and Ph.D. degrees with the Institute of Information Engineering, Taiyuan University of Technology, Taiyuan City, Shanxi Province, China.

She is currently a Professor with the Institute of Information Engineering, Taiyuan University of Technology, Shanxi, China. She has directed more than 50 graduate students and widely published paper in signal processing for communications and wireless networks. Her research interests include wireless communications, bandwidth communications and signal processing.

CluLoR: Clustered Localized Routing for FiWi Networks

Yousef Dashti and Martin Reisslein

School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, Arizona 85287-5706, USA
Email: {Yousef.Dashti, reisslein}@asu.edu

Abstract—The integration of passive optical networks (PONs) and wireless mesh networks (WMNs) into Fiber-Wireless (FiWi) networks can lead to effective access networks. Existing routing schemes for FiWi networks consider mainly hop-count and delay metrics over a flat WMN node topology and do not specifically prioritize the local network structure, i.e., the local wireless-optical network gateway. In this study, we explore a simple, yet effective routing algorithm for FiWi networks with a WMN organized into zones operating on different radio channels. We examine the effects of routing the traffic into and out of a zone through one or more cluster heads. We investigate the effectiveness of localized routing that prioritizes transmissions over the local gateway to the optical network and avoids wireless packet transmissions in zones that do not contain the packet source or destination. We find that this combination of clustered and localized routing (CluLoR) gives good throughput-delay performance compared to routing schemes that transmit packets wirelessly through “transit zones” (that do not contain the packet source or destination) following minimum hop-count routing.

Index Terms—Cluster heads, delay-sensitive traffic, fiber-wireless (FiWi) network, localized routing.

I. INTRODUCTION

Wireless and optical networking technologies at the early stages were deployed for different respective communication settings. Due to the fact that those technologies aim to solve different problems when they were initially developed, it is hard for one given technology to overcome many of the challenges arising in the access network area. The merging of optical access technologies with wireless access technologies by capitalizing on their respective advantages could lead to powerful solutions. Passive optical networks (PONs) connect several distributed optical network units (ONUs) at subscriber premises with a central optical line terminal (OLT) at high bandwidth of up to 10 Gbps [1]–[3] with reach extending over long distances [4]–[9]. We note that a plethora of studies has examined related TDM/WDM PONs, see e.g., [10]–[17]; however, they have high deployment costs. On the other hand, wireless mesh networks (WMNs) offer flexible communication and eliminate the need for a fiber drop to every user in the network, but offer only relatively low bandwidth, which is impacted by interference among ongoing wireless transmissions [18]–[28]. Fiber-wireless (FiWi) architectures that combine optical and

wireless network technologies could lower access network deployment cost while providing high bandwidth to the end users [29].

In this paper, we focus on the problem of peer-to-peer communication within a given wireless mesh network (WMN). Integrating an optical access network with the wireless mesh network could possibly lead to higher throughput and lower end-to-end packet delays. Without an optical access network, all traffic has to go through the WMN, which results in high network interference and in turn limits the network throughput. By combining the optical access network and the WMN to an integrated fiber and wireless (FiWi) network, the traffic could be routed from the source node in the WMN over wireless hops to a nearby gateway wireless router where it could be routed via the fiber network to a gateway wireless router near the destination node. This scenario would reduce interference in the wireless mesh network, and increase throughput between the two communicating peers.

As elaborated in Section II, many FiWi network architectures and routing protocols have been explored in the past few years [30]. To the best of our knowledge, the existing FiWi routing approaches mainly consider a “flat” topology for the WMN, i.e., the existing approaches do not consider a hierarchical clustering structure of the WMN nodes. Moreover, the specific local network structure, i.e., the closest local gateway from the WMN to the PON, has not been prioritized over multi-hop transmissions through the WMN. Clustering has proven very beneficial in purely wireless networks [31], [32]. In this article, we examine the combined effects of clustered localized routing. We consider a common WMN setting where the wireless nodes are organized into zones that operate on different radio channels [33]–[35]. We allow wireless nodes to send traffic to each other directly only when they are in the same zone. Otherwise, all traffic has to go through an assigned cluster head which in turn routes the traffic to the assigned gateway router (which in turn routes the traffic to the destination zone, possibly utilizing the optical network).

The remainder of this paper structured as follows. In Section II, we discuss the related work and recent research on FiWi networks. In Section III, we introduce the principles of clustered localized routing (CluLoR). In Section IV, we describe the simulation set-up for our evaluations of CluLoR. In Section V, we examine the effects of clustering by varying the number of cluster

heads in a zone and adding relay routers between adjacent zones. In Section VI, we examine the effects of the localized routing strategy by comparing CluLoR with an unlocalized routing benchmark that follows minimum hop-count routing. In Section VII, we evaluate how CluLoR behaves when the PON is stressed with background traffic. Section VIII concludes the paper and points out future research directions.

II. RELATED WORK

The recent survey [36] gives an overview of hybrid optical-wireless access networks. The Hybrid Wireless-Optical Broadband-Access Network (WOBAN) Architecture [37] is a pioneering FiWi network structure. The study [37] identified FiWi networking challenges with regard to network setup (placement of ONUs, Base Stations (BSs), and OLT to minimize the cost), and efficient routing protocols. The FiWi network planning problem has been further studied in [38]. The studies [39]–[41] proposed FiWi architectures and reconfiguration algorithms in order to serve the needs of the hybrid access network users.

Some of the first studies that examined peer-to-peer communication in a FiWi network were by Zheng et al. [42], [43]. These studies noted the significance of integrating the optical networks with the mesh networks to achieve significant performance improvements in terms of overall throughput and average packet end-to-end delays. Also, a simple routing protocol was proposed based on minimum-hop-count, which includes the gateway routers to the fiber network as part of the hop count. Li et al. [44] also studied the problem of peer-to-peer communications. The main focus was on implementing a novel arrayed waveguide grating based WDM/TDM PON structure, including wavelength assignment for groups of ONUs and a decentralized dynamic bandwidth allocation (DBA) algorithm, that supports direct communication between the ONUs without the traffic going through the OLT which could lead to improved end-to-end delay and throughput. Similarly, studies [45], [46] focused on inter-ONU communications by deploying a star coupler (SC) at the remote node (RN) to broadcast the packets of one ONU to all other ONUs, while [47] focused on the medium access control problem in radio-over-fiber networks. A WDM EPON that supports inter-ONU communications in which the polling cycle is divided into two sub-cycles was proposed in [48]. In this study, which is focused on FiWi routing, we consider a TDM PON with interleaved polling with adaptive cycle time (IPACT) with gated service dynamic bandwidth allocation [49], [50].

Routing protocols and algorithms for FiWi access networks have been the main focus of several studies, whereby some focus on routing the packets in the wireless front-end only, or routing the packets through the wireless and optical domains combined to achieve better performance. Early work that focused on routing algorithms in FiWi access networks includes the Delay-Aware Routing Algorithm (DARA) [51], Delay-Differentiated Rout-

ing Algorithm [52], Capacity-and-Delay-Aware Routing (CaDAR) [53], and Risk-And-Delay-Aware Routing (RADAR) [54]. Other recent studies on routing techniques in hybrid wireless-optical access network have focused on energy efficient routing [55], and Availability-Aware routing [56] as well as analytical frameworks for capacity and delay evaluation [57]. Most of these studies approach the routing as an optimization problem in order to find the optimum solution. However, all of them considered a flat topology, without a cluster structure, in the WMN. In contrast, this study focuses on the effects of clustered localized routing in the WMN on FiWi network performance.

A number of other studies have focused mainly on load balancing and Transmission Control Protocol (TCP) related issues in FiWi networks. Shaw et al. [58] proposed an integrated routing algorithm that adapts to the changes of the traffic demands within different regions of the wireless network in order to achieve load balancing in the hybrid network. The route assignment is located in the central hub. A hybrid TDM/WDM network with a wavelength assignment scheme that focuses on assigning a minimum number of wavelength to each group of ONUs while the maximum throughput at the ONUs is maintained was examined in [59]. The performance of multipath routing in FiWi and its effect on TCP performance due to out-of-order packets at the destination node was analyzed in [60], [61]. An integrated flow assignment and packet re-sequencing approach that obtains the probabilities of sending along the different paths with the objective of reducing the arrived out-of-order packets at the OLT was explored in [60]. A DBA technique that gives higher priority to the flows that trigger TCP fast retransmissions was proposed in [61]. We do not specifically examine TCP traffic; instead, we focus on traffic transmitted with the User Datagram Protocol (UDP).

We note for completeness that recently energy efficiency in FiWi access network has begun to attract research interest, see e.g., [62]–[66]. Survivability and protection techniques in FiWi access networks have been studied in [39], [67]–[73], while network coding in FiWi access network has been explored in [74], [75].

In summary, complementary to the existing FiWi networking literature, this study focuses on the effects of a combining (i) routing over cluster heads with (ii) prioritizing transmissions to be routed through the local WMN-PON gateway on overall FiWi network performance. While the existing FiWi routing literature has mainly considered a “flat” topology without a clustering structure of the wireless nodes, clustering techniques have been extensively studied in the area of purely wireless networking, see e.g., [31], [32]. To the best of our knowledge clustered routing in a FiWi network has so far only been studied in [76], which focused on the distribution of traffic in the downstream direction. The present study is the first to examine the benefits of clustered localized routing for peer-to-peer traffic involving both upstream and downstream PON transmissions in a FiWi network.

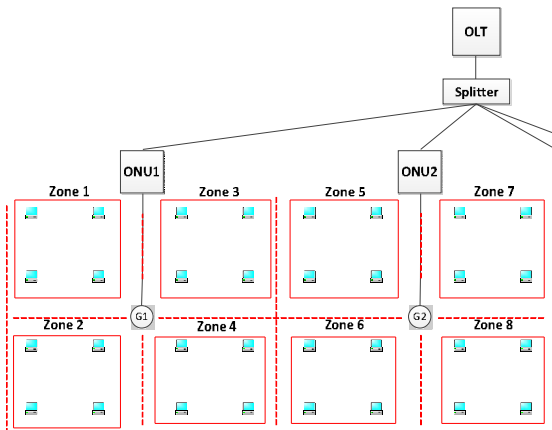


Figure 1. FiWi network structure: Wireless nodes are organized into different zones that operate on different radio frequencies (channels).

III. PRINCIPLES OF CLUSTERED LOCALIZED ROUTING (CLUFOR)

We focus on a setting where the wireless stations (nodes), which could be WiFi routers (e.g., IEEE 802.11g WiFi routers) are organized into different zones, as illustrated in Fig. 1. Each zone operates on a different radio channel than its neighboring zones [33]–[35]. There is a single gateway router that serves the zone closest to it, e.g., zones 1–4 in Fig. 1 are served by gateway router G1. Each gateway router has an Ethernet interface that is connected directly to an ONU. Within this network setting, we examine the two principles of clustered and localized routing that are outlined in the next two subsections and combined to form the CluLoR scheme.

A. Clustered Routing

In each zone, there is a node that is assigned as a cluster head, as illustrated in the upper left illustration in Fig. 2. (It is possible to have multiple cluster heads for a zone, but for ease of exposition, we initially focus on the case of one cluster head per zone.) The cluster head of a zone is the node that is located closest to the gateway router. The cluster head is responsible for routing outbound packets from the regular wireless nodes in the zone on to the gateway router and for routing inbound packets from the gateway router on to the wireless nodes in the zone.

B. Localized Routing

The routing between the wireless nodes (peers) proceeds according to the following three rules, which are summarized in the pseudo-code in Table I: (i) If the communicating peers are within the same zone, then the packet is directly wirelessly transmitted to the destination peer without going through a cluster head or gateway router. (ii) If the zone of the destination peer is serviced by the same gateway router as the zone of the source peer, then the packet is routed by the gateway router to the destination zone without going through the optical network. (iii) If the destination zone is not served by

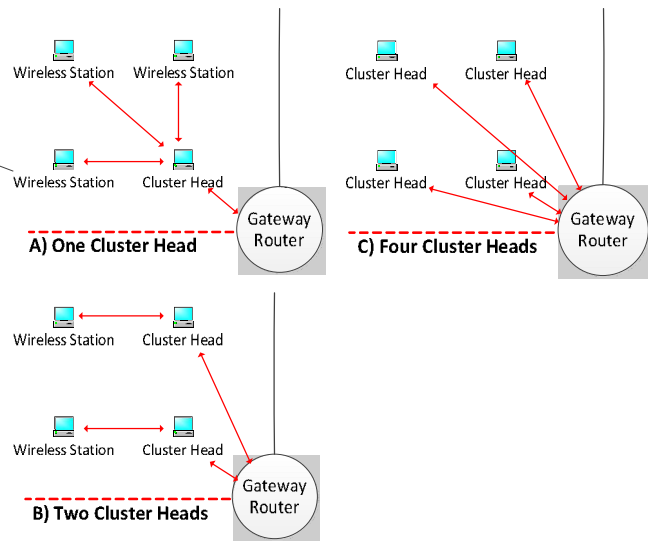


Figure 2. Illustration of clustered routing with different numbers of cluster heads in a zone: Wireless nodes direct all communication with nodes outside the zone through a cluster head. The cluster head(s) communicates with the gateway router, which is connected to an ONU. The configuration with 4 heads corresponds effectively to an unclustered benchmark as all nodes communicate directly (without going through a cluster head) with the gateway router.

TABLE I.
PSEUDO-CODE SUMMARY OF LOCALIZED ROUTING WITHOUT
RELAYS AT DIFFERENT TYPES OF NETWORK NODES.

```

Wireless Station - not a Cluster Head:
if (destination in the same Zone)
    Send packet directly to destination;
else
    Send packet to closest Cluster Head;

Wireless Station - Cluster Head:
if (destination in the same Zone)
    Send packet directly to destination;
else
    Send packet to Gateway Router;

Gateway Router:
if (destination in a Zone associated
    with the same Gateway Router)
    Send packet to Cluster Head associated
    with the destination wireless station;
else
    Send packet to Optical Network;

```

the same gateway router as the source zone, then the packet is routed through the cluster head to the source-zone gateway router, then to the optical network.

The optical network broadcasts the packet in the downstream direction, whereby the ONU connected to the destination gateway (gateway router that is closest to the destination zone) accepts it while the other ONUs discard the packet. The destination gateway router then routes the packet via the cluster head to the destination peer in the destination zone. Localized routing ensures that a packet is never wirelessly transmitted in a zone that does not contain the source or destination of the packet.

TABLE II.
QUAD MODE PAYLOAD SIZES

Ethernet encapsulated packet size	Payload size (UDP level)	Probability
64 bytes	18 bytes	60%
300 bytes	254 bytes	4%
580 bytes	534 bytes	11%
1518 bytes	1472 bytes	25%

IV. SIMULATION SETUP

In our simulations, we evaluate mean end-to-end packet delay and throughput of CluLoR in a FiWi network. The simulations are conducted in OMNeT++ 4.2.2 using INETMANET-2.0 modules. Specifically, we initially simulate a FiWi network with 64 wireless nodes and 4 gateway routers. The wireless nodes are placed uniformly in a 1600 m × 300 m region. The 64 wireless nodes are distributed evenly in 16 zones (4 nodes in each zone) resulting in each gateway managing 4 zones. Each of the 4 wireless gateway routers (IEEE 802.11g) is connected to its own ONU through an Ethernet cable with a transmission rate of 1 Gbps. All the ONUs are at a distance of around 20 km from the OLT.

Each wireless node is equipped with a single radio interface. The gateway routers are equipped with four different radio interfaces (4 radio channels), whereby each channel is assigned to a single zone that operates on the given radio channel. There are 11 different radio channels possible, whereby a given channel is reused in the furthest zones in order to minimize interference. We employ a log-distance path loss channel model with a path loss alpha value of 2. The radio sensitivity is set to -85 dBm and the signal-to-noise ratio threshold is set to 4 dB, whereby the received packet is considered noise if it is below that value. The transmitting power for the wireless routers is set to 20 mW in order for the router that is located furthest in the zone to reach the gateway router. The transmission range is around 250 m. The physical data rate is 54 Mb/s. The retransmit limit for the wireless LAN is set to its default value 7. The buffer size for the wireless interface is set to 1000 packets regardless of the packet size.

We use a quad mode model of payload sizes at the UDP level in order to reach the quad mode of encapsulated packet sizes at the Ethernet level [77], see Table II. The UDP level payload includes the UDP header of 8 bytes, the IP header of 20 bytes, and MAC level header of 18 bytes at the Ethernet layer. The maximum transmission unit (MTU) for the wireless interface is set to 1500 Bytes so as to avoid fragmentation. We consider independent Poisson packet generation processes in the wireless nodes, whereby all the wireless nodes have the same mean packet generation rate. For each generated packet at a given wireless node, any of the other wireless nodes in the network is selected as destination with equal probability. All simulation are run until the 95 % statistical confidence intervals of the performance measures are less than 5 % of the sample means.

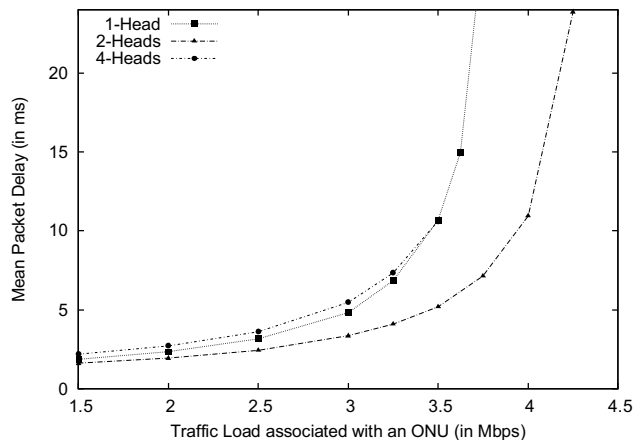


Figure 3. Clustered Routing: Mean end-to-end packet delay for different numbers of cluster heads in a zone.

V. CLUSTERED ROUTING: IMPACT OF NUMBER OF CLUSTER HEADS

A. Clustered Routing Without Relay Routers

In this section we examine the impact of the number of cluster heads in a given zone on the delay and throughput performance. We do not consider relay routers in this section; these are considered in Section V-B. As described in Section IV, the simulated wireless network is organized into different zones, whereby each zone has 4 wireless nodes. In each of the zones there is one wireless node (or multiple wireless nodes) that is (are) assigned as the cluster head (heads) of the zone and is (are) responsible for relaying the traffic from/to the gateway router. We examine the effects of having 1, 2, or 4 cluster heads (which we refer to as “heads” for brevity) in the zone, as illustrated in Fig. 2.

In the case of 1 head, the head is assigned as the wireless node that is located closest to the gateway router. In the case of 2 heads, the two wireless nodes in the zone that are closest to the gateway router are assigned as the heads. Ties in distance are broken through random selection. The outbound traffic from the other wireless nodes (that are not designated as heads) in a given zone is transmitted to the closest head; the head in turn transmits the traffic to the gateway router. Analogously, the inbound traffic is routed from the gateway router to the head that is closest to the destination node and then onwards by the head to the destination. In the case of 4 heads in a zone, all the wireless nodes in a zone are designated as heads and send their traffic directly to the gateway router. Note that the 4-head case is equivalent to unclustered routing in that all wireless nodes communicate directly with their gateway router, without a cluster hierarchy in the zone.

1) *Delay Performance:* Figure 3 shows the mean end-to-end packet delay in the FiWi network for 1, 2, or 4 heads in a zone. (The 95 % confidence intervals are too small to be visible and are omitted.) For each configuration of heads, the network traffic load is incremented until buffer overflows begin to occur; buffer overflows

are examined in detail in Section V-A.2. We observe from the figure that assigning 2-heads in the zone gives lower delays compared to the 1-head or 4-heads cases. In addition, we observe from Fig. 3 that at low loads, 1 head gives lower mean delays than 4 heads. These performance characteristics are mainly due to a trade-off between mean hop-count and transmission distance. In particular, a smaller mean hop-count implies that a packet is transmitted on average fewer times on its way from source to destination. Clearly, fewer transmissions are generally preferable as each transmission requires networking resources and incurs delay.

In the configuration with 4-heads in a zone (i.e., effectively the unclustered scenario, see Fig. 2), all four wireless nodes in a zone send directly packets to the gateway, i.e., all packets originating from the zone need only one hop to reach the gateway. Similarly, all packets arriving to the gateway for delivery to a node in the zone, reach their destination with one hop. Notice that the 4-heads configuration has the minimum mean hop-count among the three configurations illustrated in Fig. 2. As the number of heads decreases, the mean hop count increases. Specifically, the 1-head configuration requires one hop to reach the gateway from the head, but two hops to reach the gateway from any other node in the zone. Thus, to summarize, the 1-head configuration has the highest mean hop-count, the configuration with 2 heads has a moderate mean hop-count, and the 4-heads configuration has the lowest mean hop-count.

The transmission distance directly affects the received signal-to-interference and noise ratio (SINR), with transmissions propagating over longer distances being received with lower SINR. Among the considered configurations, see Fig. 2, the 4-heads configuration has the longest propagation distances, as all nodes in the zone transmit directly to and receive directly from the gateway router. Especially the propagation distance from the node in the upper left corner in the 4-heads illustration to the gateway router in Fig. 2 is the longest propagation distance among any of the three considered configurations. This long-distance transmission is particularly vulnerable to failure due to low SINR and requiring retransmissions. Notice from Fig. 2 that in comparison with the 4-heads configuration, the 1-head and 2-heads configurations both have moderate propagation distances, i.e., only moderate chances of a packet transmission being unsuccessful due to low SINR.

Returning to the interpretation of the results in Fig. 3, we note that the 4-heads configuration incurs the highest mean packet delay mainly due to the long propagation distances and the resulting packet transmission failures due to low SINR and packet re-transmissions. The lower mean-hop count cannot overcome the disadvantage of the long propagation delays and results in relatively frequent packet failures and retransmissions, which dominate the delay characteristics.

In the configuration with 1 head, the propagation distances are shorter, reducing the probability of packet failure due to low SINR. Thus, mean packet delays are

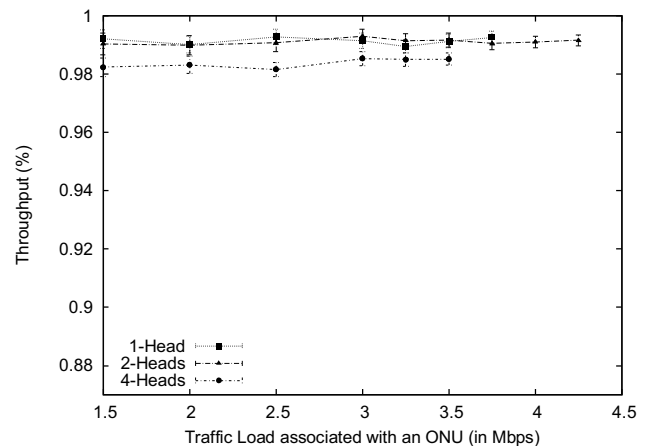


Figure 4. Clustered routing: Normalized throughput for different numbers of heads

slightly reduced compared to the 4-heads configuration. But transmissions from/to 3 wireless nodes in the zone require two hops to reach/come from the gateway router.

The configuration with 2 heads strikes a good balance between low mean-hop count and short propagation distances (i.e., high SINR) achieving the lowest mean packet delays in Fig. 3. The 2-heads configuration has similarly short propagation distances for transmissions from/to the wireless nodes in the zone as the 1-head configuration. At the same time, the 2-heads configuration has a lower mean-hop count than the 1-head configuration, since the transmissions from/to one more node in the zone, i.e., the second head, require only one hop to reach/come in from the gateway router.

2) *Throughput Performance:* Figure 4 shows the 95 % confidence intervals of the normalized mean (long-run average) throughput in terms of traffic that reaches its final destination. The traffic load is incremented until buffer overflows occur; for each curve, the rightmost point corresponds to the highest traffic load without buffer overflows. The average throughput is measured based on the number packets with their corresponding numbers of bits that are received by the destination wireless nodes. The packets (bits) received by intermediate cluster heads and gateway routers are not taken into account.

We observe from Fig 4 that the mean throughput is statistically the same for 1 head and 2 heads in the zone, whereby they both have higher throughput than the 4-heads configuration. We further observe that the 2-heads configuration accommodates higher traffic loads, up to about 4.25 Mbps before buffer overflows occur, whereas the 1-head configuration avoids buffer overflows only up to a load of about 3.75 Mbps. The explanations for these behaviors are as follows. First, the 4-heads case has lower average hop-count than the other two cases; however, the long transmission distance from the wireless node farthest from the gateway router has lower SINR than any transmissions in the 1-head and 2-heads cases. Thus, the farthest-away node relatively frequently requires packet retransmissions and hence reaches the maximum

retransmit limit relatively more often compared to the nodes in the 1-head and 2-heads configurations. As a result, more packets are dropped due to reaching the maximum retransmission limit in the 4-heads configuration compared to the 1-head and 2-heads configurations resulting in lower throughput for the 4-heads configuration at low to moderate traffic loads. Moreover, in the 4-heads configuration, the buffers fill up more due to more frequent packet retransmissions, leading to buffer overflows at lower traffic loads (3.5 Mbps); whereas, the 1-head and 2-heads configurations avoid buffer overflows up to 3.75 and 4.25 Mbps, respectively.

At low loads, both the 1-head and 2-heads configurations achieve similar throughput levels. This is because the (very slightly) shorter transmission distances (i.e., higher SINRs) with the 1-head configuration largely counterbalance its higher hop-count. Similarly, the (very slightly) longer transmission distances (i.e., lower SINRs) largely counterbalance the lower hop-count for the 2-heads configuration. As the traffic load grows high and buffer backlogs grow, the bottleneck in the 1 head leads to buffer overflows at a lower traffic rate compared to when 2 heads share the traffic load going wirelessly in and out of a zone. In fact, we have observed in our simulations that in the case of 2 heads, the buffer overflow first occurs at the gateway router as it wirelessly transmits all traffic destined into a zone to the two heads.

B. Performance with Relay Routers

Relay routers can be thought of as an extra cluster head in the zone. They are only used to relay the packets between neighboring zones, so that if the destination is in an adjacent zone, then the packet is directly transmitted to the relay router, which in turn sends the packet to the destination, as illustrated in Figure 5. The Pseudo-code for the routing algorithm is summarized in Table III. Relay routers are equipped with two different radio interfaces, which are configured to the two radio channels of the two adjacent zones. Relay routers relieve the cluster heads and the gateway routers from sending packets destined to a direct neighbor zone.

Figure 6 shows the mean packet delay with 22 relay routers added to the network configuration of Section IV and without added relay routers for the different configurations of cluster heads in a zone. We first observe that the performance with added relay routers for the different numbers of cluster heads in the zone follows the same general pattern as for the network without relays, see Section V-A. We also observe that adding relay routers results in substantially lower mean end-to-end packet delays, particularly for moderate to high traffic loads. These mean delay results illustrate the effects of bypassing the cluster heads and gateway routers, which lowers the mean hop-count. Also, the packets destined to adjacent zones avoid the queuing delays in the gateway and head routers.

Upon closer examination of Fig. 6, we notice that the relays have a slightly more pronounced effect for the 4-

TABLE III.
PSEUDO-CODE SUMMARY OF LOCALIZED ROUTING WITH RELAY ROUTERS FOR THE DIFFERENT TYPES OF NETWORK NODES.

```

Wireless Station - not a Cluster Head:
if (destination in the same Zone)
    Send packet directly to destination;
else if (destination in adjacent
        Zone & share a Relay Router)
    Send packet to Relay Router;
else
    Send packet to closest Cluster Head;

Wireless Station - Cluster Head:
if (destination in the same Zone)
    Send packet directly to destination;
else if (destination in adjacent
        Zone & share a Relay Router)
    Send packet to Relay Router;
else
    Send packet to Gateway Router;

Relay Router
    Send packet to adjacent Zone

Gateway Router:
if (destination in a Zone associated
    with the same Gateway Router)
    Send packet to Cluster Head associated
    with the destination wireless station;
else
    Send packet to Optical Network;
    
```

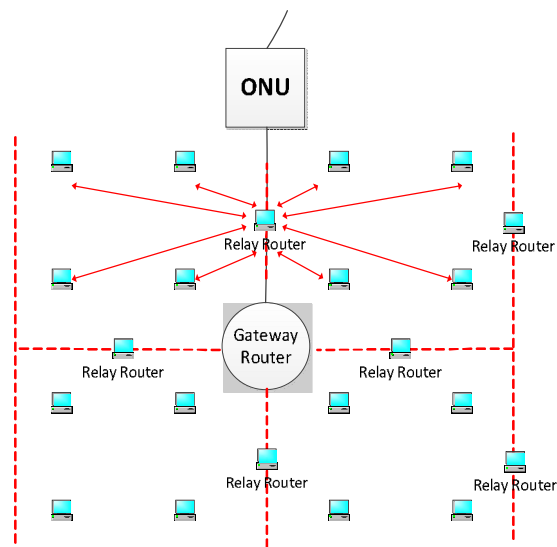


Figure 5. FiWi network structure with relays operating at the two radio frequencies of the two adjacent zones. Packets destined to an adjacent zone are routed through the relay router, bypassing the cluster heads and gateway router.

heads configuration compared to the 1-head and 2-heads configurations. This is because the average propagation distance from the wireless nodes to the relay routers is lower than to the gateway router for the 4-heads configuration. On the other hand, for the 1-head and 2-heads configurations, the propagation distances from the wireless nodes to the relay (without going through the head(s)) are somewhat longer than the distances to the cluster heads. Thus, the benefits of the relay routers

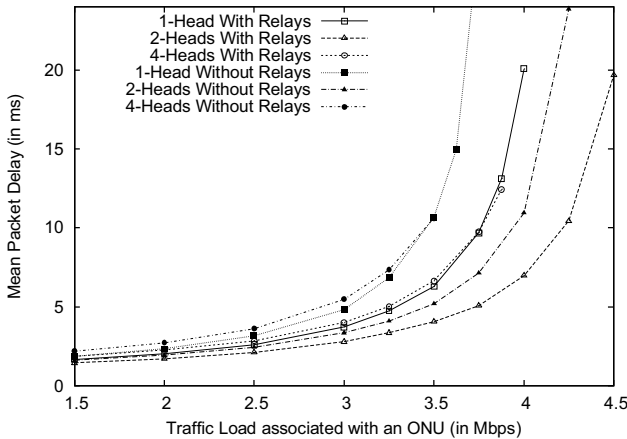


Figure 6. Clustered routing with relays: Mean packet delay as a function of traffic load for different numbers of cluster heads in a zone, with and without relays between adjacent zones.

are somewhat less pronounced with 1 or 2 cluster heads compared to the 4-heads configuration.

We observed from additional simulations for which we do not include plots to avoid clutter, that the throughput levels with relays are only very slightly elevated compared to the throughput levels without relay routers (see Fig. 4). However, the maximum traffic load that can be accommodated before buffer overflows occur is significantly increased by the relays; specifically for the 1-head configuration from 3.75 to 4 Mbps, for 2 heads from 4.25 to 4.75 Mbps, and for 4 heads from 3.5 to 3.875 Mbps.

C. Goodput for Delay Sensitive Traffic

To obtain deeper insights into the performance of the different cluster head and relay configurations, we simulated our FiWi network with a delay sensitive application. An example of delay sensitive application is online video gaming, for which packet delays should not exceed 50 ms. Higher delays disrupt the interactions between the players making the game impossible to play. In interactive video games, many of the participating players are located in the same geographic region and thus peer-to-peer traffic in a FiWi network is a reasonable model. Figure 7 shows the goodput, i.e., the portion of the normalized throughput that arrives within the 50 ms delay limit, for the different configurations. We observe from Fig. 7 that clearly the configuration with two heads in a zone combined with relays gives the highest goodput among the considered schemes. The goodput gains with relays are especially pronounced at high traffic loads. For a load of 4.25 Mbps, for instance, the relays increase the goodput by approximately 10 % for the 2-heads configuration.

VI. LOCALIZED ROUTING: COMPARISON WITH UNLOCALIZED MINIMUM-HOP-COUNT ROUTING

In this section we compare the performance of our proposed CluLoR with an unlocalized routing bench-

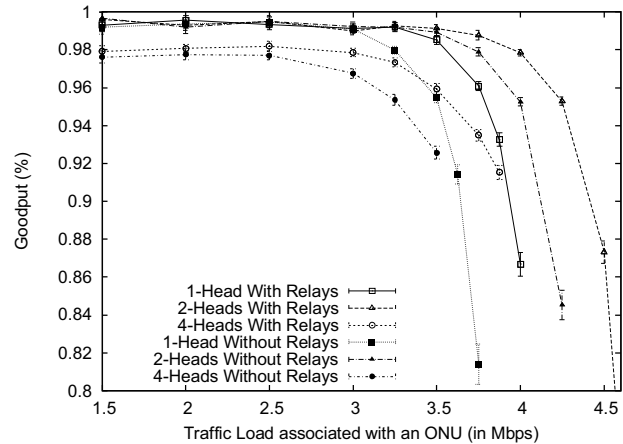


Figure 7. Clustered routing with relays: Normalized Goodput for traffic with delay limit of 50 ms

mark based on minimum-hop-count routing [42]. CluLoR transmits the traffic wirelessly only in zones that contain the source or the destination; while the traffic is routed through the fiber network from the source to the destination zone. In contrast, with unlocalized routing the traffic may be transmitted wirelessly in zones that contain neither the destination nor the source, i.e., the traffic may traverse some intermediate zones via wireless transmissions following, e.g., minimum-hop-count routing. We consider in this section the best performing clustered routing configuration from Section V, i.e., the configuration with two heads per zone and with relays.

Figure 8 illustrates CluLoR and unlocalized minimum-hop-count routing for an illustrative example with one traffic source, namely a regular wireless station, and four possible destinations. With CluLoR, the traffic is routed through the cluster head (first hop) to the gateway router (second hop), from the gateway router G1 adjacent to the source zone through the fiber network to the gateway router G2 adjacent to the destination zone (third hop), to the cluster heads (fourth hop), and regular wireless station destinations (fifth hop). Clearly, with CluLoR, the traffic is only transmitted wirelessly in a zone that includes the source or the destination; thus there is no wireless interference created in any other zones.

In contrast, unlocalized routing based on the minimum hop-count routes the traffic from the source node to the relay router between zones 1 and 2 (first hop), then the relay router transmits the packet on the wireless channel of zone 2 to reach the relay router between zones 2 and 3 (second hop), and the packet is then transmitted in turn by the relay router to reach the destinations in zone 3 (third hop).

Figure 9 compares the mean end-to-end packet delay for CluLoR with unlocalized minimum hop-count routing for the configuration with two heads per zone with relays. We observe from the figure that at lower traffic loads, the delays for both routing approaches are comparable. However, as the traffic load increases, CluLoR achieves

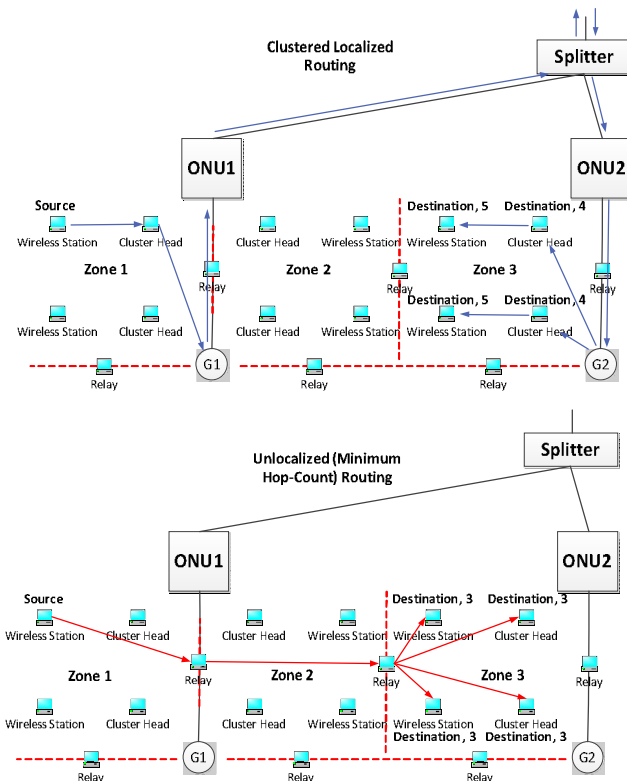


Figure 8. Illustration of clustered localized routing (CluLoR) with two heads and relays for a scenario with a regular wireless router as source: CluLoR avoids the traversal of zones that do not contain the source or destination node by routing through the fiber network. In contrast, unlocalized routing traverses zones without a source/destination (using the relays operating at both radio frequencies of adjacent zones) to achieve the minimum hop-count.

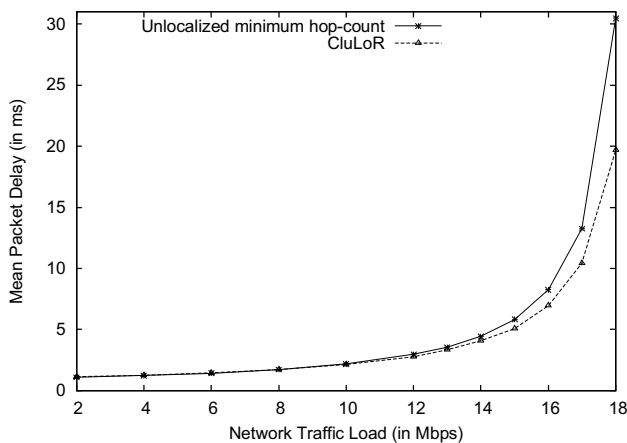


Figure 9. Localized routing: Mean packet delay for CluLoR vs. unlocalized minimum hop-count routing.

lower mean packet delays than unlocalized minimum hop-count routing. For a traffic load of 18 Mbps, the mean packet delay with CluLoR is only about two thirds of the delay with unlocalized minimum hop-count routing. The higher delay with unlocalized routing is mainly due to “transit” traffic through zones that contain neither the source nor the destination. The wireless transmissions

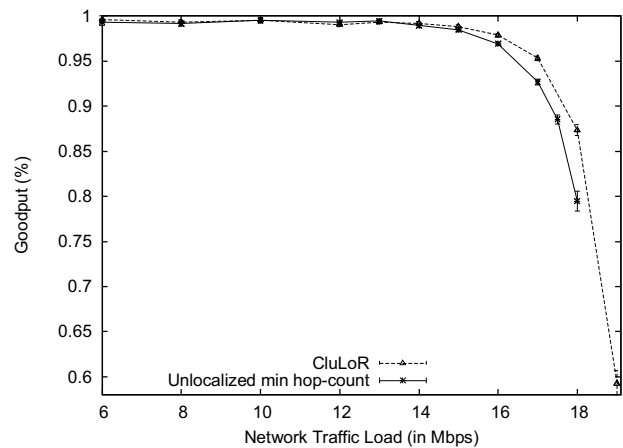


Figure 10. Localized routing: Normalized goodput for 50 ms delay limit for CluLoR vs. unlocalized minimum hop-count routing.

of this transit traffic increase the interference resulting in higher probability of packet transmissions failing due to low SINR as well as an increased chance of packet collisions. Consequently, more packet re-transmission are required, resulting in increased mean packet delays.

Both clustered localized routing and unlocalized minimum hop-count routing achieve normalized throughput levels close to 100 %, we do therefore not include a throughput plot here to avoid clutter. The only noticeable difference between CluLoR and unlocalized minimum hop-count routing is that CluLoR accommodates traffic loads up to 19 Mbps without buffer overflows compared to 18 Mbps with unlocalized routing. This behavior is mainly due to the higher interference with unlocalized routing, which causes more packets to become backlogged due to the more frequent retransmissions; hence, increasing the chance of buffer overflows.

Figure 10 compares the goodput for a delay limit of 50 ms for CluLoR with unlocalized minimum hop-count routing. We observe that clustered localized routing achieves significantly higher throughput, particularly for high traffic loads. For a traffic load of 18 Mbps, the goodput is over 8 % higher with CluLoR compared to unlocalized routing.

VII. EVALUATION FOR HIGHLY LOADED FIBER NETWORK

So far our evaluation of CluLoR has focused on networking scenarios with only peer-to-peer traffic among the wireless stations. In this section, we add a high background traffic load that traverses only the fiber network and examine the impact on peer-to-peer traffic between the wireless stations that is routed following the CluLoR approach. More specifically, we increase the ONU traffic load by adding a wired incoming traffic component. We also increase the number of ONUs to more heavily load the PON access network.

In particular, we simulate a PON network with 32 ONUs. The ONUs are divided into 8 groups; each group

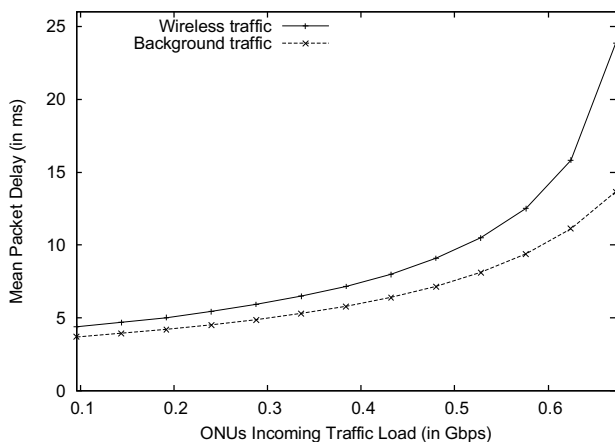


Figure 11. Mean end-to-end packet delay for PON with fiber background traffic (32 ONUs with 2:1 ratio of background traffic : peer-to-peer wireless station traffic).

has 4 ONUs. One reason for dividing the ONUs into 8 groups is to minimize the interference between the zones. Having all the 32 ONUs in one region could significantly affect the performance because wireless routers in a far-away cluster could be within the sensing range of a given cluster. The wireless nodes are uniformly distributed in a given region with a distance between each node of 50 meters. As in the set-up in Section IV, each ONU handles 16 wireless nodes. The distance separation between each region (group) of ONUs is set to be larger than 1 km. The transmission power is set to 20 mW. Each ONU is associated with 4 zones and each zone is configured with 2 cluster heads. In this section, we do not consider relay routers as our focus is to examine the impact of background traffic in the fiber network on the peer-to-peer traffic of the wireless stations. Omitting relay routers forces more of the peer-to-peer traffic through the fiber network and thus gives a worst-case assessment of the impact of fiber network background traffic.

We maintain a ratio of the incoming ONU traffic to be 2:1 for fiber network background traffic : peer-to-peer wireless node traffic. All traffic follows independent Poisson packet generation processes. For the fiber network background traffic, the destination is considered to be an Internet destination outside the PON network. We measure the delay of fiber background traffic from the instant of packet generation to the instant that the packet is completely received by the OLT.

Figure 11 shows the delay performance for background traffic and peer-to-peer wireless station traffic. The x-axis represents the total aggregate incoming traffic load at the 32 ONUs. We observe from the figure that the delays for background and peer-to-peer wireless traffic follow the same curve shape at low loads. However, for moderate to high traffic loads, a pronounced gap opens up between the wireless traffic delay and the background traffic delay. This gap grows wider with increasing traffic load.

The delay results in Fig. 11 indicate that at low traffic loads, the delays in the optical network, which is the only

network component traversed by the background traffic, dominate the wireless traffic delay. That is, the wireless transmissions to and from the gateway routers contribute relatively little to the delay experienced by the peer-to-peer wireless traffic; most of the delay comes from the PON transmissions (more specifically, the upstream transmissions, which require polling-based medium access control proceeding in polling cycles [78]). On the other hand, for high traffic loads, the probability of collisions in the random access of the wireless channels increases, which causes retransmissions that in turn increase delays. These wireless transmission delays add quite significantly to the PON delays. For a traffic load close to 0.7 Gbps, the additional wireless transmission delay experienced by the peer-to-peer wireless traffic is approximately 10 ms on top of the roughly 13 ms of the PON delay.

VIII. CONCLUSION & FUTURE DIRECTIONS

We have examined the combined effects of clustered and localized routing (CluLoR) in fiber-wireless (FiWi) networks. CluLoR is a simple routing strategy that does not require route discovery and maintenance between distant regions of the wireless mesh network (WMN) of a FiWi network. Instead, WMN nodes require only local routes to and from their nearby cluster heads, while in turn the cluster heads require only routes to their nearby gateway routers that interface the WMN with the fiber network.

Our evaluations for CluLoR in a FiWi network organized into zones operating on different radio channels revealed that the clustered routing strategy where regular wireless nodes communicate via cluster heads with the gateway router improves the throughput-delay performance compared to unclustered routing where wireless nodes directly communicate with the gateway router. Our evaluation of the localized routing strategy indicated substantial throughput-delay improvements over an unlocalized minimum hop-count routing strategy.

There are many important directions for future research on simple, yet effective FiWi routing strategies. One direction is to examine the integration of routing in FiWi networks with the routing in metropolitan area networks [79]–[83] that interconnect the FiWi network with the Internet backbone as well as the interoperation with modern cellular networking standards, such as LTE-Advanced [84], [85]. Another direction is to explore how to extend the networking service from the wireless nodes to their local area, e.g., the wireless nodes could support local body area or sensor networks [86]–[91] and help them to communicate over the access network.

REFERENCES

- [1] M. Ahsan, M. Lee, S. Newaz, and S. Asif, "Migration to the next generation optical access networks using hybrid WDM/TDM-PON," *Journal of Networks*, vol. 6, no. 1, pp. 18–25, 2011.
- [2] G. Kramer, M. De Andrade, R. Roy, and P. Chowdhury, "Evolution of optical access networks: Architectures and capacity upgrades," *Proc. IEEE*, vol. 100, no. 5, pp. 1188–1196, May 2012.

- [3] F. J. Effenberger, "The XG-PON system: Cost effective 10 Gb/s access," *IEEE J. Lightw. Techn.*, vol. 29, pp. 403–409, Feb. 2011.
- [4] H. Fathallah and A. Helmy, "Analyzing the performance of centralized polling for long-reach passive optical networks," in *Proc. of Int. Conf. Commun. and Information Techn. (ICCIT)*, 2012, pp. 166–170.
- [5] T. Jimenez, N. Merayo, P. Fernandez, R. Duran, I. de Miguel, R. Lorenzo, and E. Abril, "Implementation of a PID controller for the bandwidth assignment in long-reach PONs," *IEEE/OSA J. Optical Commun. and Netw.*, vol. 4, no. 5, pp. 392–401, May 2012.
- [6] B. Kantarci and H. Moustafah, "Delay-constrained admission and bandwidth allocation for long-reach EPON," *Journal of Networks*, vol. 7, no. 5, pp. 812–820, 2012.
- [7] A. Mercian, M. McGarry, and M. Reisslein, "Offline and online multi-thread polling in long-reach PONs: A critical evaluation," *IEEE/OSA J. Lightwave Techn.*, vol. 31, no. 12, pp. 2018–2228, June 2013.
- [8] A. Sivakumar, G. C. Sankaran, and K. M. Sivalingam, "A comparative study of dynamic bandwidth allocation algorithms for long reach passive optical networks," *IETE Techn. Rev.*, vol. 29, no. 5, pp. 405–413, 2012.
- [9] O. C. Turna, M. A. Aydin, T. Atmaca, and A. Zaim, "A novel dynamic bandwidth allocation algorithm based on half cycling for EPONs," in *Proc. Int. Conf. on Emerging Netw. Intelligence*, Oct. 2010, pp. 38–43.
- [10] G. Belleffi, G. Incerti, L. Porcari, S. D. Bartolo, M. Guglielmucci, A. Teixeira, L. Costa, N. Wada, J. Prat, J. Lazaro, and C. P., "Reducing complexity and consumption in future networks," *J. Netw.*, vol. 5, no. 11, pp. 1310–1314, 2010.
- [11] A. R. Dhaini, P.-H. Ho, and X. Jiang, "QoS Control for Guaranteed Service Bundles Over Fiber-Wireless (FiWi) Broadband Access Networks," *IEEE/OSA J. Lightw. Techn.*, vol. 29, no. 10, pp. 1500–1513, May 2011.
- [12] K. Grobe and J.-P. Elbers, "PON in adolescence: From TDMA to WDM-PON," *IEEE Comm. Mag.*, vol. 46, no. 1, pp. 26–34, Jan. 2008.
- [13] Y. Luo and N. Ansari, "Bandwidth allocation for multi-service access on EPONs," *IEEE Commun. Mag.*, vol. 43, no. 2, pp. S16–S21, Feb. 2005.
- [14] M. McGarry and M. Reisslein, "Investigation of the DBA algorithm design space for EPONs," *IEEE/OSA J. Lightwave Techn.*, vol. 30, no. 14, pp. 2271–2280, July 2012.
- [15] A. Razmkhah and A. G. Rahbar, "OSLG: A new granting scheme in WDM ethernet passive optical networks," *Opt. Fiber Techn.*, vol. 17, no. 6, pp. 586–593, Dec. 2011.
- [16] F. Slaveski, J. Sluss, M. Atiquzzaman, H. Nguyen, and D. Ngo, "Optical fiber wavelength division multiplexing," *IEEE Aerospace and Electronic Systems Mag.*, vol. 18, no. 8, pp. 3–8, 2003.
- [17] J. Zhang and N. Ansari, "Scheduling hybrid WDM/TDM passive optical networks with nonzero laser tuning time," *IEEE/ACM Trans. Netw.*, vol. 19, no. 4, pp. 1014–1027, Aug. 2011.
- [18] A. Alsarhan and A. Agarwal, "Cluster-based spectrum management using cognitive radios in wireless mesh network," in *Proc. IEEE ICCCN*, 2009, pp. 1–6.
- [19] H. Cheng, N. Xiong, G. Chen, and X. Zhuang, "Channel assignment with topology preservation for multi-radio wireless mesh networks," *Journal of Communications*, vol. 5, no. 1, pp. 63–70, 2010.
- [20] T. Chen, H. Zhang, G. Maggio, and I. Chlamtac, "Topology management in CogMesh: a cluster-based cognitive radio mesh network," in *Proc. IEEE ICC*, 2007, pp. 6516–6521.
- [21] M. Iqbal, X. Wang, D. Wertheim, and X. Zhou, "Swan-Mesh: A multicast enabled dual-radio wireless mesh network for emergency and disaster recovery services," *Journal of Communications*, vol. 4, no. 5, pp. 298–306, 2009.
- [22] K.-C. Lan, Z. Wang, R. Berriman, T. Moors, M. Hassan, L. Libman, M. Ott, B. Landfeldt, Z. Zaidit, and A. Seneviratne, "Implementation of a wireless mesh network testbed for traffic control," in *Proc. ICCCN*, 2007, pp. 1022–1027.
- [23] V. Loscri, "On the interaction between multiple paths and wireless mesh networks scheduler approaches," *Journal of Networks*, vol. 3, no. 7, pp. 64–77, 2008.
- [24] J. Nunez-Martinez and J. Mangués-Bafalluy, "A survey on routing protocols that really exploit wireless mesh network features," *J. Commun.*, vol. 5, no. 3, pp. 211–231, 2010.
- [25] Y. Qin and R. Zhu, "Efficient routing algorithm based on decision-making sequence in wireless mesh networks," *Journal of Networks*, vol. 7, no. 3, pp. 502–509, 2012.
- [26] V. Borges, D. Pereira, M. Curado, and E. Monteiro, "Routing metric for interference and channel diversity in multi-radio wireless mesh networks," in *Ad-Hoc, Mobile and Wireless Networks*, ser. Lecture Notes in Computer Science, P. Ruiz and J. Garcia-Luna-Aceves, Eds. Springer, 2009, vol. 5793, pp. 55–68.
- [27] R. Zhang, Y. Song, F. Chu, and B. Sheng, "Study of wireless sensor networks routing metric for high reliable transmission," *Journal of Networks*, vol. 7, no. 12, pp. 2044–2050, 2012.
- [28] J. Zhang, B. Wang, and X. Jia, "Relative-closest connect-first method for topology control in wireless mesh networks," in *Proc. IEEE Globecom*, 2009, pp. 1–6.
- [29] M. Xia, Y. Owada, M. Inoue, and H. Harai, "Optical and wireless hybrid access networks: Design and optimization," *IEEE/OSA J. of Optical Commun. and Netw.*, vol. 4, no. 10, pp. 749–759, 2012.
- [30] N. Ghazisaidi and M. Maier, "Fiber-wireless (FiWi) networks: Challenges and opportunities," *IEEE Network*, vol. 25, no. 1, pp. 36–42, Jan./Feb. 2011.
- [31] R. Rajaraman, "Topology control and routing in ad hoc networks: a survey," *ACM SIGACT News*, vol. 33, no. 2, pp. 60–73, 2002.
- [32] J. Y. Yu and P. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Comm. Surv. Tut.*, vol. 7, no. 1-4, pp. 32–48, 2005.
- [33] H. Li, Y. Cheng, C. Zhou, and W. Zhuang, "Minimizing end-to-end delay: A novel routing metric for multi-radio wireless mesh networks," in *Proc. IEEE INFOCOM*, 2009, pp. 46–54.
- [34] F. Theoleyre, B. Darties, and A. Duda, "Assignment of roles and channels for a multichannel MAC in wireless mesh networks," in *Proc. IEEE ICCCN*, 2009, pp. 1–6.
- [35] J. Wang, M. Song, G. Hsieh, and C. Xin, "Minimum cost broadcast in multi-radio multi-channel wireless mesh networks," in *Proc. Int. Conf. Mobile Ad-hoc and Sensor Networks*, 2011, pp. 238–247.
- [36] L. Kazovsky, S.-W. Wong, T. Ayhan, K. Albeyoglu, M. Reiberio, and A. Shastri, "Hybrid optical-wireless access networks," *Proc. IEEE*, vol. 100, no. 5, pp. 1197–1225, May 2012.
- [37] S. Sarkar, S. Dixit, and B. Mukherjee, "Hybrid wireless-optical broadband-access network (WOBAN): A review of relevant challenges," *IEEE Journal of Lightwave Technology*, vol. 25, no. 11, pp. 3329–3340, Nov. 2007.
- [38] Y. Liu, C. Zhou, and Y. Cheng, "S²U: An efficient algorithm for optimal integrated points placement in hybrid optical-wireless access networks," *Computer Communications*, vol. 34, no. 11, pp. 1375–1388, 2011.
- [39] Y. Liu, L. Guo, B. Gong, R. Ma, X. Gong, L. Zhang, and J. Yang, "Green survivability in fiber-wireless (FiWi) broadband access network," *Optical Fiber Technology*, vol. 18, no. 2, pp. 68–80, Mar. 2012.
- [40] S. Bhandari and E. Park, "Hybrid optical wireless networks," in *Proc. ICN/ICONS/MCL*, 2006, pp. 1–5.
- [41] S.-W. Wong, D. Campelo, N. Cheng, S.-H. Yen, L. Kazovsky, H. Lee, and D. Cox, "Grid reconfigurable optical-

- wireless architecture for large scale municipal mesh access network,” in *Proc. IEEE Globecom*, 2009, pp. 1–6.
- [42] Z. Zheng, J. Wang, and J. Wang, “A study of network throughput gain in optical-wireless (FiWi) networks subject to peer-to-peer communications,” in *Proc. IEEE ICC*, June 2009, pp. 1–6.
- [43] Z. Zheng, J. Wang, and X. Wang, “ONU placement in fiber-wireless (FiWi) networks considering peer-to-peer communications,” in *Proc., IEEE GLOBECOM*, Nov./Dec. 2009, pp. 1–7.
- [44] Y. Li, J. Wang, C. Qiao, A. Gumaste, Y. Xu, and Y. Xu, “Integrated fiber-wireless (FiWi) access networks supporting inter-ONU communications,” *IEEE J. Lightw. Techn.*, vol. 28, no. 5, pp. 714–724, Mar. 2010.
- [45] N. Nadarajah, M. Attygalle, A. Nirmalathas, and E. Wong, “A novel local area network emulation technique on passive optical networks,” *IEEE Phot. Techn. Letters*, vol. 17, no. 5, pp. 1121–1123, May 2005.
- [46] A. V. Tran, C. J. Chae, and R. S. Tucker, “Bandwidth-efficient PON system from broadband access and local customer internetworking,” *IEEE Photonics Techn. Letters*, vol. 18, no. 3, pp. 670–672, March 2006.
- [47] G. Kalfas and N. Pleros, “An agile and medium-transparent MAC protocol for 60 GHz radio-over-fiber local access networks,” *IEEE/OSA Journal of Lightwave Technology*, vol. 28, no. 16, pp. 2315–2326, 2010.
- [48] W. Chang, H. Lin, S. Hong, and C. Lai, “A novel WDM EPON architecture with wavelength spatial reuse in high-speed access networks,” in *Proc. ICON*, Nov. 2007, pp. 155–160.
- [49] G. Kramer, B. Mukherjee, and G. Pesavento, “IPACT a dynamic protocol for an Ethernet PON (EPON),” *IEEE Communications Magazine*, vol. 40, no. 2, pp. 74–80, February 2002.
- [50] F. Aurzada, M. Scheutzow, M. Herzog, M. Maier, and M. Reisslein, “Delay analysis of Ethernet passive optical networks with gated service,” *OSA Journal of Optical Networking*, vol. 7, no. 1, pp. 25–41, Jan. 2008.
- [51] S. Sarkar, H.-H. Yen, S. Dixit, and B. Mukherjee, “A novel delay-aware routing algorithm (DARA) for a hybrid wireless-optical broadband access network (WOBAN),” *IEEE Network*, vol. 22, no. 3, pp. 20–28, Jan./Feb. 2008.
- [52] X. Chen, A. Reaz, L. Shi, P. Chowdhury, Y. Zhang, R. Wang, and B. Mukherjee, “Delay-differentiated routing algorithm to enhance delay performance of WOBAN,” in *Proc., COIN*, July 2010, pp. 1–4.
- [53] A. Reaz, V. Ramamurthi, S. Sarkar, D. Ghosal, S. Dixit, and B. Mukherjee, “CaDAR: An efficient routing algorithm for a wireless-optical broadband access network (WOBAN),” *IEEE/OSA J. Optical Commun. and Netw.*, vol. 1, no. 5, pp. 392–403, Oct. 2009.
- [54] S. Sarkar, H.-H. Yen, S. Dixit, and B. Mukherjee, “RADAR: Risk-and-delay aware routing algorithm in a hybrid wireless-optical broadband access network (WOBAN),” in *Proc., OFC/NFOEC*, Mar. 2007, pp. 1–3.
- [55] P. Chowdhury, M. Tornatore, S. Sarkar, and B. Mukherjee, “Building a green wireless-optical broadband access network (WOBAN),” *IEEE J. Lightw. Techn.*, vol. 28, no. 16, pp. 2219–2229, Aug. 2010.
- [56] X. Shao, Y. K. Yeo, L. H. Ngho, X. Cheng, W. Rong, and L. Zhou, “Availability-aware routing for large-scale hybrid wireless-optical broadband access network,” in *Proc., OFC*, Mar. 2010, pp. 1–3.
- [57] F. Aurzada, M. Levesque, M. Maier, and M. Reisslein, “FiWi access networks based on next-generation PON and gigabit-class WLAN technologies: A capacity and delay analysis,” *IEEE/ACM Trans. Netw.*, in print, 2014.
- [58] W.-T. Shaw, S.-W. Wong, N. Cheng, K. Balasubramanian, X. Zhu, M. Maier, and L. Kazovsky, “Hybrid architecture and integrated routing in a scalable optical wireless access network,” *IEEE/OSA J. Lightwave Techn.*, vol. 25, no. 11, pp. 3443–3451, Nov. 2007.
- [59] S. Dai, Z. Zheng, J. Wang, and X. Zhang, “Wavelength assignment scheme of ONUs in hybrid TDM/WDM fiber-wireless networks,” in *Proc. IEEE ICC*, May 2010, pp. 1–5.
- [60] J. Wang, K. Wu, S. Li, and C. Qiao, “Performance modeling and analysis of multi-path routing in integrated fiber-wireless networks,” in *Proceedings of IEEE INFOCOM*, March 2010, pp. 1–5.
- [61] S. Li, J. Wang, C. Qiao, and Y. Xu, “Mitigating packet reordering in FiWi networks,” *IEEE/OSA J. Optical Commun. Netw.*, vol. 3, no. 2, pp. 134–144, February 2011.
- [62] S. Chen, A. Dhaini, P.-H. Ho, B. Shihada, G. Shen, and C.-H. Lin, “Downstream-based scheduling for energy conservation in green EPONs,” *Journal of Communications*, vol. 7, no. 5, pp. 400–408, 2012.
- [63] B. Kantarci, M. Khair, and H. Mouftah, “Power saving clusters for energy-efficient design of fiber-wireless access networks,” in *Proc. HONET*, 2010, pp. 73–78.
- [64] B. Kantarci and H. T. Mouftah, “Energy efficiency in the extended-reach fiber wireless access networks,” *IEEE Network*, vol. 26, no. 2, pp. 28–35, March/April 2012.
- [65] A. Reaz, V. Ramamurthi, M. Tornatore, and B. Mukherjee, “Green provisioning of cloud services over wireless-optical broadband access networks,” in *Proc., IEEE GLOBECOM*, Dec. 2011, pp. 1–5.
- [66] A. Reaz, V. Ramamurthi, and M. Tornatore, “Cloud-over-WOBAN (CoW): an offloading-enabled access network design,” in *Proc. IEEE ICC*, June 2011, pp. 1–5.
- [67] N. Ghazisaidi, M. Scheutzow, and M. Maier, “Survivability analysis of next-generation passive optical networks and fiber-wireless access networks,” *IEEE Trans. Reliab.*, vol. 60, no. 2, pp. 479–492, June 2011.
- [68] Y. Liu, Q. Song, R. Ma, B. Li, and B. Gong, “Protection based on backup radios and backup fibers for survivable fiber-wireless (FiWi) access network,” *J. Network and Computer Appl.*, in print, 2013.
- [69] M. Maier, “Survivability techniques for NG-PONs and FiWi access networks,” in *Proc., IEEE ICC*, June 2012, pp. 1–6.
- [70] T. Feng and L. Ruan, “Design of a Survivable Hybrid Wireless-Optical Broadband-Access Network,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 3, no. 5, pp. 458–464, May 2011.
- [71] B. Kantarci and H. T. Mouftah, “Reliable and fast restoration for a survivable wireless-optical broadband access network,” in *Proc. ICTON*, June/July 2010, pp. 1–4.
- [72] Y. Liu, L. Guo, R. Ma, and W. Hou, “Auxiliary graph based protection for survivable fiber-wireless (FiWi) access network considering different levels of failures,” *Opt. Fiber Techn.*, vol. 18, no. 6, pp. 430–439, 2012.
- [73] Z. Yubin, H. Li, X. Ruitao, Q. Yaojun, and J. Yuefeng, “Wireless protection switching for video service in wireless-optical broadband access network,” in *Proc. IC-BNMT*, 2009, pp. 760–764.
- [74] K. Fouli, M. Maier, and M. Médard, “Network coding in next-generation passive optical networks,” *IEEE Commun. Mag.*, vol. 49, no. 9, pp. 38–46, Sept. 2011.
- [75] J. Zhang, W. Xu, and X. Wang, “Distributed online optimization of wireless optical networks with network coding,” *IEEE/OSA J. Lightwave Techn.*, vol. 30, no. 14, pp. 2246–2255, July 2012.
- [76] M. Honda, H. Nishiyama, H. Nomura, T. Yada, H. Yamada, and N. Kato, “On the performance of downstream traffic distribution scheme in fiber-wireless networks,” in *Proc. IEEE WCNC*, 2011, pp. 434–439.
- [77] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, “On the self-similar nature of Ethernet traffic (extended

- version),” *IEEE-ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, February 1994.
- [78] F. Aurzada, M. Scheutzow, M. Reisslein, N. Ghazisaidi, and M. Maier, “Capacity and delay analysis of next-generation passive optical networks (NG-PONs),” *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1378–1388, May 2011.
- [79] A. Bianco, T. Bonald, D. Cuda, and R.-M. Indre, “Cost, power consumption and performance evaluation of metro networks,” *IEEE/OSA J. Opt. Comm. Netw.*, vol. 5, no. 1, pp. 81–91, Jan. 2013.
- [80] M. Maier and M. Reisslein, “AWG-based metro WDM networking,” *IEEE Communications Magazine*, vol. 42, no. 11, pp. S19–S26, Nov. 2004.
- [81] M. Scheutzow, M. Maier, M. Reisslein, and A. Wolisz, “Wavelength reuse for efficient packet-switched transport in an AWG-based metro WDM network,” *IEEE/OSA J. Lightwave Techn.*, vol. 21, no. 6, pp. 1435–1455, June 2003.
- [82] H.-S. Yang, M. Maier, M. Reisslein, and W. Carlyle, “A genetic algorithm-based methodology for optimizing multiservice convergence in a metro WDM network,” *IEEE/OSA J. Lightwave Techn.*, vol. 21, no. 5, pp. 1114–1133, May 2003.
- [83] M. Yuang, I.-F. Chao, and B. Lo, “HOPSMAN: An experimental optical packet-switched metro WDM ring network with high-performance medium access control,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 2, no. 2, pp. 91–101, Feb. 2010.
- [84] J. Seo and V. Leung, “Design and analysis of backoff algorithms for random access channel in UMTS-LTE and IEEE 802.16 system,” *IEEE Trans. Vehicular Techn.*, vol. 60, no. 8, pp. 3975–3989, Oct. 2011.
- [85] R. Tyagi, F. Aurzada, K.-D. Lee, S. Kim, and M. Reisslein, “Impact of retransmission limit on preamble contention in LTE-Advanced network,” *IEEE Systems J.*, in print, 2014.
- [86] T. Andrade, N. Fonseca, L. Oliveira, and O. Branquinho, “MAC protocols for wireless sensor networks over radio-over-fiber links,” in *Proc. IEEE ICC*, 2012, pp. 254–259.
- [87] M. Hossen and M. Hanawa, “Network architecture and performance analysis of MULTI-OLT PON for FTTH and wireless sensor networks,” *Int. J. Wireless & Mobile Networks*, vol. 3, no. 6, pp. 1–15, 2011.
- [88] M. Hossen, K.-D. Kim, and Y. Park, “A PON-based large sensor network and its performance analysis with Sync-LS MAC protocol,” *Arabian J. Science Eng.*, pp. 1–9, 2013.
- [89] L. Oliveira and J. Rodrigues, “Wireless sensor networks: a survey on environmental monitoring,” *Journal of Communications*, vol. 6, no. 2, pp. 143–151, 2011.
- [90] A. Seema and M. Reisslein, “Towards efficient wireless video sensor networks: A survey of existing node architectures and proposal for a Flexi-WVSNP design,” *IEEE Comm. Surv. & Tut.*, vol. 13, no. 3, pp. 462–486, Third Quarter 2011.
- [91] X. Yu, Y. Zhao, L. Deng, X. Pang, and I. Monroy, “Existing PON infrastructure supported hybrid fiber-wireless sensor networks,” in *Proc. OFC*, 2012, pp. 1–3.

Martin Reisslein is a Professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University (ASU), Tempe. He received the Ph.D. in systems engineering from the University of Pennsylvania in 1998. His research interests are in the areas of multimedia networking, optical access networks, and engineering education.

Yousef Dashti received the B.S. and M.S.E. degrees in electrical engineering with emphasis on signal processing and communications from Arizona State University, Tempe, U.S.A., in 2003 and 2008, respectively. Since 2011 he has been pursuing his Ph.D. degree in electrical engineering at Arizona State University, Tempe, USA. He is an assistant teacher at the College of Technological Studies, The Public Authority for Applied Education & Training, Kuwait. His research interests include peer-to-peer communications, fiber-wireless (FiWi) networking, and optical networking.

Water Quality Monitoring and Control for Aquaculture Based on Wireless Sensor Networks

Daudi S. Simbeye* and Shi Feng Yang

College of Electronic Information and Automation, Tianjin University of Science and Technology, 1038 Dagu South Road, Hexi District, Tianjin 300222, P. R. China

*Corresponding author, Email: daudi.simbeye@gmail.com; yangsf@tust.edu.cn

Abstract—We have designed and presented a wireless sensor network monitoring and control system for aquaculture. The system can detect and control water quality parameters of temperature, dissolved oxygen content, pH value, and water level in real-time. The sensor nodes collect the water quality parameters and transmit them to the base station host computer through ZigBee wireless communication standard. The host computer is used for data analysis, processing and presentation using LabVIEW software platform. The water quality parameters will be sent to owners through short messages from the base station via the Global System for Mobile (GSM) module for notification. The experimental evaluation of the network performance metrics of quality of communication link, battery performance and data aggregation was presented. The experimental results show that the system has great prospect and can be used to operate in real world environment for optimum control of aquaculture environment.

Index Terms—Water Quality Monitoring; Wireless Sensor Network; Aquaculture; ZigBee; LabVIEW; GSM

I. INTRODUCTION

Aquaculture is increasingly considered as an integral component in the search for global world food security and economic development. The vast majority of aquaculture production takes place in China. The automation of aquaculture systems will allow the industry to improve environmental control, reduce catastrophic losses, reduce production cost, and improve product quality [1]. The most important parameters to be monitored and controlled in an aquaculture system include temperature, dissolved oxygen, pH, ammonia, nitrates, salinity, and alkalinity, since they directly affect animal health, feed utilization, growth rates and carrying capacities [2].

Water temperature affects the feeding pattern and growth of fish. Fish generally experience stress and disease breakout when temperature is chronically near their maximum tolerance or fluctuates suddenly. Warm water holds less dissolved oxygen than cool water. Oxygen consumption is directly linked to size of fish, feeding rate, activity level and pond temperature. The amount of dissolved oxygen in water increases as temperature reduces, and decreases when salinity increases. Low dissolved oxygen concentration is

recognized as a major cause of stress, poor appetite, slow growth, disease susceptibility and mortality in aquaculture animals [3]. It is generally accepted that the minimum daily dissolved-oxygen concentration in pond culture systems is of greatest concern. Not only is dissolved oxygen important for fish respiration, it is also important for the survival of phytoplankton, the organism which breaks down toxic ammonia into harmless forms. The acceptable range of pH for fish culture is usually between pH 6.5 to pH 9.0. When water is very alkaline (> pH 9), ammonium in water is converted to toxic ammonia, which can kill fish. On the other hand, acidic water (< pH 5) leaches metals from rocks and sediments. These metals have an adverse effect on the fish's metabolism rates and ability to take in water through their gills, and can be fatal as well [4]. Since failure of any component can cause catastrophic losses within a short period of time, the system must be reliable and constantly monitored. Thus, precise measurements and controls are necessary for the success of an intensive aquaculture system [5]-[6].

However, there are few applications of systems which could carry out real-time water quality monitoring continuously in China. According to the conventional methods of water quality monitoring, samples of water are taken and transported to a chemical laboratory to analyze the hazardous substances. On the other hand, the maintenance of the measurements and control process is manual influenced by the personal experience [7].

How to realize real-time data collection in a secure, robust, manageable and low-cost manner without long-distance cable connections is still a bottleneck in the development of information monitoring in fish culture. Modern aquaculture environment detection and control technology achieves high-quality, high yield, improves the basic environmental conditions and is one of the key means to promote fish production through the integrated application of bio engineering and computer technology to make the appropriate adjustments, according to the variation of indicators, increase production, and guarantee reliable income [8]-[9]. A properly-controlled system will also be energy efficient since production can be optimized with respect to the various inputs. So a sustainable development of aquaculture environmental factors monitoring and control system for intensive fish farming is inevitable.

Wireless sensor network (WSN) is an important and exciting new technology with great potential for improving current applications in intensive aquaculture [10]. In contrast to wired sensors, the obstacle has been to develop hardware that is capable of transmitting data under difficult circumstances and developing low-cost, long-term energy sources for the sensor nodes. WSN are in intimate connection with the immediate physical environment allowing each sensor to provide detailed information on environment of material that is otherwise difficult to obtain by means of traditional wired instrumentation [11]. In this work, ZigBee wireless communication technology (IEEE 802.15.4) is preferred over other technologies for the development of wireless sensor network due to its low cost and low power consumption property.

This work focuses on the use of multiple sensors to monitor and control the water quality parameters of temperature, dissolved oxygen, pH and water level in aquaculture in real-time. The sensor nodes collect the water quality parameters and transmit them to the base station host computer through ZigBee wireless communication standard. Several measurement and performance analysis to evaluate the reliability, feasibility and effectiveness of the network performance metrics of quality of communication link, battery performance and data aggregation was presented. The system was tested at Tanggu fish farm demonstration base in Tianjin for six months. The key water quality indicators was precisely controlled by relevant actuators, taking timely measures to improve the stability of variety of factors, greatly savings of electrical energy consumption, providing continuous data that can be used to identify trends and improve production and hence increasing income of aquaculture farming.

II. RELATED WORKS

WSN become an important issue in environmental monitoring. The relatively low cost of the devices allow the installation of nodes that can adequately represent the variability present in the environment [11]. WSNs was applied successfully for monitoring of soil water content, temperature and salt in a cabbage farm of Spain semi-arid regions Murcia [12]-[13]. The design of the WSNs included four types sensor networks topology structure nodes deployed in the field. They were soil node, environmental node, water node and gateway node. Furthermore, the software and hardware of each node were given. The management and real time measurement of the whole system were carried by the central processing computer in the farm management office. System testing was carried in two stages, including the laboratory test and the field test. The laboratory test has analyzed mainly the function of the system devices, network performance and energy consumption; measurement range, robustness and reliability of system test were mainly in the field test.

In [14] a ZigBee WSN was developed for monitoring an experimental aquaculture recirculating system. Temperature, dissolved oxygen, water and air pressure as

well as electric current sensors were included in the setup. Modules for reading and transmitting sensor values through a ZigBee wireless network were developed and tested. The modules were installed in an aquaculture recirculating system to transmit sensor values to the network coordinator. A monitoring program was created in order to display and store sensor values and to compare them with reference limits. E-mail and an SMS message alert can also be sent to the cellular phone of the system administrator so that immediate action can be taken. A web interface allows internet access to the sensor values.

A WSN based on ZigBee in aquaculture was presented by [15]. The aquaculture monitoring environment has characteristics of multi-measuring points, long measuring time and high complexity measuring conditions. This system achieves the goals of collect, transmit and display multi-parameters such as dissolved oxygen and temperature. In [16] a WSN for continuous monitoring water quality in aquaculture farm was developed. Multi-parameter water quality node, temperature chain node, routing node and an on-site monitoring center were designed and implemented. Multi-parameter water quality node was created for measurement of dissolved oxygen, water level and temperature in sea cucumber ponds. The routing node used to extend the range of continuous monitoring in aquaculture farm. Reference [17] developed a WSN based traceability system for recirculation aquaculture (RATS). The system enables rapid deployment and can acquire water temperature, salinity, dissolved oxygen and pH and achieve real-time data transmission. The RATS was mainly developed using C# in Microsoft Visual Studio 2008 integrated with the real-time monitor chart powered by the Matlab M-language dynamic link library. The structure of the WSN to collect and continuously transmit data to the monitoring software was designed by [18]. Then they accomplished the configuration model in the software that enhances the reuse and facility of the monitoring project. Moreover, the monitoring software developed to represent the monitoring hardware and data visualization, and analyze the data with expert knowledge to implement the auto control. The monitoring system has been realization of the digital, intelligent, and effectively ensures the quality of aquaculture water.

Moreover, the use of the ZigBee standard is often seen in agriculture through the use of WSN in order to monitor or control various parameters [19]. Reference [20] conducted a study in real time with the remote measurement of humidity, temperature and brightness of the ambient air. In addition to detect water pollution in irrigation, a ZigBee WSN was installed in agricultural production. In [21] a novel methodology for the monitoring of the agricultural production process based on wireless sensor networks was developed. The authors proposed a methodology consisting of a set of well-defined phases that cover the complete life cycle of WSN applications for agricultural monitoring. An online water monitoring system based on ZigBee and GPRS was developed by [22]. The sensor data were collected and transmitted via ZigBee and GPRS. The data process

procedure was implemented by LabVIEW software. Reference [23] developed a distributed measurement system based on networked smart sensors to monitor aquaculture factors in multi-environment. The system consists of four parts: data collection nodes, routing nodes, on-site monitoring center and remote monitoring center; and can bring-out real-time monitoring water quality parameters and meteorological parameters.

However, the application of their proposed system is still limited by its rather complicated operational requirements and high maintenance cost. Further, none of these studies analyzed battery behavior of sensor nodes in an outdoor environment like our work. Moreover, the systems were not integrated with actuators in nodes for remotely correcting environmental parameters such as dissolved oxygen and water valves. In our work, the graphical user interface (GUI) was designed by LabVIEW software platform so that the users can observe and modify the related values of aquaculture environment. Still, there are many challenges that arise when one want to get the best performance of the network installed in this wide variety of locations. Problems of control and actuation, information packet loss, battery consumption, as well as aspects related to the real-world environments. Very few results exist to date regarding meeting real-time requirements in WSN. Many other functions must also meet real-time constraints including data fusion and data transmission.

III. SYSTEM DESIGN

The wireless sensors used in this experiment monitor temperature, pH, dissolved oxygen and water level. The sensors measure these parameters at specific time intervals and transmit the data wirelessly to a receiver station. The sampling time interval was set to roughly every 3 minutes in order obtain a long effective transmission communication range and 2.405 GHz was selected as the communication frequency for this application. As shown in Fig. 1, the entire system has a sensor node, communication device and the base station. Accordingly, each sensor node is designed to communicate to the base station via ZigBee communication technology. The base station host computer acts as the central monitoring platform for data analysis, processing and presentation. While the sensor nodes acts as the remote monitoring platform for data acquisition.

A. Sensor Nodes

The sensor nodes consists of data acquisition, data transformation and transmission, and water quality control components. Data acquisition component collects non electricity signals of the most important environmental factors by using various sensors. With the current measurement, pH value is measured by glass electrode method, temperature by thermometer sensing technology and dissolved oxygen sensor by membrane electrode technique. Dissolved oxygen sensors collects fish ponds dissolved oxygen information and converts into electrical signals to provide the necessary condition for subsequent processing circuit. The system uses the

dissolved oxygen sensor with a rigid solid structure for automatically compensating sensor membrane permeability due to temperature changes and automatic pressure balance to prevent the diaphragm deformation and to provide material conditions for the accurate collection of information.

The data transformation and transmission is composed of the signal conditioning circuit, data acquisition board, core processing chip and communication module. The low level input signals directly from sensors are converted to serial digital output voltage standard in the range of 0-5V by high performance AD7705 converter chip. This chip uses sigma-delta conversion technique to achieve 16-bit code performance. Also, the device includes self-calibration and system calibration options to eliminate gain and offset errors of the device or system. The signal is then transmitted to the microcontroller (MCU) chip. The sensor nodes uses ATmega16L microcontroller as the core for data acquisition and processing. ATmega16 has an advanced RISC architecture, 113 instructions, 32 general-purpose registers and work at 16 MHz performance up to 16 MIPS, two cycle hardware multipliers, 8-channel 10-bit Analog to Digital Converter (ADC), 32 programmable Input/Output (I/O) ports, 16KB system programmable watchdog timer with independent on chip oscillator, power on reset, programmable power failure detection and on chip calibrated RC oscillator [24]. Then the MCU chip through a comparison with standard parameters issues control signals to drive relays for each actuator motor device. Data and control signals generated are aggregated and transmitted through ZigBee network to the gateway which is connected to the host computer base station. The host computer has priority, records these data for the user to query at any time according to the actual needs and issues commands through a gateway and ZigBee network to control the actuators in case environmental parameters are outside the preset threshold. The data processing microcontroller chip pack the data into the wireless transmission module and enables the gateway to receive the data and transmit them to a host computer for further analysis. The sensor nodes are powered by battery 5V DC. The CC2520 single chip of ZigBee processes the data and then communicates between antenna and gateway wirelessly. CC2520 is the second generation ZigBee/IEEE 802.15.4 RF transceiver in TI Company used for industrial, scientific and medical (ISM), 2.4 GHz band. CC2520 can work at 125 °C which provides the excellent sensitivity, connectivity and can work at low-voltage operation as well. CC2520 supports frame processing, data buffering, burst transmission, data encryption, data authentication, idle channel detection, link quality display and frame timing information thereby reducing the load of main controller.

B. Gateway

The gateway receives command packets, preprocesses and analyzes the data from the sensor nodes and then send to the host computer. The gateway is connected with the base station with serial RS-232 cable. In case the users have not yet received the response data from the

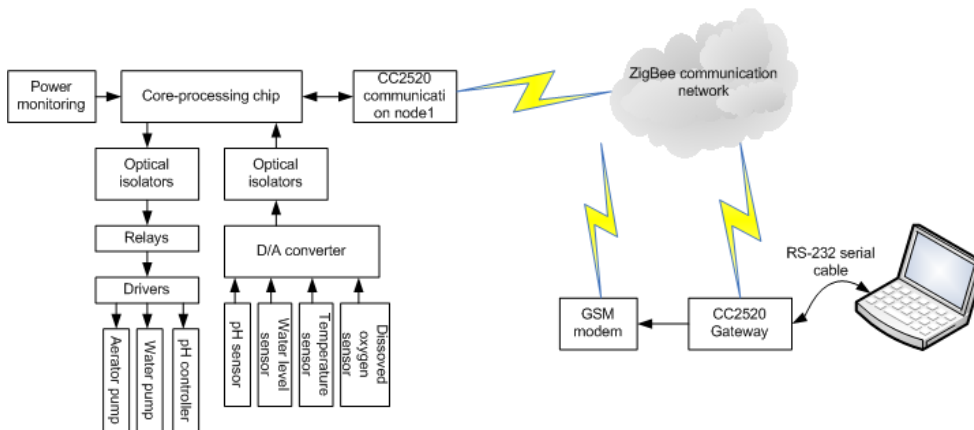


Figure 1. Overall system block diagram

sensor nodes within the time frame set, the sensor nodes will be considered to failure and then text messages will be sent to owners via the base station to notify about maintenance. The monitoring program is installed in the host computer base station that displays results and warns stakeholders through early short message warning. The environmental parameters of the fishpond will be sent to owners through short messages from the host computer via the global system for mobile (GSM) module. In view of the communication requirements between host computer and owners, the use of the GSM network to communicate not only reduces the cost, but also expands the communication range and space. Moreover, the base station sends data on a regular basis to the sensor nodes capture command when the transfer is complete and wait for the sensor node to return the data. If normal environmental parameters from the sensor nodes are received, the base station displays results on the display device. If otherwise the set value text messages are sent to the owners for notification and automatically open the corresponding actuator pump for correcting the respective environmental parameters. With the user-friendly interface, the host computer allows the owners to carry out a number of parameter settings to facilitate monitoring. Also is possible to set a manual command so as to achieve reasonable adjustment and control of systems diversity.

The gateway node comprises of 4 basic modules including communication module, RS-232/USB interface module, MAX-232 and power module. The communication module chosen for this module is the same with that in the sensor node, CC2520. The ZigBee CC2550 uses the serial peripheral interface (SPI) to connect the couple of control lines, and then connect to USB chip through a serial cable. This node receives power from the computer through the SPI bus, which ensures that the node is online all the time. Personal Computer (PC) is used as a processor in place of microcontroller. So in this work, microcontroller is not used at gateway node. However, PC performs the functionality of processor and is used to receive the data from the transmitting node and issue commands to control remote actuator devices.

C. Software Design

The software design of the sensor nodes is mainly carried out using ICCAVR compiler, one of the third-party C compilers recommended by ATMEL Corporation. ICCAVR comply with the ANSI C standard language to develop a suitable tool for the MCU program, easy to use, good technical support and basically has the following characteristics:

ICCAVR is an integrated editor and project manager integrated working environment (IDE).

The source files have all been organized to the project, document editing and project building, also in this environment the errors are displayed in the status window and when one clicks on the compile error, the cursor automatically jumps to the wrong line.

The project manager can also directly generate INTEL.HEX format, the format file most programmers are supported for downloading to the chip.

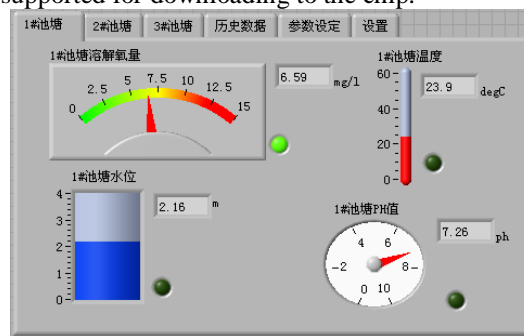


Figure 2. PC interface

The status of the system running the host computer is monitored in real time and can analyze, compute, display, print and process data. There are many platforms for development of PC monitoring screen such as KINGVIEW, C, C#, C++, and LabVIEW. In this study, LabVIEW is chosen for software realization. LabVIEW software user interface is as shown in Fig. 2. LabVIEW is a graphical G-language whereas the resulting program is in block diagram form. The production line technology staff can easily learn and use simple procedures which help to maintain, master and apply to practice in a very short period of time. On one hand, the host computer uses LabVIEW software to monitor and process water quality

data and on the other, to control actuator devices on remote site. The software can also be set for each of the ponds in order to ensure that the parameters in different species under different weather conditions have a suitable environmental growth [26]-[27].

D. Data Aggregation

Data aggregation, which combines data from multiple sensors, is performed in the firmware of the sensor nodes and monitoring program of the gateway. The network data aggregation can reduce the data packet size, the number of data transmissions and the number of nodes involved in gathering data from a WSN. The most dominating factor for consuming energy of WSNs is communication, i.e., transmitting and receiving messages [28]. Therefore, reducing generation of unnecessary traffics in WSNs enhances their lifetime. In addition, involving as many sensor nodes as possible during data collections by the sink node can utilize maximum resources of every sensor node. Aggregated result of sensor data at the sink node is used for making important decisions. Because WSNs are not always reliable, it cannot be expected that all nodes reply to all request. Therefore, the final aggregated result must be properly derived. For this, the information of the sensor nodes (Node Identifications, IDs) contributing to the final aggregated result must be known by the sink node. And the communication cost of transmitting IDs of all contributed sensor nodes along with the aggregated data must be minimized. Following are some promising reasons for transmitting IDs of sensor nodes along with their sensed data.

To know the exact picture of sensors data by identifying which sensor nodes are sending their data for data aggregation.

Data loss due to collision is inevitable in WSNs. Therefore, IDs of sensor nodes are needed to deal with data loss resiliency and accuracy of the final aggregated result of sensors data at the gateway node.

To know either a sensor node is providing service or not (survivability of a sensor node).

Hence, a gateway node must be aware of node IDs of those sensor nodes which contribute in aggregated value of sensors data in order to derive exact result of the collected data in WSNs. This is possible only when if there exists such a scheme which can transmit IDs of all the participating sensor nodes to the sink node. In this work, each sensor node has the capabilities of sensing, aggregating and forwarding data and it can send fixed-length data packets to the gateway node periodically. Finally, the sensor nodes can switch into sleep mode or a low power mode to preserve their energy when they do not need to receive or send data.

We used decentralized fusion architecture whereby data fusion occurs locally at each sensor node on the basis of local observations and the information obtained from sensors. This scheme has the advantage of scalability and tolerant to the addition or loss of sensing nodes or dynamic changes in the network. Due to their energy constraints, sensors need to perform efficient data fusion to extend the lifetime of the network. Lifetime of a

sensor network is the number of rounds of data fusion it can perform before the first sensor drains out. This is known as the Maximum Lifetime Data Aggregation (MLDA) problem.

Given: the location & energy of each sensor and the base station (BS). The goal is to find an efficient manner to collect & aggregate reports from the sensors to the BS [29].

System model

n sensor nodes (1..n)
Base station (n+1)
Fixed data packet size: k bits
Initial energy of a sensor i: ϵ_i
Receive energy,

$$RX_i = \epsilon_{elec} * k \quad (1)$$

Transmission energy,

$$TX_{ij} = \epsilon_{elec} * k + \epsilon_{amp} * d_{ij}^2 * k \quad (2)$$

where ϵ_{elec} is the electronic energy and ϵ_{amp} is amplifier energy.

Algorithm

Phase 1:

Sensors are grouped into clusters in a node.

Each sensor node consists of a minimum no. of sensors.

The energy of a sensor node is the sum of the energy of all the sensors within it.

Distance between two sensor nodes is the maximum distance between two sensors where, each resides in a different sensor node.

Apply the MLDA algorithm.

Software instruction-level parallelism (ILP) is employed to find a near-optimal admissible flow network.

Objective: maximize the lifetime of network (T) under the energy constraints.

Generate schedule(s) from the admissible flow network.

Phase Two:

Initialize {Aggregation Schedule} = 0.

Life Time, T = 0.

Choose a scheduler from phase 1.

Initialize aggregation tree, A with the BS.

Visit each clusters and add the nodes such that, the residual energy at each edge is maximized.

Add A to the aggregation scheduler.

Increment T by 1.

Repeat steps 3-7 until a node drains out.

Comments

Provides a set of data fusion schedules that maximize the lifetime of the network.

Clustering of nodes reduces the time needed to solve the ILP.

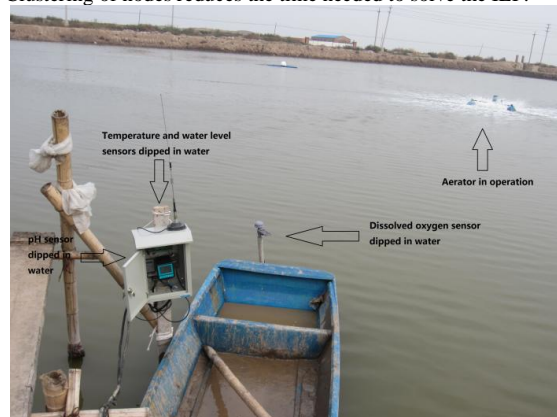


Figure 3. Sensor installation

E. Testing Environment

The system has been tested in Tanggu district at Tianjin intensive aquaculture base for six months (from

June 4, 2012 to November 25, 2012. The data presented here was taken during the whole testing time. The pond area is 2 acres divided into 4 ponds, pool 3 meters deep. Photos of the project setup and hardware testing for real time data acquisition, analysis and presentation are shown in Fig. 3, Fig. 4, and Fig. 5. Fig. 3 shows dissolved oxygen sensor installation at the Tangu fish pond. Furthermore, Fig. 4 shows the user interface device for setting user's mobile number and displaying values from sensor nodes and Fig. 5 shows dissolved oxygen aerator pump controlled automatically by the sensor nodes.



Figure 4. User interface device



Figure 5. Dissolved oxygen aerator pump working at Tangu fish pond

The GUI of the monitoring center allows monitoring data obtained from the sensor nodes and observe the behavior of the network in terms of quality of the radio link of each sensor node, data aggregation and battery status. The received signal strength is expressed in dBm. We used two nodes in the test, each being installed in different fish pond.

Considering the project requirements of cost, stability, accuracy, durability and other indicators, this system uses the following sensors:

Temperature sensor: DS18B20 thermometer, operating voltage range (3-5.5V), temperature range detection of $-55\sim+125^{\circ}\text{C}$ (-67 to $+257^{\circ}\text{F}$) and accuracy up to ± 0.5 degrees Celsius.

pH value sensor: PH400/450 series, pH display controller, the pH value measuring range -2.00 to 16.00 , pH resolution of 0.01 pH and accuracy is ± 0.01 pH.

Dissolved oxygen sensor: D-6800 intelligent dissolved oxygen detector, measuring range: $0\sim 20.00\text{mg/L}$, automatic range switching and temperature compensation: $0\sim 60^{\circ}\text{C}$, resolution: 0.01mg/L , precision: $\pm 0.5\%$.

Water level sensor: UXI-LY pressure type level transmitter, range of $1\sim 70\text{m}$, accuracy: $0.3\%\text{FS}$ and temperature range: $-10\sim 70^{\circ}\text{C}$.

IV. RESULTS AND DISCUSSION

The experimental results of sensor readings, battery performance and communication performance (signal strengths) recorded in every three minutes were monitored during the six months period. Fig. 6 shows monitored data of temperature with fluctuation by more than 5°C during a 24-hour period for the whole experimental duration. Two sets of data sampled automatically from node one and two placed in different fishponds have been compared. The curves correlate well but do not match, due to local biomass conditions in the fish ponds. During daylight hours, energy from the sun warms the water, while heat is lost to the cooler atmosphere at night. Wind and storms affect temperature by breaking up stratification, mixing the water and equally distributing the heat throughout the water column. Fig. 7 shows the monitored data of dissolved oxygen from node one and two respectively. The maximum values of dissolved oxygen recorded was 10.7 Mg/L and 9.93 Mg/L from node one and two respectively. Whereas, the minimum values of dissolved oxygen recorded was 4.50 Mg/L and 4.91 Mg/L from node one and two respectively. It can be noted that dissolved oxygen content did not fall below 4.5 Mg/L set even at night times. This can be explained by the fact that at this level the fishponds were sustained by aerators, thus meeting the objective of preventing fish mortality. These values demonstrate the ability of the controller to maintain the desired set points. Dissolved oxygen normally increases during day light hours when photosynthesis is occurring and decreases at night when respiration continues. Dissolved oxygen curves observed here from node one and two are different due to difference in aquatic animals and availability of phytoplankton in the ponds. It can be seen in Fig. 8 that the pH was relatively stable with standard deviation of ± 0.21 and ± 0.42 respectively. It can be observed that by being within range, there was no need for controlling the pH, since the values acquired by the software was within the preset range limit. The pH tends to decline in fishponds as bacteria produce acids and carbon dioxide is generated by the fish, algae and phytoplankton. Carbon dioxide reacts with water to form carbonic acid which drives the pH downward. Below a pH of 6.8 the nitrifying bacteria are inhibited and do not remove toxic nitrogen wastes. Optimum pH range in fishponds is maintained through the addition of alkaline buffers. The most commonly used buffers are sodium bicarbonate and calcium carbonate but calcium hydroxide, calcium oxide, and sodium hydroxide have been utilized. These curves change consistently and reasonably. The acquisition data reflects temperature, dissolved oxygen, pH and water level trend appropriately. These figures

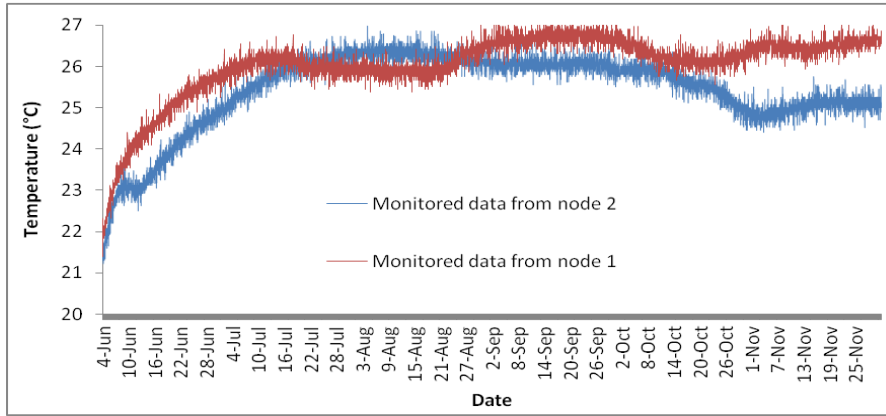


Figure 6. The monitored data of temperature collected from 4 June to 25 November, 2012

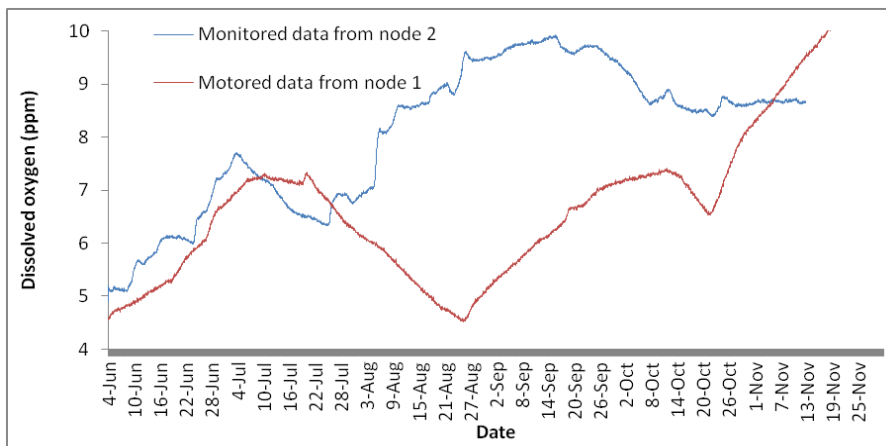


Figure 7. The monitored data of dissolved oxygen collected from 4 June to November, 2012

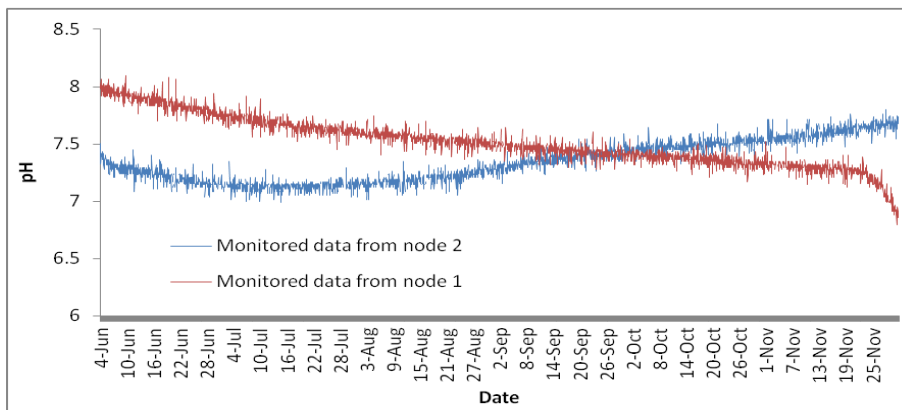


Figure 8. The monitored data of pH collected from 4 June to November, 2012

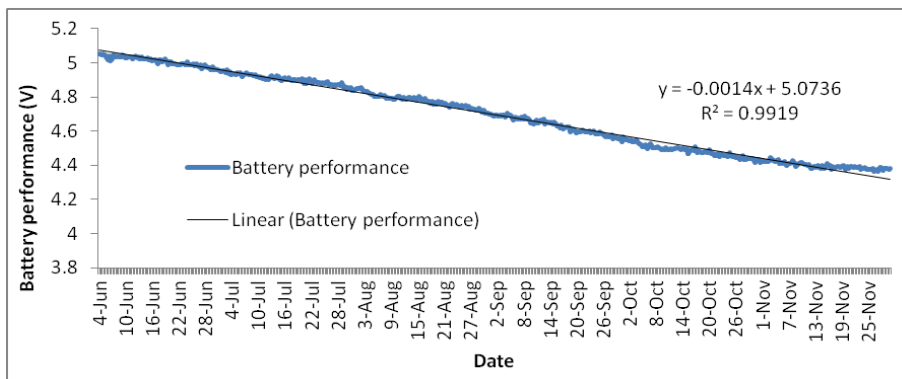


Figure 9. The monitored battery performance data collected from 4 June to November, 2012

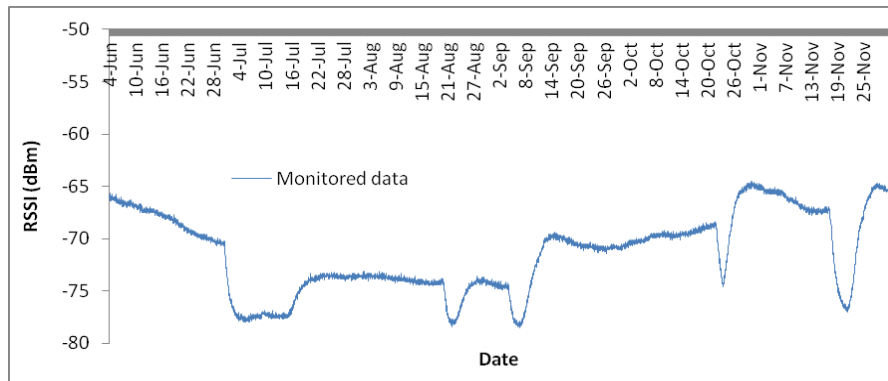


Figure 10. The monitored RSSI data collected from 4 June to November, 2012

show the correctness and feasibility of the fish pond monitoring system. The detailed changes of water level measured by our proposed system are in the same way with satisfactory results.

Sensor nodes were powered with Duracell alkaline battery size Lantern, number MN908, 6V 11.5 Ampere hours. The battery level of all the sensor nodes stayed very stable in the range of 5.05V to 4.39V for the whole experiment duration (Fig. 9). Regression analysis of the battery performance experimental data with a linear fit gave a determination coefficient, $R^2 = 0.9919$, showing that a close relation between the values exists. Part of the power is consumed by the sensors, microcontroller ATmega16L, and the communication module CC2520. The actuators and the gateway are powered by the mains. Since the sensor nodes are operated with batteries, the power supply is very limited and so power saving is therefore of utmost importance in designing, implementing and operating a WSN based monitoring system. In our case, energy consumption is reduced by using low power hardware (sensors, microcontrollers, radio chips) for implementing sensor nodes that consume typically significantly less energy. The hardware and software presented here are designed specifically to address the needs of WSNs namely efficiency power consumption, low cost and scalability that integrates detection, processing and storage. The batteries are unable to supply enough current to power the node once the voltage drops below 2.5V.

The Received Signal Strength Indicator (RSSI) curve for the whole experimental duration is shown in Fig. 10. A low RSSI value represents a bad radio link, a high value a good radio link. The typical receiving sensitivity is -94 dBm. In general the signal strength that reaches the receiver antenna is dependent on the orientation of the antenna and the distance between the transmitter and receiver. This represents an additional source of error which may significantly influence the accuracy of the received signal strength. In order to achieve high signal strength the antennas should be placed in line of sight at distance $\leq 1/r^2$ for avoiding attenuation of antenna signal power in outdoor environment (where r is the distance between transmitter and receiver). However, in real-world environments, this indicator is highly influenced by noises, obstacles, and the type of antenna, which makes it hard to model mathematically. In this case it is important

to make a system calibration, where values of RSSI and distances are evaluated ahead of time in a controlled environment.

Similar findings with a remote wireless system for water quality web based monitoring in intensive fish culture were reported by [5, 17]. Rather precise and constant regulation of dissolved oxygen, temperature, water level and pH has been achieved by this system. For example, in six months test of using this system, these environmental parameters were kept at optimal levels where almost all aquatic organisms can survive indefinitely provided other environmental parameters are within allowable limits. Whereas the fish are reasonably comfortable and healthy at 5-6 mg/L dissolved oxygen concentrations, which is in agreement with our findings. Ideally, fish ponds should be at or near oxygen saturation at all times. This system has a structure of receiving and storing water quality information sent from respective sensors in real time and links it with GSM module so that the user can have access to fish pond status at any place in time. Information stored in host computer can be displayed as a graph, with which a user can understand the status of the fish pond in real-time, and user can take corrective action for any possible problems at proper time.

Continuously monitoring real-time environmental parameters and alarms will automatically notify the user of any out of bounds condition that could signal an equipment failure, improper settings, extreme weather conditions etc.

V. CONCLUSIONS

This study provides the design of water quality monitoring and control system for aquaculture based on wireless sensor networks and single chip computer technology as a base in the actual operation. It realizes the monitoring of the water environmental parameters for intensive aquaculture and alarm notification through short message when monitored variables take anomalous values and is suitable for long-term stability under growth conditions thus increasing yield per unit area. The system can monitor the data of temperature, dissolved oxygen, pH, and water level continuously and in real-time. Two nodes have been implemented for six months to evaluate the system feasibility. The sensor data, battery performance and network performance metrics have been analyzed and presented. The pump can be set to auto-start

according to the parameter values and avoid common fisheries problems such as dissolved oxygen depletion due to high water temperature, cloudy weather and pond turn over's hence avoiding fish kills. The pump working hours will be greatly reduced thus efficient energy consumption and reducing labor cost.

Future works should be enhancing the system remote access to the sensor nodes using internet and data transmission for further analysis. More network performance metrics need to be studied and evaluated to make the system more robust and scalable.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 60771014 and Tianjin Agricultural Scientific Committee Foundation under Grant 201101190. The authors would like to thank members of Tangu fish pond demonstration base for the assistance on the setting up of the experiment and collection of water quality data.

REFERENCES

- [1] Phillip G. Lee, 1995. A Review of Automated Control Systems for Aquaculture and Design Criteria for Their Implementation, *Aquacultural Engineering*, Vol. 14, No. 3, pp. 205-227.
- [2] P. Fowler, D. Baird, R. Buklin, S. Yerlan, C. Watson & F. Chapman, 1994. Microcontrollers in Recirculating Aquaculture Systems, University of Florida, EES-326.
- [3] W. J. S. Mwegoha, M. E. Kaseva and S. M. M. Sabai, 2010. Mathematical modeling of dissolved oxygen in fish ponds, *African Journal of Environmental Science and Technology*, Vol. 4(9), pp. 625-638. [Online] Available: <http://www.academicjournals.org/AJEST>.
- [4] Summerfelt, Robert C. (n.d.) Water Quality Considerations for aquaculture, Aquaculture Network Information center, (<http://aquanics.org>).
- [5] Yang Shifeng, Ke Jing, and Zhao Jimin, 2007. Wireless monitoring system for aquaculture environment, in *Proc. Radio-Frequency Integration Technology, RFIT 007, IEEE*, pp. 274-277.
- [6] Yang Shifeng and Li Yang, 2011. Dissolved oxygen remote monitoring system based on the internet, *Electronic Measurement Technology*, (7): 88-90.
- [7] Xiuna Zhu, Daoliang Li, Dongxian He, Jianqin Wang, Daokun Ma, Feifei Li, 2010. A remote wireless system for water quality online monitoring in intensive fish culture, *Computers and Electronics in Agriculture*, (715), 53-59. www.elsevier.com/locate/compag.
- [8] Seungjoon Lee, Bennett L. Ibey, Gerard L. Coté Michael V. Pishko, 2008. Measurement of pH and dissolved oxygen within cell culture media using a hydrogel microarray sensor, *Sensors and Actuators B*, 128, 388-398. [Online]. Available: www.sciencedirect.com.
- [9] Yanle Wang, Changsong Qi, Hongjun Pan, 2012. Design of Remote Monitoring System for Aquaculture Cages Based on 3G Networks and ARM-Android Embedded System, *Procedia Engineering*, 29, 79-83. [Online]. Available: www.sciencedirect.com.
- [10] Stankovic, J., 2008. When sensor and actuator networks cover the world, *ETRI Journal*, 30(5), 627-633.
- [11] Luis Ruiz-Garcia, Loredana Lunadei, Pilar Barreiro and Jose Ignacio Robla, 2009. A review of wireless sensor technologies and applications in agriculture and food industry: *state of the art and current trends*, *Sensors*, 9(6), 4728-4750. doi:10.3390/s90604728.
- [12] J.A. López Riquelme, F. Soto, J. Suard áz, P. Sánchez, A. Iborra, J.A. Vera, 2009. Wireless Sensor Networks for precision horticulture in Southern Spain, *Computers and Electronics in Agriculture*, 68, 25-35.
- [13] Xiaoqing Yu, Pute Wu, Wenting Han, and Zenglin Zhang, 2012. The research of an advanced wireless sensor networks for agriculture, *African Journal of Agricultural Research* Vol. 7(5), pp. 851-858.
- [14] Francisco J. Epinosa-Faller, Guillermo E. Rendon-Rodriguez, 2012. A ZigBee wireless sensor network for monitoring an aquaculture recirculating system. *Journal of Applied Research and Technology*, Vol. 10 (3), 380-387.
- [15] Xingqiao Liu and Liqiang Cheng, 2012. Wireless sensor network based on ZigBee in aquaculture, *Advances in Intelligent and Soft Computing*, Vol. 148, pp. 553-558.
- [16] Ma Daokun, Ding Qisheng, Li Daoliang, Zhao Linlin, 2010. Wireless Sensor Network for Continuous Monitoring Water Quality in Aquaculture Farm, *Sensor Letters*, Volume 8, Number 1, pp. 109-113(5).
- [17] Qi Lin, Zhang Jian, Mark Xu, Fu Zetian, Chen Wei, Zhang Xiaoshuan, 2011. Developing WSN-based traceability system for recirculation aquaculture. *Mathematical and Computer Modelling*, (53), 2162-2172.
- [18] Mingfei Zhang, Daoliang Li, Lianzhi Wang, Daokun Ma, Qisheng Ding, 2011. Design and Development of Water Quality Monitoring System Based on Wireless Sensor Network in Aquaculture, *IFIP Advances in Information and Communication Technology*, Volume 347, pp 629-641.
- [19] Liu H., Liao G., Yang F., 2008. Application of wireless sensor network in agriculture producing Changsha: *Institute of Agricultural Information, Hunan Agricultural University*, 2008, v. 11.
- [20] Haifeng W., Binglian W., Xianglong K., Qiang G., 2008. Agricultural environment measure system based on Zigbee *Network and Algae Cell Sensors. In: Control Conference*, 27, Kunming. *Proceedings*. pp 209 – 213.
- [21] Soledad Escolar D áz, Jesús Carretero Pérez, Alejandro Calderán Mateos, Maria-Cristina Marinescu, Borja Bergua Guerra, 2011. A novel methodology for the monitoring of the agricultural production process based on wireless sensor networks, *Computers and Electronics in Agriculture*, Volume 76, Issue 2, Pages 252-265.
- [22] Xin Wang, Longquan Ma, Huizhong Yang, 2011. Online water monitoring system based on ZigBee and GPRS, *Procedia Engineering*, (15), 2680-2684.
- [23] Tai Haijiang, Liu Shuangyin, Li Daoliang, Ding Qisheng, Ma Daokun, 2012. A multi-environmental factor monitoring system for aquiculture based on wireless sensor networks, *Sensor Letters*, Vol. 10 (1-2), 265-270 (6).
- [24] Atmel Corporation, 2010. ATmega16-16L datasheet, 2466T-AVR-07/10.
- [25] Claude E. Boyd and Terry Hanson, 2010. Dissolved-Oxygen Concentrations In Pond Aquaculture, *Global aquaculture advocate*, January/February 2010.
- [26] Lopa Ghosh and G. N. Tiwari, 2008. Computer Modeling of Dissolved Oxygen Performance in Greenhouse Fishpond: An Experimental Validation, *International Journal of Agricultural Research*, 3 (2): 83-97.
- [27] Ilan Halachmi, Yitzchak Simon, Rami Guetta, Eric M. Hallerman, 2005. A novel computer simulation model for design and management of re-circulating aquaculture systems, *Aquacultural Engineering*, 32, 443-464.
- [28] Rabindra Bista and Jae-woo Chang, 2010. Energy efficient data aggregation for wireless sensor networks, sustainable

wireless sensor networks, Yen Kheng Tan (Ed.), ISBN: 978-953-307-297-5, In Tech, Available from: <http://www.intechopen.com/books/sustainable-wireless-sensor-networks/energy-efficient-dataaggregation-for-wireless-sensor-networks>.

- [29] K. Dasgupta, K. Kalpakis and P. Namjoshi, 2003. An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks, *IEEE WCNC*.

Daudi S. Simbeye received the B. E. degree in electronics and microelectronics from Moscow Power Engineering Institute (Technical University) and the M. E. degree in design and technology in electrical engineering from Southern Federal University-Taganrog Institute of Technology, Russia, in 2006 and 2008 respectively. Currently, he is pursuing PhD in light industry information technology and engineering at Tianjin University of Science and Technology, China. From 2008 to date, He is an assistant lecturer in the department of computer studies at Dar Es Salaam Institute of Technology, Tanzania. His areas of interest include embedded systems, automation and control, and Wireless sensor networks.

Shi-Feng Yang received the B. E. degree in agriculture mechanization and M. S. degree in computer application in agricultural engineering from Hebei Agricultural University (Baoding city) in 1981 and 1991 respectively, and the Ph.D degree in computer application in agricultural engineering from China Agricultural University (Beijing) in 1998. From 1992 to 1998, he was an associate professor in the department of mechanical and electrical engineering at the Hebei Agricultural University. From 1998 to 1999, he was a professor in the department of agricultural automation engineering at the China Agricultural University. Since 1999 to date, he is a professor in the department of automation at Tianjin University of Science and Technology, China. In 1994 he received the fourth award of science and technology of Hebei Province for microcomputer based on instrumentation system to measure seedmeter performance. In 1995 he received the first award of Hebei Agricultural University for “a system for monitoring and controlling the growing environment parameters in the greenhouse”, and in 1997 he received the agricultural engineering prize. His areas of interest include research of information detection and intelligent technology.

An NFC-based Scenic Service System

Jie Ma and Jinlong E*

College of Software, Nankai University, Tianjin 300071, China

*Corresponding author, Email: majie1765@nankai.edu.cn, ejinlongnk@163.com

Abstract—Based on the characteristics that NFC devices can read and write specific tags and readers, three programs are proposed using NFC mobile phones to interact with the deployed tags and readers for acquiring information of the current attraction and attractions nearby. A kind of check-in and scoring mechanism is also designed which can be efficient and effective for users to upload scores for popular attractions recommendation, and its correctness has been verified by formula derivation. Then a cross-platform interactive NFC-based service system is designed and implemented on platforms of Android phones, servers and PC clients according to these programs and the mechanism. By testing the time consumption of each operation and comparing three programs, the system performance proves to be fine and can meet actual use.

Index Terms—Near Field Communication (NFC); Location Based Services (LBS); Tags; Information Push Services (IPS)

I. INTRODUCTION

With the popularity of smart phones and the Internet of Things (IOT), Near Field Communication (NFC), a new technology, has been adopted as the basic configuration of the system on more and more smart phones. The development of this technology makes the idea of integrating the function of smart RF cards into mobile phones possible. Nowadays, more and more people start using LBS (Location Based Services). This business has gradually developed into an essential part of the mobile Internet. One of its key factors is that the access to the user's current location, which needs some kinds of positioning technology, should be considered first before a series of services based on location are carried out. When it comes to positioning technology, the most well-known is GPS (Global Positioning System), which is used as the main positioning method in most LBS applications such as Google Maps, Foursquare, etc. But due to the dependence of the satellite signal, it is almost impossible to use GPS in an indoor environment. Positioning based on network such as cellular positioning and WiFi positioning can be used both indoors [1] and outdoors [2]. But cellular positioning accuracy is too low, and generally used as the alternative of GPS. WiFi positioning performs well in the building where hotspots are deployed and signals of all reference points are collected beforehand. Microsoft RADAR system [3] is a successful application in this respect. But when users are far away from the building, they can't obtain a good performance again due to the weakness of the wireless network signal.

Tourist services have now developed into a popular LBS application. When tourists visit the various attractions in a scenic spot, they all hope to keep abreast of detailed descriptions of the current attraction as well as popular attractions nearby. The traditional GPS positioning uses latitude and longitude to identify the location, but these absolute coordinates are not readily convertible to the relative position of the various attractions within a small area. The WiFi positioning based on the signal strength fingerprint map is commonly used in general scenic spots. This positioning method requires the prior acquisition of the WiFi signal strength in each attraction and formation of the fingerprint map. When a user needs to position with his phone, the attraction with the closest signal will be selected as the current position. In most scenic spots, the devices that are used to explain the information of attractions are designed based on this positioning method. But there is a large error in this method, and the immediate result is that the explainer begins to introduce the relevant information when there is still some distance to the attraction. In addition, most explainers are designed to introduce each attraction's information once only. These factors make users visiting the attractions unsynchronized with acquiring the information, and the desired effect cannot be reached. If NFC technology can be used in the scenic service system, and a number of NFC tags and NFC readers connected to the terminal systems are deployed in attractions, it is not only easy for users to acquire information of attractions, but also easy to check in and score with interactions of phones and readers, to recommend popular attractions nearby. Some studies, such as [4], [5], design mobile guide systems, which make smart posters can be placed near the collection items to deliver multimedia contents to users. Some people present a design of a multimedia mobile guide for visitors of a museum, with NFC technology applied to enhance UX (User Experience) of the system [6]. However, few of them propose specific programs with NFC tags, readers and mobile phones. This paper will present three programs applying NFC tags and NFC readers to the scenic service system and a scoring mechanism, implement the "NFC-based scenic service system" combining these programs and the mechanism, as well as test and compare the time consumption of each program.

The rest of the paper is organized as follows: Section II describes key technologies in the system, proposing three programs of acquiring attractions' information, system

backstage management and the scoring and recommendation mechanism. Section III presents the design and implementation of the system. Section IV analyzes and compares the results of performance in experiments. Section V, the last section, summarizes the paper and gives an outlook of the future.

II. DESCRIPTIONS OF KEY TECHNOLOGIES IN THE SYSTEM

A. Characteristics of NFC

NFC technology is proposed by Philips, SONY, Nokia, etc. It is a new short-range wireless communication technology, which evolves from the combination of Radio Frequency Identification (RFID) technology and traditional near field interconnection technologies such as Bluetooth, WiFi, etc. This technology makes two devices communicate with each other by touching in a very close range (about 10cm). It works in the 13.56MHz band, with transmission a rate such as 106kb/s, 212kb/s and 424kb/s which can be chosen. Compared to RFID and other near field interconnection technologies, it has the characteristics of near transmission distance, high bandwidth, low power consumption, etc. NFCIP-1, identified as Standard ISO/IEC 18902, elaborates control principles of NFC devices [7]. NFCIP-2, identified as Standard ISO/IEC 21481, defines a flexible gateway system to detect and select the 3 operating modes of the NFC technology: tag-emulation mode, reader/writer mode, and peer-to-peer communication mode [8]. In this way, NFC devices can be used as electronic tickets and electronic wallets, and they can read smart posters and transmit data peer-to-peer by touching.

NFC is applied for communication in a very short distance (10cm~20cm). Such a short distance limits the potential eavesdropping and access by hackers. Therefore, the technology has a very high security. In addition, NFC logic link layer also includes an encryption and authentication procedure and an anti-collision mechanism. It can choose the only target to communicate in the initialization process, to exclude the third-party from controlling the link as the role of "middleman". In the case of sensitive applications such as mobile payment in the tag-emulation mode, the AES encryption algorithm and Triple DES encryption algorithm can also be added, which are adopted by the standard smart card, to the upper application [9].

B. Three NFC-based Attractions' Information Interaction Programs

As a scenic service system, the most important function is users' access to information of attractions. Taking use of NFC mobile phones interacting with NFC tags and readers, three attractions' information interaction programs are proposed, namely "reading attractions' information from NFC tags" ("reading NFC tags"), "reading attractions' information from NFC readers" ("reading NFC readers"), "writing identity to NFC readers and waiting for attractions' information pushing" ("writing NFC readers"). With the mutual collaboration

of these three programs, the system can have stronger service capabilities, and enhance users' experience effects.

1) "Reading NFC Tags" Program

For this program, NFC tags should be hung or posted at each attraction. It is necessary to write the current attraction's name, BSSID of the scenic WiFi LAN, LAN password (in WPA2 mode), the IP addresses and port numbers of the system server both within the WiFi LAN and by the cellular network through router address mapping to the tags by NFC devices. After a user's NFC mobile phone reads the NFC tag, the mobile client application can first try to make a request to the scenic WiFi hotspot for joining the WiFi LAN with the acquired BSSID and password (if required). Then the WiFi hotspot authenticates the mobile phone, and assigns DHCP dynamic IP address to it. After the mobile phone joins the WiFi LAN, it can acquire information of attractions from the server with the IP address and port number of the server in the WiFi LAN. If the mobile phone fails to join the scenic WiFi LAN, the cellular network can be used to access the server with the IP address and port number mapped to the external network in the tag. In this program, the communication between mobile phones and tags belongs to the "reader/writer" mode which is one of the three NFC communication modes from the perspective of mobile phones. Due to the limited capability of a tag (usually only 1KB), it cannot save a lot of information but only an attraction's name. Users can further inquire the detailed descriptions of the current attraction and information of nearby recommended attractions from the server with the name in the tag. A user's operation can be described by the flowchart in Figure 1.

2) "Reading NFC Readers" Program

For this program, a terminal system is set up to connect NFC readers in each attraction, and the readers are set as writing mode, i.e. when RFID cards or NFC phones come close, the readers can write data to them. Unlike the "reading NFC tags" program, this program maintains the attraction's information in the terminal system rather than pre-writes messages to the readers, which makes the data can be pushed to NFC phones in real time. The communication between mobile phones and readers in this program seems to belong to the "tag-emulation" mode, but actually it still belongs to the "peer-to-peer communication" mode, i.e. two devices communicate through a specific form of commands encapsulated in NDEF (NFC Data Exchange Format) [10]. In the interaction, the reader transmits the attraction's information, and the format of interactive commands varies due to the reader. No longer limited by the capability, the detailed descriptions of the current attraction and information of nearby recommended attractions can be transmitted to the user's mobile phone together with the required network parameters. In this way, users do not have to further inquire information of attractions from the server, which reduces interaction time consumption as well as the server's response payload. After users get information of attractions, they can also join the scenic WiFi LAN with the acquired

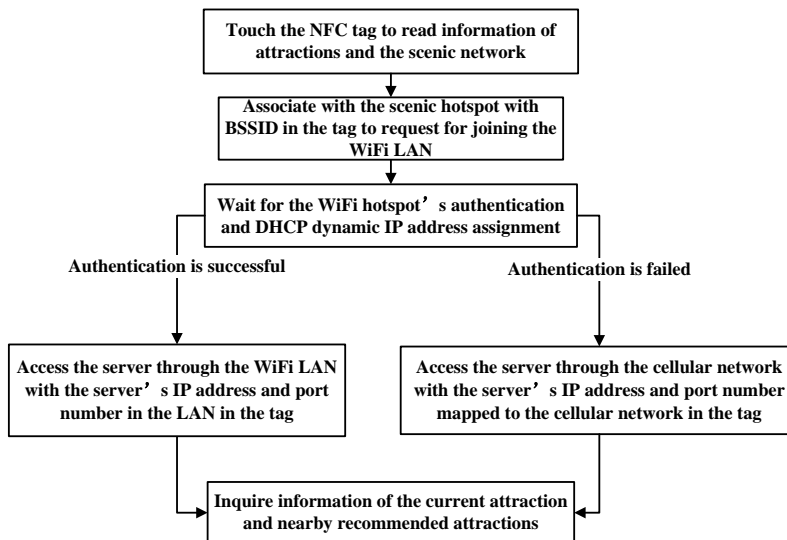


Figure 1. The flowchart of a user's operation in the "reading NFC tags" program

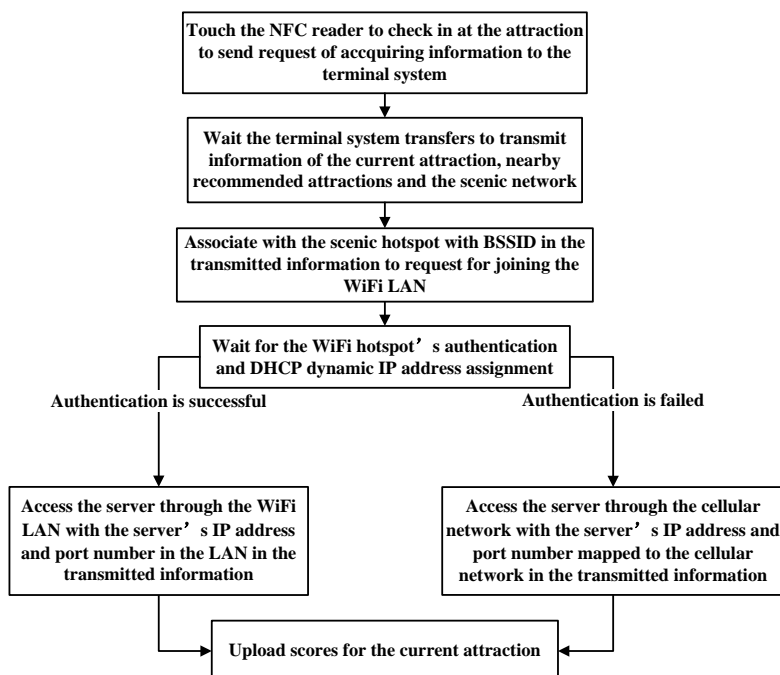


Figure 2. The network architecture of system

network parameters, and access the server through WiFi LAN or the cellular network to score for the current attraction. Compared with the off-line "reading NFC tags" program, another advantage of this online program is that it can collect the users' access number as the popularity of an attraction in order to recommend attractions to users. Specifically, the system uses a parameter to save the access number, which adds 1 when the attraction's information is written to the user's mobile phone successfully every time. The access number can be written to a particular file when there is a need to suspend the service in order to maintain the system, and then it is read into the system again and continues to accumulate when the service is restarted. The above process can be represented by the flowchart in Figure 2.

3) The "Writing NFC Readers" Program

For this program, each attraction also needs a terminal system connecting with NFC readers. The readers are set as reading mode, and wait for users' mobile phones writing identity information to it. NFC phones and readers interact through specific formatting commands. The user's mobile phone first reads its phone number. The IP address will also be read from the phone system and a random unused port number will be generated if the phone has joined the WiFi LAN. All these parameters need to be encapsulated into specific format and written to the readers as the user's identity. The terminal system in each attraction maintains an ID file, which takes the user's phone number and the current date as an ID, to reflect the user's access in a day. For the first time when a user's ID is written to the system, the user's information is uploaded to the server, and the system determines

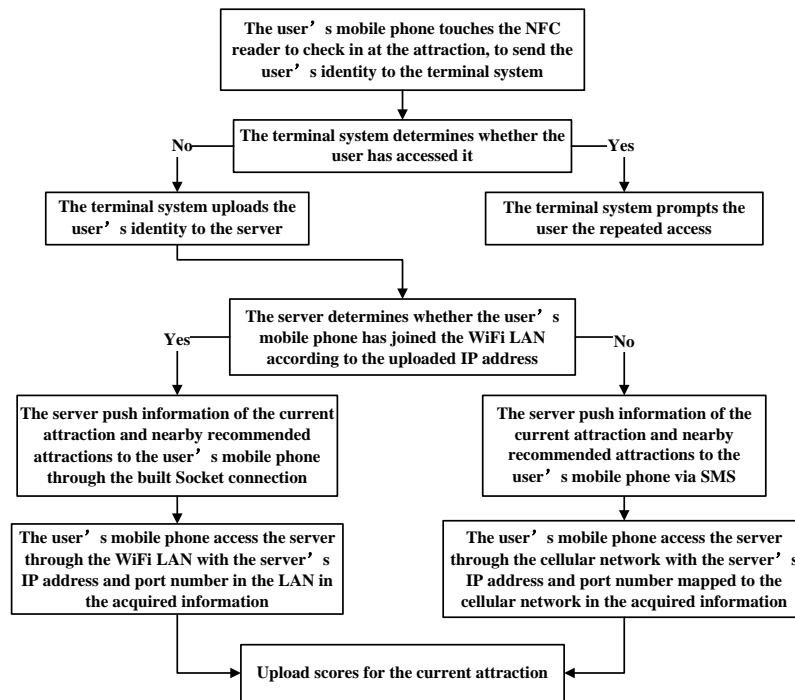


Figure 3. The flowchart of the “writing NFC readers” program

whether the phone has been joined the WiFi LAN depending on whether there are legal IP addresses. For the mobile terminal which has joined the LAN, it can be used as Socket server waiting for a connection. The system server, as Socket client, connects the phone via TCP with the uploaded IP address and port number, to push information of attractions to the user. By contrast, for the mobile terminal not joined the LAN, the server will push information of attractions, the IP address and port number available from the cellular network to the user via SMS. If the user has written his or her identity to the reader on the same day, he or she will be prompted that information of attractions has been already pushed and repeat push is not available. Such a mechanism can effectively avoid over-loaded response caused by someone’s malicious check-ins. Similar to the “reading NFC readers” program, the terminal system collects the users’ access number every time they check in. The different part is that no further increase in the access number for a duplicate user ID. After getting information of attractions, the user can choose WiFi or the cellular network and use the corresponding IP address and port number to access the server and score for the current attraction. The process of the program can be described as the flowchart in Figure 3.

C. System Backstage Management

The backstage management of the system involves many aspects, and the most important two are described as follows.

1) Initial Information Collection through Mobile Terminals

In the process of initializing the system, administrators collect information in various attractions under the circumstance that their mobile terminals have joined the scenic WiFi LAN and they know the IP address and port

number of the server. They need to upload each attraction’s name, introduction, detailed descriptions, IP address and port number as well as the distances and routes between adjacent attractions and other information to the server, and add them to the database. It is also necessary to write each attraction’s name and network parameters of the WiFi hotspot and the server to NFC tags in the respective attraction in a specific format through NFC devices, which is useful for the “reading NFC tags” program.

2) Periodic Maintain of Terminal Systems

The “reading NFC readers” program needs each terminal system to maintain an attraction’s information file, which provides the attraction’s descriptions, the average score, the access number, nearby attraction’s distances, routes and other information to users through readers. As users’ access and scoring, each attraction’s access number and average score will be constantly changing, and these changes are maintained by the server in real time. The attraction’s information file in each attraction will be outdated and cannot reflect the current situation of the attraction, which affects users’ judgment. Therefore, the server needs to update information of attractions maintained by all terminal systems periodically. This system sets timers in both the server and terminal systems, and select a period (such as 2~3 a.m.) per day as the maintenance period. During this period, all terminal systems suspend interactions with mobile phones, and act as Socket servers waiting for TCP connection from the system server. The system server will be triggered by the timer, to query each attraction’s information stored in the database, and transmit these data to each attraction in turn via TCP according to the stored each attraction’s IP address and port number. After a terminal system receive the information and update the

information file, it will resume the normal services. In addition, the updated users' access number is also transmitted through the periodically Socket communication in the "reading NFC readers" program. By contrast, in the "writing NFC readers" program, users' access number is uploaded to the server together with a user's information every time a user accesses the attraction.

D. Attractions' Scoring and Recommendation Mechanisms

As described in the previous subsection, this system can aggregate users' access number in each attraction, and allows users to score for attractions. In this case, similar to scoring for goods on a shopping website, attractions' scoring and recommendation mechanisms can be added to the system.

1) Scoring Mechanisms

For the scoring part, some specific mechanisms are mainly used as follows:

a) A user can select a number from 1 to 5 as the score for an attraction and the default score is 3. It is necessary to average the scores of all users who access an attraction as the average score of the attraction. An attraction's recommendation will depend on both the average score of the attraction and users' access number which is regarded as the attraction's popularity.

b) Since all attractions adopt those three attractions' information interaction programs at the same time and the latter two programs can aggregate users' access number, users who acquire information of attractions by these two programs are allowed to score for attractions and upload the scores to the server.

c) The terminal system in an attraction is connected with two NFC readers, which aggregate users' access numbers with the latter two programs respectively. The system takes the sum of these two access numbers as the total users' access number, i.e. the attraction's popularity.

d) In order to ensure the attraction's average score and popularity showed to the first user are not 0, avoiding the system's "cold start-up", administrators can score for each attraction once as the initial score when they collect and upload information of attractions and set both two NFC readers' access numbers to 1.

e) The specific scoring process is as follows: When a user access an attraction, the terminal system uploads the default score (i.e. 3) to the server. If the user scores for the attraction thereafter, the difference between the new score and 3 will be uploaded to the server to update the average score of the attraction stored in the database. Otherwise, the user's score for the attraction is regarded as 3.

The formula for calculating the average score can be deduced according to the mechanisms. The total users' access number here is set as n , in which the access number by reading a reader is n_1 , and the access number by writing a reader is n_2 , i.e. $n = n_1 + n_2$. The i -th user's score is S_i , wherein $i = 1, 2, \dots, n$. Then the average score of the attraction is

$$S_{avg} = \sum_{i=1,2,\dots,n} S_i / n \tag{1}$$

When the terminal system uploads the updated users' access number to the system server, for the "reading NFC readers" program, if the updated users' access number is n_1' , then the updated average score of the attraction is

$$S_{avg}' = [(n_1' - n_1) * 3 + S_{avg} * n] / (n_1' - n_1 + n) \tag{2}$$

And the users' access number by reading readers is updated to n_1' , i.e. the total users' access number is $n' = n_1' + n_2$. Similarly, for the "writing NFC readers" program, if the updated users' access number is n_2' , then the updated average score of the attraction is

$$S_{avg}' = [(n_2' - n_2) * 3 + S_{avg} * n] / (n_2' - n_2 + n) \tag{3}$$

And the users' access number by reading readers is updated to n_2' , i.e. the total users' access number is $n' = n_1 + n_2'$. To sum up, the formula

$$S_{avg}' = (S_{avg} * n + 3 * \Delta n) / (n + \Delta n) \tag{4}$$

can represent the updated average score, wherein $\Delta n = n_1' - n_1$ or $\Delta n = n_2' - n_2$ is the added users' access number. When a user uploads his or her score for the attraction to the server, the attraction's average score will be updated as

$$S_{avg}'' = (S_{avg}' * n' + S_k' - 3) / n' \tag{5}$$

where in S_k' is the k -th user's new score. This formula is used to re-calculate the average score whenever a user uploads a score.

If there are m more users who access an attraction, these m users' score are S_i' respectively, wherein $i = 1, 2, \dots, m$. The new access number will be $n' = n + m$. Based on Formula (1), after these users upload their scores, the updated average score of the attraction is

$$S_{avg}''' = (S_{avg}' * n + \sum_{i=1,2,\dots,m} S_i') / (n + m) \tag{6}$$

Combining Formula (4) and Formula (5), it can be verified that the two-time scores uploading in the proposed scoring mechanisms have the same effect as updating the average score in one time, and the former is more practical. Based on Formula (4),

$$S_{avg}''' = (S_{avg}'' * n' + \sum_{i=1,2,\dots,m} S_i' - 3 * m) / n' \tag{7}$$

can be deduced. When Formula (5) is put into Formula (7), Formula (6) can be deduced as follows:

$$S_{avg}''' = [(S_{avg}'' * n + 3 * m) / n' * n' + \sum_{i=1,2,\dots,m} S_i' - 3 * m] / n' \\ = (S_{avg}'' * n + \sum_{i=1,2,\dots,m} S_i') / (n + m)$$

Through the above derivation, the correctness of the proposed scoring mechanisms can be verified.

2) Recommendation Mechanisms

When users acquire information of attractions by those three programs, nearby attractions within the selected

range are sorted based on the average score and the users' access number and showed to users. This system takes the product of the average score and the users' access number (i.e. the total score) as the sort criteria of nearby attractions, so a higher total score means a higher degree of recommendation. More users accessing an attraction means the attraction is much more popular and more worth of recommendation compared with a few users giving high scores. Therefore, sorting the nearby attractions according to the product (i.e. the total score) adopted by the system can evaluate the attractions objectively.

III. DESIGN AND IMPLEMENTATION OF THE "NFC-BASED SCENIC SERVICE SYSTEM"

A. Design of System Framework

This system is used to provide users' mobile clients with information of attractions in a scenic spot and recommends nearby popular attractions. With the system, it is easy for users to visit the attractions they are interested in on purpose according to the prompted routes in a large scenic spot and it can prevent users from getting lost or wandering aimlessly in the scenic spot. As a consequence, it will save users' time and energy, and make a good scenic service experience for users. The system includes two aspects: acquiring information of attractions and scoring for the current attraction. There platforms are used, namely the server, terminal systems in attractions and users' mobile clients.

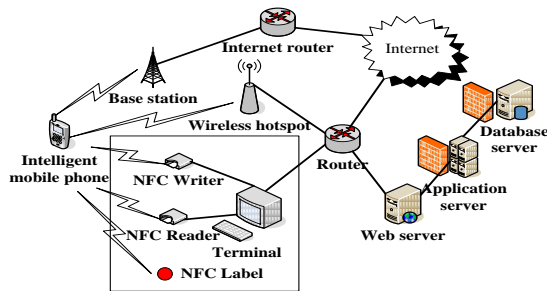


Figure 4. The network architecture of system

Due to the limitation of embedded devices' computing power, storage capacity and battery power consumption, the work of "inquiring information of the current attraction and nearby recommended attractions", "updating attractions' scores" and "transmitting updated information to each attraction periodically" should be completed on the server. Terminal systems in attractions complete the work of "pushing information of attractions to users' mobile phones", "verifying users' identities" and "transmitting users' information to the server". Mobile clients only need to collect and acquire attractions' information, and upload scores for attractions. The server can be a physical server, but it had better be divided into front-end web server, application server used for positioning business, and back-end database server, which is more conducive to the load balancing. The client can be connected to the Internet via the scenic WiFi LAN or the cellular network to access the server and terminal systems, and the three platforms transfer information by

HTTP, Socket and NFC peer-to-peer communication. The overall network architecture is shown in Figure 4.

B. Implementation of Three Platforms

1) Mobile Client Platform

A mobile client accesses the remote server and terminal systems with NFC readers to acquire information of the current attraction and nearby recommended attractions and upload a score for the current attraction. The client system can be divided into two modules in the role of administrators, namely, "collecting information of attractions" and "collecting route information between adjacent attractions", and three modules in the role of users, namely, "reading NFC tags", "reading NFC readers" and "writing NFC readers". The specific functions of these modules have been described in the previous section, so they will not be repeated here.

2) Terminal System Platform

Terminal systems in attractions, as the medium between module clients and the server need to deal with interactions with those two platforms. This platform can be divided into three modules, including "sending NFC data", "receiving NFC data" and "updating information of attractions". The "sending NFC data" module corresponds to the "reading NFC readers" module while the "receiving NFC data" module corresponds to the "writing NFC readers" module. The specific functions have also been described in the previous section. The noticeable point is that terminal systems upload users' access numbers aggregated by NFC readers in both reading and writing mode in the process of "uploading the user's identity" and "updating information periodically" respectively.

3) Server Platform

The server platform can be developed with the lightweight J2EE Servlet +Spring +Hibernate framework, deployed on Apache Tomcat server, and implemented as a three-tier architecture. The persistence layer and data access layer, the lowest two layers of the architecture, map entity classes and database tables by Hibernate technology. The database tables involve "information of attractions" and "route information between adjacent attractions". Middle-tier business logic layer interacts with the upper and lower layers of the Spring framework, and mainly deals with 4 aspects of business including "collecting information of attractions and route information between adjacent attractions", "responding to the inquiry from users or terminal systems", "updating attractions' scores and users' access numbers" and "transmitting updated information of attractions to terminal systems periodically". The top-level presentation layer provides the interface corresponding to the four aspects above implemented by Servlet technology, for receiving requests and data from mobile clients and terminal systems, and then returning the corresponding result.

C. Implementation of Communication between Client and Server

The mobile clients access the Internet through the cellular network or the scenic WiFi wireless hotspots,

while terminal systems in attractions and the server accesses the Internet through wired LAN. If mobile clients are not in the scenic WiFi LAN, they have to access the server via the cellular network, so it is necessary to do port mapping on the router of the LAN where the server is, and then the client can access the server via a fixed IP address and port number. Most operations in the system adopt HTTP communication except a few particular operations which adopt communications such as Socket and NFC peer-to-peer. For these operations with HTTP communication, the client sends an HTTP POST or HTTP GET request to the server, and the server returns result data encapsulated in the JSON (JavaScript Object Notation) data interchange format, which is easy for the client to resolve and display to the user. The collected information of attractions and users' scores for attractions are uploaded to the server by POST method, with data contents as parameters. The server then returns a JSON Object to the client to notify the processing results. As for the attraction's information inquiring process, the client sends a request by GET method, with the name of the current attraction as the parameter. The server encapsulates the results into a JSON Array and returns it to the client. Figure 5 shows the architecture of the server and its interaction with mobile clients.

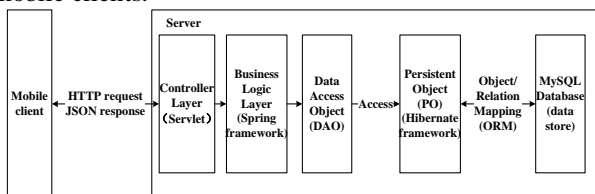


Figure 5. The architecture of the server and the interaction of system

IV. SYSTEM PERFORMANCE TESTING EXPERIMENTS

The “NFC-based scenic service system” provides users with three programs to acquire information of attractions just as previously proposed, namely, the “reading NFC tags” program, the “reading NFC readers” program and the “writing NFC readers” program. The specific operations of various programs are tested here to compare advantages and disadvantages, and the proper programs in various cases will be recommended to users. Throughout the testing process, the server is Lenovo Centre M5100t (CPU AMD Phenom II X4 2.59GHz; Memory RAM 3.0GB). The terminal system platform in each attraction is HP Pavilion dv4-1104TX (CPU Intel Core2 Duo P7350 2.00GHz; Memory RAM 2.0GB), connected with two ACR122U NFC readers (one for reading data, the other for writing data). Some SONY Xperia LT26i NFC tags are also needed. The mobile platform is Samsung Galaxy Nexus (CPU 1228MHz, Memory RAM 1GB) with Android 4.0 OS. The mobile phone is configured with NFC reading and writing functions.

In order to reduce errors caused by the different delay time of devices touch or finger click in each time of the test and the interference by other unknown events in the system environment, 30 groups of tests are carried out

and the test results are averaged. The time consumption comparisons of the various operations in all three programs are shown in Table I. For each program, it has been considered that mobile phones interact with the server through both the scenic WiFi LAN and the cellular network. For some operations which are not influenced by the type of network, the table displays the average result of all the test data. Similarly, for an operation which belongs to two or three programs, the result shown in the table is also the average value of the operation's test data in all programs. In this way, the impact of measuring error can be eliminated as possible when comparing the three programs, and then the total time consumption of each program can be analyzed and compared more accurately.

As can be seen from Table I, these operations can be divided into five categories, and the biggest differences among the three programs are the process of NFC touch and attractions' information interaction. In the process of NFC touch, reading NFC tags and readers are all reading operation for an NFC mobile phone. When the data has been prepared on the tag or the reader, a simple touch can complete the data transmission. Among them, the process of reading readers, as a mutual reading process, saves more time than the process of reading tags, in which the tags act as passive devices. By contrast, writing a user's information to a reader needs a click on the user's NFC mobile phone to confirm the data transmission to the reader after the phone touches the reader. Manual operations will increase the time greatly, which makes the total time above 1 second.

For the “reading NFC tags” program and the “reading NFC readers” program, a user's mobile phone joins the scenic WiFi LAN with BSSID and a password (if required). If the WiFi hotspot works, and the user turns on his or her mobile phone's WiFi function and has not joined the WiFi LAN, the whole process of joining the network, including hotspot authentication, DHCP dynamic IP address assignment and so forth, takes more than 2 seconds. Otherwise, the process of joining the network fails quickly, after which the user has to access the server through the cellular network to acquire information of attractions and upload scores. In the process of attractions' information interaction, the “reading NFC tags” program and the “reading NFC readers” program acquire the information from the server and the terminal system respectively. The former process has a large difference in using the WiFi LAN and the cellular network to access the server, and the time is tens of milliseconds and over 1 second respectively. The difference is caused by the weak wireless signal of the cellular network and multi-hop routers forwarding packets. The latter process refers to reading attraction files from the terminal system and writing information to a reader, which takes a very short time and does not depend on the type of network. It is followed by the process of NFC touch, in which the user's mobile phone accesses information of attractions directly by reading the reader. In contrast, in the “writing NFC readers” program, the server needs to push information of attractions to

TABLE I. THE TIME CONSUMPTIONS OF ALL OPERATIONS IN THE THREE PROGRAMS (MS)

Operation category	Specific operation	Time consumption (ms)		Programs including the operation
		WiFi LAN	Cellular network	
NFC Touch	Reading information of attractions in the tag	2.6		Reading NFC tags
	Reading information of attractions in the reader	1.2		Reading NFC readers
	Writing the user's information into the reader	1514.1		Writing NFC readers
Joining the network	Joining the WiFi LAN	2518.4	2.7	Reading NFC tags and reading NFC readers
Attractions' information interaction	Acquiring information of attractions from the server	58.4	1127.1	Reading NFC tags
	Acquiring information of attractions from the terminal system	1.25		Reading NFC readers
	Receiving information of attractions from the server	547.6	145278.6	Writing NFC readers
Scoring	Uploading scores for the attraction	154.3	1293.2	Reading NFC readers and writing NFC readers
System management	Writing information of attractions to the tag	217.7		Reading NFC tags
	Uploading information of attractions to the server	141.5		All programs
	Uploading route information between adjacent attractions to the server	144.4		All programs
	Updating information of attractions in the terminal system	3595.7		Reading NFC readers

TABLE II. COMPARISON OF THE TOTAL TIME CONSUMPTIONS BETWEEN THREE PROGRAMS (MS)

Program name	WiFi LAN	Cellular network	System management
Reading NFC tags	2515.7+63.7n	1132.4n	217.7m
Reading NFC readers	2.5n(+2515.7+157n)	2.5n(+1295.9n)	3595.7m
Writing NFC readers	2515.7+2064.4n(+154.3n)	146795.4n(+1293.2n)	0

users' mobile phones according to each user's identity. If the phone has joined the WiFi LAN, it takes about half a second via Socket communication. Otherwise, there will be a great delay via SMS, usually in more than 2 minutes. The programs of "reading NFC readers" and "writing NFC readers" allow users to upload scores for attraction recommendation when users can check in. This process will also take users over 1 second to access the server through the cellular network, much longer than that through the WiFi LAN.

Finally, it is also necessary to test the time consumptions of data acquisition and system maintenance process, and they can be considered as additional time consumptions in the comparison of programs. The "writing information of attractions to tags" operation provides tag information for the "reading NFC tags" program. And the "uploading information of attractions and route information between adjacent attractions to the server" operation collects data for the system. All these operations take a short time, in approximately 200 milliseconds. In addition, terminal systems of the "reading NFC readers" program should be maintained periodically, with downloading the latest information from the server, and the updating process each time needs nearly 4 seconds.

Base on the analysis of time consumptions of all operations, the total time of the three programs under two types of network can be calculated as shown in Table II.

The data in brackets in Table II is the additional time of uploading scores, and the time will not be counted into the total time of each program when the user cancels the uploading process. The n is set to be the number of times of accessing information, and m is the number of times of updating tags' or readers' information. Since joining the scenic WiFi LAN will occur when a user reads an NFC

tag or an NFC reader for the first time, but a repeated process is not needed, the total time of completing all operations of a program through the WiFi LAN is not proportional to the number of interactions. Compared with using the cellular network, the more number of times of completing all operations of a program, the more savings in the total time via WiFi. For the "reading NFC tags" program and "the "reading NFC readers" program, when completing the program for the first time, the time consumption is longer than that through the cellular network due to the time cost joining the WiFi LAN. According to the test results in Table II, when $n \geq 3$, the time of completing all the operations in a program through the WiFi LAN will be less than that through the cellular network. For the "writing NFC readers" program, although it does not include the process of joining the WiFi LAN, its operations such as interactions with the server need users' mobile phones to join the WiFi LAN as a precondition. Therefore, when considering the total time of all the operations in the program, it is still necessary to add the time of joining the WiFi LAN. The time of completing all the operations in the program through the WiFi LAN is obviously less than that through the cellular network regardless of the size of n, due to the long time of data transmission via SMS compared with Socket. By comparing the two network types, it can be concluded as follows: If users' mobile phones include the WiFi function, they should be prompted to turn on their WiFi, and join the scenic WiFi LAN by reading tags or readers. In this way, users can acquire information of attractions and upload scores quickly in various attractions only by joining the network once.

In comparison of the three programs, the "writing NFC readers" program takes too long time, but the program has the advantage of obtaining information about all

aspects of users, so that the system can analyze the user behavior and provide more suitable recommendation for them. The “reading NFC readers” program can complete operations in the least time among these programs both through the WiFi LAN and through the cellular network if uploading scores is unnecessary. However, the additional time of system management increases the total time apparently. According to the test results in the table, if both programs need one time of maintenance everyday (i.e. $m=1$), when $n \geq 15$ for the WiFi LAN and $n \geq 3$ for the cellular network, the total time of these n users by reading readers will be shorter than that by reading tags. Moreover, the system usually updates data at midnight, which does not take users’ time to access attractions. Therefore, in the same condition, the “reading NFC readers” program takes the least total time. Nevertheless, users generally upload their scores to the server in this program, which increases its time consumption and makes it more time-consuming than the “reading NFC tags” program. By comparing the total time consumptions of the three programs, it can be concluded as follows. The “writing NFC readers” program is used to analyze users’ identities and obtain the user behavior. Therefore, the “writing mode” readers should be deployed in attractions as less as possible, and a user is limited to access this kind of readers in an attraction only once a day. The “reading NFC readers” program is the fastest program to acquire information of attractions. The “reading mode” readers can be deployed in certain amount, to provide users with information and collect users’ access number. The “reading NFC tags” program is suitable for the users who are unwilling to check in and upload scores, by which they can acquire information easily by reading tags.

VII. CONCLUSION

This paper proposes three programs using NFC mobile phones to interact with NFC tags and readers, and to acquire information of the current attraction and nearby recommended attractions. It also presents a check-in and scoring mechanism for recommending attractions and verifies its correctness through mathematical derivation. A scenic service system using these programs and the mechanism is then designed and implemented. After that, the paper tests and calculates the time consumptions of various operations, and compares and analyzes the total time consumptions of the three programs in the two network types. The system’s performance can meet users’ need basically, but some aspects still need improvement, such as too long time of the “writing NFC readers” program. Further analysis of the user behavior and characteristics of the three programs will be necessary in the future, to design a comprehensive multi-program selection strategy.

ACKNOWLEDGMENT

The authors wish to thank the members of the Embedded System and Information Security Lab, College

of Information Technical Science, Nankai University for their support and help. This work was supported in part by a grant from China’s Ministry of Science and Technology: Science and Technology based SME Technology Innovation Fund (Granted No. 10C26211200143).

REFERENCES

- [1] K. Pahlavan, X. R. Li, J. P. Makela, “Indoor geolocation science and technology”, *Communications Magazine, IEEE*, vol. 40, no. 2: pp. 112-118, 2002.
- [2] J. Hightower, G. Borriello, “Location systems for ubiquitous computing”, *Computer*, vol. 34, no. 8: pp. 55-66, 2001.
- [3] P. Bahl, V. N. Padmanabhan, “RADAR: An in-building RF-based user location and tracking system”, *In Proceedings of the IEEE INFOCOM, Tel Aviv*, pp. 775-784, 2000.
- [4] M. Isomursu, P. Isomursu, M. Komulainen-Horneman, “Touch to access the mobile Internet”, *In Proc. of the 20th Australasian Conf. on Computer-Human Interaction: Design for Habitus and Habitat*, pp. 17-24, 2008
- [5] M. A. Ayu, T. Mantoro, S. A. Ismail, N. S. Zulkifli, “Rich information service delivery to mobile users using smart posters”, *In Proc. of the DICTAP 2012, Second Int. Conf. on Digital Information and Communication Technology and it’s Applications*, pp. 149-153, 2012
- [6] U. B. Ceipidor, C. M. Medaglia, V. Volpi, etc, “NFC technology applied to touristic-cultural field: A case study on an Italian museum”, *In Proc. of 2013 5th International Workshop on Near Field Communication, NFC 2013*, 2013
- [7] NFC Forum, Near Field Communication Interface and Protocol (NFCIP-1) 2nd Edition, 2004. 12
- [8] NFC Forum, Near Filed Communication Interface and Protocol (NFCIP-2) 1st Edition, 2003.12
- [9] T. Ghanname, How NFC can to speed Bluetooth transactions today, EETimes, 2006. 2. 14, <http://www.eetimes.com/design/communications-design/4012606/How-NFC-can-to-speed-Bluetooth-transactions-151-today>
- [10] NFC Forum, NFC Data Exchange Format (NDEF) Technical Specification 1.0, 2006. 7. 24

Jie Ma received his B.S degree in Physics in 1982 from Nankai University, China. He is current a professor in College of Software, Nankai University, China. He publishes many research papers in journals. A number of his studies have won the provincial and ministerial awards of china. He has a number of social part-time jobs and many projects in research supported by national, provincial and ministerial funds. His research interests are New Media, Embedded System and Wireless Networks.

Jinlong E received his B.S. degree in Software Engineering in 2007 from Nankai University, China. He is current a master student in Computer Software and Theories, College of Software, Nankai University, China. His research interests are Wireless Networks, Mobile Computing and Internet of Things (IOT).

Research on Delay and Packet Loss Control Mechanism in Wireless Mesh Networks

Qiuling Yang^{1,2}, Zhigang Jin¹, and Xiangdang Huang^{2*}

1. School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China

2. College of Information Science & Technology, Hainan University, Haikou 570228, China

*Corresponding author, Email: yql0515@163.com, zgjin@tju.edu.cn, xiangdanghuang@126.com

Abstract—In wireless mesh networks, the performance of TCP was degraded rapidly due to the interference in wireless channels. To deal with this problem, A TCP control mechanism based on the character of delay distribution and wireless packet loss is proposed in this paper. Firstly, this delay model can capture the delay exactly that a packet experiences at one hop transmission with rigorous theoretic derivation and lower overhead, and computational complexity. Then we analyze the character of the wireless packet loss. Furthermore, this mechanism points out the control method at transport layer to deal with the different type of packet loss. The simulation results show that our mechanism can decrease the packet loss rate efficiently.

Index Term—Wireless Mesh Networks; Delay Distribution; TCP Packet Loss Control

I. INTRODUCTION

Multimedia service, a type of delay sensitive services, has strict constraints on time delay, which means the data packets exceeding the time threshold to be abandoned. There are multi-kinds of reasons for packet loss, such as network queue overflowing, maximum number of retransmission exceeding and beyond time threshold, etc. In wireless mesh networks, data from source need to be sent to gateway through backbone network with multi-hop access, and finally to the Internet. One delay-sensitive service is usually constrained by maximum tolerable end-to-end time delay, while the end-to-end time delay is the accumulation of time delay in each hop. As a result, the control algorithm of end-to-end delay in wireless mesh networks must start from ensuring the single-hop delay. As for the inevitable problem of packet loss, this paper proposes a method to decrease the packet loss rate by adjusting the TCP congestion window dynamically.

In recent years, researchers have built models, analyzed the network congestion and data packet transmission delay by using Queuing Theory. Guihai et al. in [1] organized the queue of data packets at the wireless mesh networks using M/D/1 model, but this model assumed the time of processing the data packets in the gateway is a constant, and it failed to consider the impacts on delay by different QoS in different types of services and network congestion, So it is not accurate enough for the estimation of time delay. Bisink et al. modeled and analyzed the queue at the interfaces in MAC

layer using G/G/1 model in [2], it deducing the average delay of data packet transmission between nodes in wireless mesh networks. Though wireless mesh networks are mainly used for Internet services, this model also provided methods and instances for subsequent research on time delay of backbone networks. Gupta et al. in [3] modeled the time interval between two successful transmissions of one node based on Markov chain. Omer in [4] utilized discrete-time Markov chain model to analyze the “hunger” queue produced by contention of the gateway bandwidth by nodes having different hops to the gateway. Pan Lei in [5] analyzed packet loss caused by different factors using discrete-time Markov chain model, but it didn’t give the corresponding TCP congestion control method.

The existing works mentioned above provide some methods and instances reference for the analysis on the delay and its characteristics of distribution, but all of these cases do not well match the actual wireless network scenarios and applications. Focusing on the random characteristics of the data packets arrival time and service time, the delay analysis of this paper takes avoidance-contention mechanism in MAC layer, probability of data transmission attempt and link throughput into account, and grasps the delay distribution on network layer, MAC layer and Physical layer exactly with the rigorous theory and probability analysis.

In the current implementation of TCP in wireless mesh networks, there have been many solutions proposed for improving the performance of TCP. We classify the related work into two parts. One is to timeout packet loss, and the others are to congestion packet loss. Jacobson et al. in [6] used the TCP timestamp option for detecting TCP retransmission timeout, it compared the extra information in the Acks in the sender and the arrival to detect and eliminate retransmission timeout. The general idea of DSACK algorithm [7] is that if a retransmitted packet has been acknowledged for the second time, the sender assumes that the earlier retransmission was spurious, but the slow reaction of DSACK judgment makes the TCP retransmit several packets mistakenly, or the worst case, a whole window of packet has to be retransmitted.

These TCP control schemes could distinguish spurious packet loss from congestion packet loss. However, these schemes can not well deal with the performance

degradation of TCP. To the current review of literatures, the study of detecting non-congestion packet loss and congestion has been carried out by MI-Young et al in [8], this scheme identified the non-congestion packet loss by the estimation of rate of queue usage during the retransmission, thereby avoiding of unnecessary congestion window size reduction, but this scheme could not solve the problems of random packet loss efficiently. Sreekumari et al in [9] applied the improved explicit congestion notification mechanism to detect congestion packet loss and reduced the size of congestion window following the algorithm of TCP NewReno, thereby improvement of TCP performance in wireless mesh networks, however, this mechanism classified the non-congestion into random packet loss and spurious packet loss, increasing the complexity of algorithm, and that had not further discussion on the timeout packet loss and the size of congestion window systematically. By considering the limitations of this mechanism, we develop a new control mechanism for non-congestion packet loss with lower computational complexity.

II. QUEUING DLELAY MODEL OF NETWORK LAYER IN WIRELESS MESH NETWORKS

Definition 1: The time interval between a data packet reaches the waiting queue of a node and becomes the head of the queue is defined as queue waiting delay, described by $d_{i-queue}$.

A. Establishment of Queuing Model

This paper utilizes M/M/1 queuing model to simulate the waiting process of data packets in a router. The model uses λ (the arrival rate of data packets/sec) to describe the randomness of data packets arriving at router (or gateway). Assuming the initial performance of each router is the same, and packet sizes and priorities of different services are unequal, then the required service time is different. This paper uses μ , the transmission rate of data packets/sec, to describe the probability characteristics.

Assume the number of data packets queuing in buffer waiting for transmission in current router node is n , where the data packet with label No.1 is the head element of the queue and the time for transmission is t_1 , and so on for the other packets. The data packet just arrived (label is $n+1$) will be head of the queue after waiting for $T=t_1+t_2+\dots+t_n$, so $d_{i-queue}=E(T)$. From characteristics of M/M/1 model and definition of gamma distribution, $T=t_1+t_2+\dots+t_n$ is consistent with the probability distribution characteristics of gamma distribution, and the expected value of time interval T which is the period from starting queuing to becoming the head can be obtained through the following derivations.

B. Solution of Queuing Model

Let B represent that queue in a node is not empty, that is, the new arrival data packets need waiting before transmission, and $P(B)$ represents the probability of B correspondingly. Let $P(n)$ represent the probability of the queue length being n , and the exact values of $P(B)$ and $P(n)$ can utilize the existing conclusion of the M/M/1

model. This paper obtained the expected value of T by establishing the relationship of probability distribution of queue length in saturation state and density function of waiting time T, and the procedure is as follows:

$$P(B)f(t) = \sum_{n=1}^{\infty} P(n)f_1(t) * f_2(t) * \dots * f_n(t)$$

$$t \text{ subject to } f_1(t) * f_2(t) * \dots * f_n(t) = \begin{cases} \frac{\mu e^{-\mu t} (\mu t)^{n-1}}{\Gamma(n)}, t \geq 0 \\ 0, t \leq 0 \end{cases}, \Gamma(n) = (n-1)!$$

In this equation, $f(t)$ and $f_i(t)$ represent density functions of T and t_i , * represents convolution. According to the M/M/1 model, $f(t)$ is derived:

$$f(t) = \frac{\sum_{n=1}^{\infty} P(n) f_1(t) * f_2(t) * \dots * f_n(t)}{\sum_{n=1}^{\infty} P(n)}$$

$$= \frac{\sum_{n=1}^{\infty} \left[\left(\frac{\lambda}{\mu}\right)^n \frac{\mu e^{-\mu t} (\mu t)^{n-1}}{(n-1)!} \right]}{\sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n}$$

Finally, the expected value of T, $E(T(\lambda, \mu, n))$ is obtained:

$$E(T(\lambda, \mu, n)) = \int_0^{+\infty} t f(t) dt = \frac{1}{\sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n} \frac{1}{\mu} \sum_{n=1}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n$$

Through the derivation from equation (1) to (3), the queuing delay time of the data packet with label 'n+1' in one hop at network layer is $d_{i-queue} = E(T(\lambda, \mu, n))$. From the expression of T, the queue waiting delay of a data packet is determined by data packet arrival rate, router transmission rate and the queue length.

III. CONTENTION DELAY OF MAC LAYER IN WIRELESS MESH NETWORKS

A. Contention Delay of MAC Layer $d_{i-content}$

Definition 2: The time interval from becoming head element to gaining physical channel through contention is called contention accessing delay, expressed by $d_{i-content}$, which is decided by channel state of MAC layer sensed by the head data packet of the queue which is about to achieve services.

B. Channel State

There are three kinds of channel state at node i: busy, idle, and back-off. Here some contention accessing delays in different kinds of channel states are presented in this section, where all the mentioned data packets are referred to head packet.

1) Idle Channel State

In this paper, the idle state of channel is defined as: when queues of node i and its neighbor nodes contending for the same channel are all empty, the channel state is idle. Otherwise, channel state sensed by nodes should be busy or back-off. At this state, a new arrival data packet can be transmitted immediately after a DIFS (Distributed Inter Frame Space), and this delay time is expressed as d_i^{idle} , that is:

$$d_i^{idle} = DIFS \tag{4}$$

2) *Busy Channel State*

Busy channel state means data in node i or its neighbor nodes is occupying the channel. If the channel is in busy state when a data packet arrives, this delay is defined to be busy delay (let d_i^{busy} denote it), no matter the transmission is successful or conflicted. With the mechanism of CSMA/CA, the node has to go through a complete back-off procedure before achieving channel. Assume the time for transmitting a data frame successfully is T_t , and the time of a transmission collision is T_c , and the number of data of successful transmission in back-off stage of node i and the number of collision are C_{i-suc} , C_{i-col} respectively. Then the delay of data when the channel is under busy state is:

$$d_i^{busy} = C_{i-suc} \times T_t + C_{i-col} \times T_c + DIFS \tag{5}$$

3) *Back-off State*

Definition of back-off state: the channel is in the idle state at this stage, but the waiting queue of node i or its neighbor nodes are not empty, as time goes on, channel will provide service to the node with the counter decreasing to zero first. Because the time duration of back-off stage is extremely short compared to busy state or idle state^[10], the delay caused by back-off is ignored in this paper.

C. *Computation of Contention Accessing Delay in MAC Layer*

As mentioned above, the contention delay in MAC layer is mainly decided by the channel state perception of the head data packet. Synthesizing the equation (4) and (5), the contention delay of a node in MAC layer is:

$$d_{i-conten} = DIFS + \begin{cases} 0 & (d_{i-conten} = d_i^{idle}, \text{ idle state}) \\ C_{i-suc} T_t + C_{i-col} T_c & (d_{i-conten} = d_i^{busy}, \text{ busy state}) \end{cases} \tag{6}$$

IV. TRANSMISSION DELAY OF PHYSICAL LAYER IN WIRELESS MESH NETWORKS

Definition 3: Going through delay of the above two stages, data starts obtaining the right to the use of the channel. The time interval from the data achieving the right for the first time to transmission successfully is called transmission delay (denotes as $d_{i-trans}$), which can be interpreted as the time of transmitting in physical media, including the time of collision transmission.

A. *Bianchi's Node Throughput Model*

The model in this paper is based on the improved Bianchi throughput model proposed in [10]. Here we introduce the Bianchi throughput model firstly. A channel time slot is an average form weighted by channel state probability in each stages using synthetic index method in this model. Node throughput is obtained in combined with the time slot and successful transmission payload based on node transmitting attempt probability. Assuming ζ is the probability of transmitting attempt in one certain time slot of one node, this model derives the available throughput x_{n_i} of one single node in a single collision domain:

$$x_{n_i} = \frac{\zeta (1-\zeta)^{N-1} L_i}{N\zeta (1-\zeta)^{N-1} T_{suc} + P_{col} T_{col} + P_{idl} T_{idl}} \tag{7}$$

$$= \frac{\zeta (1-\zeta)^{N-1} L_i}{\Lambda}$$

where L_i is data payload, Λ is length of one channel time slot, consisting three parts: successful transmitting time T_{suc} , collision time T_{col} , idle time T_{idl} [10], and the probabilities are P_{suc} , P_{col} , P_{idl} respectively.

B. *Derivation of the Link Throughput Model*

As described in previous section, the throughput of a certain node n_i in a single collision domain in Bianchi model can be expressed as:

$$node\ throughput = \frac{data\ payload\ size}{time\ of\ one\ channel\ time\ slot} \tag{8}$$

This section derives the throughput x_{l_i} in one-hop link by improving the Bianchi throughput model in single collision domain. Bianchi model [10] assumes that there is only one export link of each node, but data of multiple nodes may need to be forward by this node in actual networks. This means that a node may have multiple export links. Because performance parameters of each link are different, their throughput is not average value. This section expands the Bianchi node throughput model to a situation where one node has multiple export links, and derives the throughput in a particular link. Therefore, following assumptions are made:

1) Node n_i has to forward data through its E export links. We denote the Attempt transmitting probability and collision probability corresponding of link i as ζ_i and p_{col-i} respectively.

2) The arrival rate and service rate of link i are denoted as λ_i and μ_i respectively. So the total data packet arrival rate and total service rate could be expressed as $\lambda = \sum_{i=1}^E \lambda_i$ and $\mu = \sum_{i=1}^E \mu_i$ respectively.

3) Define the probability where a data packet is transmitted through link i is $\psi_i = \frac{\lambda_i}{\lambda}$.

The back-off mechanism adopts binary exponential back-off algorithm. Chatzimisios shows that the number of average back-off time slots gone through by a data packet at the k_{th} back-off stage is $\frac{W_k + 1}{2}$ in [11]. Then the total number of time slots needed when a data frame goes through the link i is expressed as:

$$cou_i = \sum_{k=1}^m \frac{W_k + 1}{2} P_{col-li}^k \quad (9)$$

where W_k is the size of contention window of the K_{th} back-off stage; P_{col-li}^k is the collision probability of link i at the K_{th} back-off stage. Thus the probability that a channel slot time is occupied by link i can be obtained, denoted as u_i :

$$u_i = \frac{\psi_i cou_i}{\sum_{j=1}^E (\psi_j cou_j)} \quad (10)$$

The probability ζ that node n_i attempts to transmit data can be expressed by u_i :

$$\zeta = \sum_{i=1}^E u_i \times \zeta_i \quad (11)$$

The probability P_{col} that node n_i conflicts can be expressed as:

$$P_{col} = \frac{\sum_{i=1}^E u_i \times \zeta_i \times P_{col-li}}{\zeta} \quad (12)$$

The transmission probability ζ_i and collision probability P_{col-li} of link i can be derived according to equation (12). At last, referring to Bianchi model, assuming L_i is the size of data frame and x_{li} is the average throughput of link i , the throughput of link i can be obtained:

$$x_{li} = \frac{u_i \zeta_i (1 - P_{col-li}) L_i}{\Lambda} \quad (13)$$

Hence the transmitting delay of a data frame through link i can be expressed by:

$$d_{i-trans} = \frac{L_i}{x_{li}} \quad (14)$$

Now, the queuing delay, contention delay and transmitting delay of a data packet are all obtained. And the total delay can be obtained by adding up the equation (3), (6) and (14).

V. ANALYSIS ON CHARACTER OF THE WIRELESS PACKET LOSS AND CONTROL MODE IN TRANSPORT LAYER

In wireless mesh networks supporting delay-sensitive services, the reasons for packet loss could be classified into two types: one is congestion, and the other is timeout.

Assuming the node has finite size of buffer L and limited maximum time of retransmission C , when the queue is full, the arriving data packets will be abandoned. Moreover, the data packets which exceed the maximum number of retransmission caused by channel error or interference would also be abandoned. Such is called packet loss by congestion. In addition, as to delay-sensitive services, data packets arriving beyond the time threshold will be abandoned directly. Such is called packet loss by timeout.

A. Character of Wireless Packet Loss

We assume that the channel state could be classified into good state (denoted as "1" state) and poor state (denoted as "0" state). According to these assumptions, the transfer probability could be expressed as $p_{ij} | \{i, j \in \{0,1\}\}$. In this paper, we used the following formula to analyze channel state:

$$y(t) = (u, v) | (0 \leq u \leq C, 0 \leq v \leq L) \quad (15)$$

where u represent the times of retransmission is $L - u$, and v represent the size of buffer. From the above definition of the congestion packet loss, the congestion packet loss due to buffer overflow would happen when $v = L$ and $u \geq 2$, exceeding maximum number of retransmission when $u = 1$ and $v > 0$ would also cause the congestion packet loss. The remaining packet loss happen when $u \geq 2$ and $v < C$, which caused by the time threshold of the delay-sensitive services. The channel transfer probability were shown in table 1, table 2 and table 3 when the packet loss happened.

TABLE I. TRANSITION PROBABILITY WHEN CONGESTION LOSS DUE TO BUFFER OVERFLOW

$y(0, L)$	$y(u-1, L)$
P_{10}	P_{11}

TABLE II. TRANSITION PROBABILITY WHEN CONGESTION LOSS BECAUSE OF EXCEEDING THE MAXIMUM NUMBER OF RETRANSMISSION

$y(0, v-1)$	$y(0, v)$	$y(C, v-1)$	$y(C, v)$
$\bar{\lambda} p_{10}$	λp_{10}	$\bar{\lambda} p_{11}$	λp_{10}

TABLE III. TRANSITION PROBABILITY WHEN PACKET LOSS DUE TO TIMEOUT

$y(0, v)$	$y(o, v+1)$	$y(u-1, i)$	$y(u-1, v+1)$
$\bar{\lambda} p_{10}$	λp_{10}	$\bar{\lambda} p_{11}$	λp_{10}

The influence of different packet arrival rates on the packet loss were analyzed by the Markov-Model [12, 13] of channel transfer state according to the transfer probability. In this section, we demonstrates the comparison between congestion packet loss and timeout packet loss in terms of packet arrival rate ranging from 0.2 to 0.9, corresponding to different value of packet error rate (PER). As shown in Fig. 1, Fig. 2.

On the basis of the analysis on the increasing trend of the packet loss shown in Fig. 1 and Fig. 2, then the following conclusions are drawn: 1) when the packet arrival rate is low, the proportion of congestion packet

loss and timeout packet loss are almost equal, on the contrary, the congestion packet loss occupy main proportion; 2) with rising of packet arrival rate, the congestion packet loss rate show a significant exponential increase trend, and the timeout packet loss show a linear increase trend, respectively.

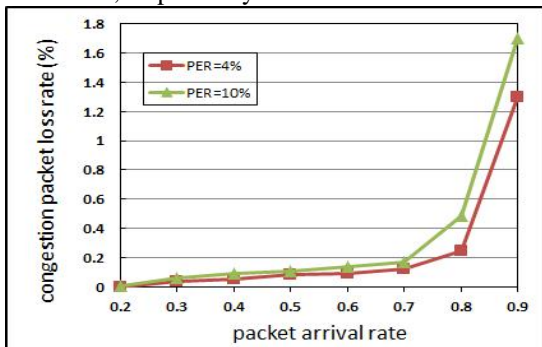


Figure 1. The influence of packet arrival rate on congestion loss

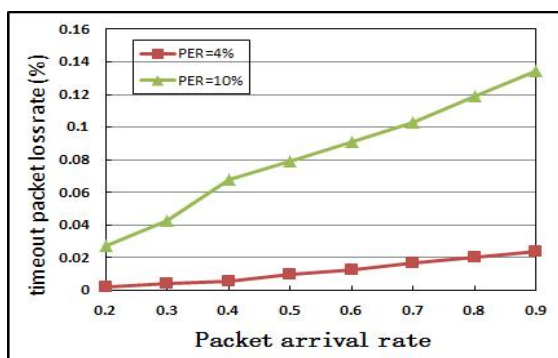


Figure 2. The influence of packet arrival rate on timeout loss

B. Proposed Control Mechanism in Transport Layer

When packet loss occurs, the source can receive the feedback and adjust the data packet arrival rate. The data packet arrival rate is controlled by TCP congestion window, and the unified mechanism of window halving is not suitable for different types of packet loss [5]. This paper proposes a dynamic control method of TCP congestion window according to probabilities of different type of packets loss.

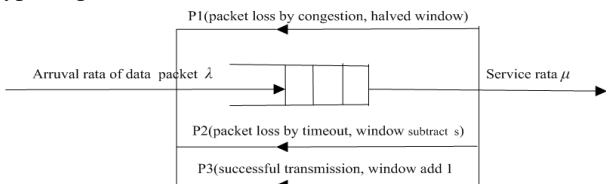


Figure 3. M/M/1/L Probability Model

To improve the ability of contending for channels of delay-sensitive services, this paper assumes when packet loss occurs, the congestion window subtract a certain value s , instead of halving the window. In this paper, we build an $M/M/1/L$ model for computing value s in different types of packet loss rate, as showed in Fig. 3. In this model, data sending rate can be expressed by w/RTT (w represents the current size of the window). Assuming probabilities of packet loss due to congestion

and timeout are $p1$, $p2$ respectively; and probability of successful transmission [12] is $p3$. Assuming the optimal sending rate is set to $p0$, which means the effective transmission. So the value of data packet arrival rate should be controlled at $p0$. According to the Queuing Theory, the balance equation can be obtained:

$$p0 * p1 * 1/2 + (p0 - s / RTT) * p2 + (p0 + 1 / RTT) * p3 = p0 \quad (16)$$

From equation (16), the value s could be obtained.

VI. ANALYSIS ON SIMULATION

In order to verify the efficiency of our mechanism, in this section, the performance improvement in packet loss is compared to that of related work TCP-NRT [9] and the traditional TCP halved window mechanism when $\lambda = 0.2$ and $\lambda = 0.9$ (λ denotes the packet arrival rate). The simulation scenario is that CBR data stream of phone-to-phone, queue length of MAC layer interface, maximum retransmission times and end-to-end delay constraint value are 10 packets, 7 times and 0.15s respectively. The simulation time is 200s. Comparisons of packet loss rate in terms of varying PER ranging from 1% to 10% are shown in Fig. 4 and Fig. 5.

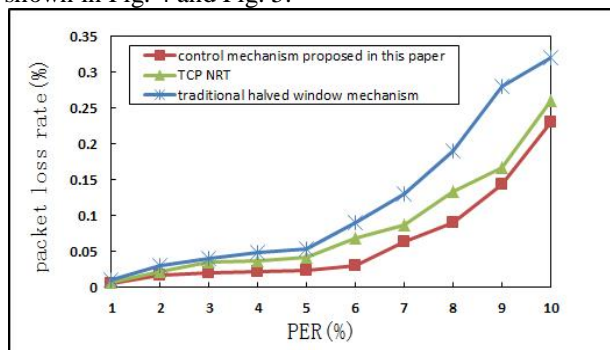


Figure 4. Packets loss rate with $\lambda = 0.2$

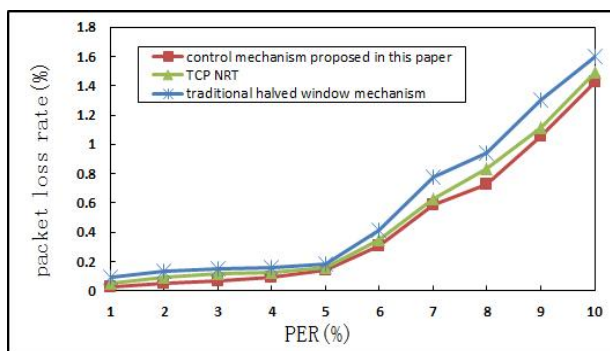


Figure 5. Packets loss rate with $\lambda = 0.9$

The simulation results and analysis show that our proposed control method can effectively support delay sensitive services with strict delay constraints. Specific analyses are as follows:

1) As shown in Fig. 4, when the packet arrival rate is low, the packet loss rate based on the control mechanism we proposed is significantly lower than the TCP-NRT [9] and traditional TCP halved window mechanism. Based on the analysis above in section V, the proportion of timeout packet loss is slightly higher while the packet

arrival rate is low, the simulation results draws a conclusion that our mechanism which focus on the timeout packet loss can support the delay-sensitive services with rigorous time threshold effectively;

2) As shown in Fig. 5, when the packet arrival rate is high, the packet loss rate based on the control mechanism proposed in this paper is slightly lower than the TCP-NRT [9] and traditional window halved window mechanism. Based on the analysis above in section V, the congestion packet loss occupy main proportion when the packet arrival rate is high, moreover, with rising of packet arrival rate, the congestion packet loss rate show a significant exponential increase trend, therefore, traditional window halved window mechanism has certain control meaning toward the congestion packet loss, but meanwhile, the rise of congestion packet loss could also result in the rise of timeout packet loss, they demonstrate a causal relationship. Therefore, the control mechanism proposed in this paper which focus on the timeout packet loss is still feasible when the packet arrival rate is high.

VII. CONCLUSIONS

This paper researches and analyzes the delay experienced on one hop of the data transmission in wireless mesh networks, using the classical M/M/1 queue theory model to describe the queuing delay of data at the network layer. Contention delay based on channel state is given by analyzing the IEEE802.11 DCF mechanism. It improves the throughput model proposed by Bianchi [10] with the combination of the principle of probability theory, and gets the throughput of different export links, helping obtain the delay of data transmission through the link. On the basis of the different types of packet loss probability, The influence of different packet arrival rates on the packet loss were obtained by the analysis on the character of the wireless packet loss through the Markov model. Finally, this paper points out the feedback control method at transport layer to adjust the size of the congestion window dynamically according to different types of packet loss after grasping of the delay distribution accurately, and the simulation results show that this mechanism can decrease the packet loss rate efficiently. That is, some conclusions are drawn in the end of this paper, the control mechanism at transport layer proposed in this paper which combine the character of delay distribution and wireless packet loss can provides more accurate delay metric and effective theory method for delay sensitive services in wireless Mesh network.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper. This paper is support by the national natural science foundation of China (Grant No.61261024) and by the special social service fund of Hainan University, China (Grant No.HDSF201301), I appreciate for that.

REFERENCES

- [1] X. Wu, J. Lu, Guihai. Analysis of bottleneck delay and throughput in wireless Mesh networks, In *2006 IEEE International Conference on Mobile Ad Hoc and Sensor Systems*. Vancouver, Canada, October 2006, pp. 765-770.
- [2] N. Bisink, A. Abouzeid. Delay and throughput in random access wireless Mesh networks, In *2006 IEEE International Conference on Communications (ICC 2006)*, Istanbul, Turkey, June 2006, pp. 403-408.
- [3] N. Gupta, P. R. Kumar, A performance analysis of the 802.11 wireless LAN medium access control, *Communications in Information and Systems*, 2009, 3(4), pp. 279-304
- [4] G. Omer, Vincenzo, J. Shi, et al., Measurement and modeling of the origins of starvation of congestion-controlled flows in wireless Mesh networks, *IEEE/ACM Transactions on Networks*, 2009, 17(6), pp. 1832-1845.
- [5] Pan Lei, Wu hongyi. Design and analysis of Prioritized Medium Access Control protocol for backbone routers in wireless mesh networks. *Tsinghua Science and Technology*, 2012, 17(5) pp. 537-552.
- [6] V. Jacobson, R. Braden, D. Borman. TCP extensions for high performance, RFC 1323. (1992). <http://www.ietf.org/rfc/rfc1323.txt>. Accessed 23 March 2012.
- [7] E. Blanton, M. Allman. Using TCP duplicate selective acknowledgment (DSACK) and stream transmission control protocol, RFC Information-3707, 2004. <http://tools.ietf.org/html/rfc3707>. Accessed 10 March 2012.
- [8] P MI-Young, C Sang-Hwa. Detecting TCP retransmission timeouts non-related to congestion in multi-hop wireless networks. *IEICE Transactions on information and Systems*, 2010, E93-D(12), pp. 3331-3343.
- [9] P. Sreekumari, M. Lee. TCP NRT: a new TCP algorithm for differentiating non-congestion retransmission timeouts over multi-hop wireless networks. *EURASIP Journal on wireless communications and networks*, 2013, (1): 172.
- [10] G. Bianchi. Performance analysis of IEEE 802.11 distributed coordination function, *IEEE Journal on Selected Areas in Communication*, 2000, 18(3), P535-547.
- [11] P. Chatzimisios, V. Vitsas, AC. Boucouvalas. Throughput and delay analysis of IEEE 802.11 protocol/Networked Appliances, 2002. *Liverpool. Proceedings. 2002 IEEE 5th International Workshop on. IEEE, UK, October 2002* pp. 168-174.
- [12] TM. Hoang, M. A. Hoang. Novel Analytical Model to Identify Link Quality in 802.11 Mesh Networks//Computational Intelligence, Communication Systems and Networks (CICSyN), *2012 Fourth International Conference on. IEEE, 2012*: 129-136.
- [13] L. Lei, J. Zhou, X. Chen, et al. Modelling and analysing medium access delay for differentiated services in *IEEE 802.11 s wireless mesh networks. Networks, IET*, 2012, 1(2) pp. 91-99.

Qiuling Yang, received her B.E degree in computer science and technology from Shenyang Aerospace University, China, in June 2003 and her M.E degree in computer science and technology from Guangxi University, China, in June 2010. She is currently working towards her Ph.D. degree in telecommunication engineering at Tianjin University, China. She is an associate professor at computer science and technology, Hainan University, China. Her current research interests include the VoIP performance enhancements in wireless Mesh networks and the wireless networks security.

Zhigang Jin, received his B.E degree and M.E degree in computer architecture from Tianjin University, Tianjin, China, in 1994 and 1996, respectively, and Ph.D. degree in signal & information processing, Tianjin University, China, in 1999. He is a professor and doctoral tutor at telecommunication engineering, Tianjin University, China. His current research interests include wireless Mesh networks, network and information security and the internet of things.

Xiangdang Huang, received his B.E degree in computer science and technology from Yanan University, China, in June 2002 and his M.E degree in software engineering from University of electronic science and technology, China, in June 2011. He is an associate professor at computer science and technology, Hainan University, China. His current research interests include the wireless mesh networks and the vehicle networks.

Compressed Wideband Spectrum Sensing with Partially Known Occupancy Status by Weighted l_1 Minimization

Zha Song and Huang Jijun

School of Electronic Science and Engineering, National University of Defense Technology, Changsha, China
Email: zhasong1987@hotmail.com, huangjj1989@sina.com

Li Ning

Unit 77108, People's Liberation Army, Chengdu, China
Email: 285170852@qq.com

Abstract—This paper considers the problem of compressed wideband spectrum sensing in wideband cognitive radio when partial occupancy status is known. While performing wideband spectrum sensing, incomplete prior information on the occupancy status may be obtained from coexisting narrowband detectors, collaborative sensing nodes or remote database. In this paper, we present a new optimization model in order to incorporate this prior information and then propose a particular weighting strategy in the reconstruction algorithm based on weighted l_1 minimization to solve it. Numerical simulation results demonstrate that the use of partially known occupancy status leads to an improvement in detection performance and the proposed approach exploits such prior information effectively. As incorrect prior information is unavoidable in practical situation, cases in which the prior information is non-ideal are also investigated via simulations.

Index Terms—Cognitive Radio; Wideband Spectrum Sensing; Compressed sensing; Partially Known Occupancy Status; Weighted l_1 Minimization; Non-ideal Prior Information

I. INTRODUCTION

Spectrum sensing, whose objectives are detecting signal of licensed users (LUs) and identifying the spectrum holes for dynamic spectrum access (DSA), is an important enabling technology for cognitive radio (CR), a leading choice for efficient utilization of spectrum resource [1]-[3]. Nowadays, wideband applications have received significant attentions since they not only offer high throughput, but also purport pronounced spectrum access opportunities for CR users. Meanwhile, wideband spectrum sensing (WSS) entails considerable challenges in practice, especially its very high signal acquisition costs [2]-[4].

Recently, compressed sensing (CS) theory [5]-[7] has been introduced to alleviate the heavy pressure on conventional analog to digital converter (ADC) technology in WSS by utilizing the low percentage of spectrum occupancy – a fact that motivates dynamic

spectrum access [8], [9]. Compressed sensing theory shows that sparse or compressible signal can be reconstructed from much fewer samples than that suggested by the Shannon-Nyquist sampling theorem. Although existing frameworks of compressed wideband spectrum sensing (CWSS) [10], [11] show powerful ability of reducing signal-acquisition complexity, they are quite vulnerable to noise and the performance of CWSS degrades severely when signal to noise ratio (SNR) is low.

Different from traditional sparse signal recovery algorithms in which sparsity is the only prior information on signal characteristic, other forms of prior information about the signal's structure, such as partial support knowledge [12]-[15], support probability [16], connected tree structure [17], block-sparsity structure [17], etc., have been introduced into the reconstruction process. It has been demonstrated that the further exploitation of signal model, in addition to sparsity, would give birth to recovery performance enhancement.

While performing wideband spectrum sensing, partial occupancy status may be known from coexisting narrowband detectors, collaborative sensing nodes or remote database [2]. In this paper, we study how to exploit partially known occupancy status (PKOS) to improve the detection performance of CWSS. In order to incorporate this type of prior information, a new CWSS model is presented and then a particular weighting strategy in the reconstruction algorithm based on weighted l_1 minimization is proposed to solve it. The key idea of our proposed approach, named CWSS with PKOS (CWSS-PKOS), is to separated the spectrum whose occupancy status is known into two disjoint parts, the part known to be occupied and that known to be unoccupied, then impose relative small weights on the former which encourage nonzero entries in the reconstructed signal and relative large weights on the latter which discourage nonzeros. Simulation results demonstrate that the introduction of partially known occupancy status provides an improvement in detection performance of compressed wideband spectrum sensing, and that the

proposed CWSS-PKOS approach can exploit such prior information effectively. In addition, we exemplify cases in which the prior information is non-ideal, namely some prior information is incorrect, and such cases are more significant in practical situation than cases with ideal prior information.

The remainder of the paper is organized as follows. The signal model and the spectrum sensing problem of interest are explained in Section II and a typical representative of the traditional compressed wideband spectrum sensing approach is provided in Section III. In order to incorporate partially known occupancy status, Section IV presents the proposed CWSS model and the particular weighting strategy. Performance of proposed approach is demonstrated by numerical experiments in Section V and we draw conclusions in Section VI.

Throughout this paper, boldfaced characters denote matrices and vectors. The superscripts of $(\cdot)^T$ represent the operations of transpose. The notation $\|r\|_k$ denotes the l_k norm of the vector r . For a set S , $|S|$ denotes its size (cardinality).

II. SIGNAL MODEL

Consider a slot-segment model [10], [11], [18] of wideband spectrum, where the monitored spectrum is divided into N non-overlapping narrowband sub-bands (also known as slots). Despite that the locations of these sub-bands are known in advance, their power spectral density (PSD) levels are unknown and dynamically varying, depending on whether they are occupied or not. Those temporarily unoccupied sub-bands are termed spectrum holes and they are available for opportunistic spectrum access by CR users. Suppose that the Nyquist-rate discrete form of received wideband signal at CR is denoted by an $N \times 1$ vector r_t and r_f is its frequency-domain discrete versions. The relationship between r_t and r_f is given by

$$r_f = F_N r_t \tag{1}$$

where F_N is the N -point unitary discrete Fourier transform (DFT) matrix.

Many investigations have shown that the radio spectrum is in a very low utilization ratio [1]-[4]. Recently a survey of a wide range of spectrum utilization across 6GHz of spectrum in some places of New York demonstrated that the maximum utilization was only 13.1% [8]. Therefore, it is reasonable, by using sparse or compressible structure, to model the received wideband signal of CR, at any given time and spatial region, which suggests that the received signal is inherently sparse or compressible in frequency domain, i.e. r_f is a sparse or compressible signal. This is exactly the motivation for introducing CS theory into WSS.

As to hierarchical access model [1], overlay spectrum sharing protocol is adopted in which CR user avoids transmitting at any occupied sub-bands. In this case, the

goal of spectrum sensing task is to determine which sub-bands are occupied and it is equivalent to detecting the support of sparse or compressible signal r_f .

For sparse case, the support of r_f is defined as follow:

$$\text{supp}(r_f) \triangleq \{1 \leq i \leq N : |r_f[i]| \neq 0\} \tag{2}$$

where $r_f[i]$ is the i -th element of r_f .

If r_f is a compressible signal, the concept should be replaced by $\alpha\%$ -energy support [13], which is defined as

$$\text{supp}(r_f) \triangleq \{1 \leq i \leq N : |r_f[i]|^2 > \rho\} \tag{3}$$

where ρ is the largest real value for which $\text{supp}(r_f)$ contains at least $\alpha\%$ of the signal energy.

Therefore the true occupancy status $d \in \{0, 1\}^{N \times 1}$, whose i -th element indicates whether the corresponding sub-band is occupied or not, is given by

$$d_i = \begin{cases} 1, & i \in \text{supp}(r_f) \\ 0, & i \notin \text{supp}(r_f) \end{cases} \quad i = 1, \dots, N \tag{4}$$

where $d_i = 1$ means that the i -th sub-band is occupied, while $d_i = 0$ means that the i -th sub-band is unoccupied.

III. TRADITIONAL COMPRESSED WIDEBAND SPECTRUM SENSING

According to CS theory, compressed measurements are in fact the linear projections of the received signal r_t onto a $M \times N$ measurement matrix $\Phi = [\phi_1^T, \dots, \phi_M^T]^T$ with $M < N$, where ϕ_1, \dots, ϕ_M are sensing waveforms. It makes sense that only M samples need to be measured instead of N samples. The compression ratio, which is defined as $\lambda = M/N$, reflects the reduced number of samples M collected at CR receiver, with reference to the number N needed in full-rate Nyquist sampling. The matrix format for collecting compressed measurements can be formulated as

$$y_t = \Phi r_t = \Phi F_N^{-1} r_f \tag{5}$$

where F_N^{-1} is the inverse DFT matrix.

Notice that, the inverse problem of (5) is an underdetermined problem and the high-dimensional signal r_f cannot be recovered from low-dimensional measurements y_t . However, it is shown in [5]-[7] that the reconstruction problem will stop being underdetermined if the matrix ΦF_N^{-1} satisfies restricted isometry property (RIP). It is proved that random matrix Φ whose entries are chosen according to a Gaussian distribution will satisfy the RIP with high probability provided M is sufficiently large. More significantly, the

RIP will be preserved for $\Phi\Psi$ if Φ is Gaussian random matrix and Ψ is arbitrary orthonormal basis.

The problem of reconstructing the spectrum estimate \hat{r}_f can be formulated as a combination of a l_0 norm minimization and a linear measurement fitting constraint. However, the exact solution of l_0 norm minimization is NP-hard and it requires an intractable combinatorial search. There are mainly two practical and tractable alternatives [6], [7]: greedy algorithms, such as various matching pursuits and convex relaxation algorithms. Both of them have advantages and disadvantages when applied to different scenarios. A brief assessment of their differences would be that convex relaxation algorithms require fewer measurements while greedy algorithms have less computation complexity.

An approximate solution that is largely used is obtained by solving the following convex relaxation leading to l_1 -norm minimization problem.

$$\begin{aligned} \hat{r}_f &= \arg \min_{r_f} \|r_f\|_1 \\ \text{s.t.} \quad &y_t = \Phi F_N^{-1} r_f \end{aligned} \quad (6)$$

In practice the received signal is inevitably polluted by noise. In this case, a conic constraint is required, i.e. the optimization problem in (6) needs to be changed to

$$\begin{aligned} \hat{r}_f &= \arg \min_{r_f} \|r_f\|_1 \\ \text{s.t.} \quad &\|y_t - \Phi F_N^{-1} r_f\|_2 \leq \varepsilon \end{aligned} \quad (7)$$

where ε bounds the amount of noise energy in the compressed measurements. Note that both of (6) and (7) are convex optimization problem [19] and a number of Matlab toolboxes, such as CVX [20], SeDuMi [21], l_1 Magic [22], can be used to efficiently solve the problem.

After we get spectrum estimate \hat{r}_f from (7), the energy detection [2], [3] can be used to make the decision on spectrum occupancy status. The i -element of the decision \hat{d} is made as follow:

$$\hat{d}_i = \begin{cases} 1, & \text{if } |\hat{r}_f(i)| \geq \eta \\ 0, & \text{if } |\hat{r}_f(i)| < \eta \end{cases} \quad \text{for } i = 1, \dots, N \quad (8)$$

where $\hat{r}_f[i]$ denotes the i -element of \hat{r}_f and η is a decision threshold which is chosen according to the desired probability of false alarm. After simple thresholding, the spectrum holes for DSA can be clearly given.

IV. THE PROPOSED COMPRESSED WIDEBAND SPECTRUM SENSING

It has been demonstrated that further exploitation of signal model, in addition to sparsity, would give birth to recovery performance enhancement [13]-[17]. As mentioned above, traditional CWSS exploits only the prior that the received signal is sparse or compressible in frequency domain, and does not assume any additional

knowledge about the unknown sparse spectrum. However in the practical implementation of wideband spectrum sensing, there could be other knowledge about the monitored wideband spectrum. Incorporating additional knowledge would potentially improve spectrum sensing performance. In this paper, partial occupancy status is assumed to be known in advance from coexisting narrowband detectors, other collaborative sensing nodes or remote database, and we study how to exploit such prior information to improve spectrum sensing performance.

Now we consider the spectrum reconstruction of r_f with partially known occupancy status from compressed measurements y_t . The part of sub-bands, whose occupancy status is known and index set is denoted by S , can be divided into two parts: the known occupied part indexed by S_o and the known unoccupied part indexed by S_u . Thus, S_o and S_u are disjoint and $S = S_o \cup S_u$. Therefore, a new CWSS model incorporating partially known occupancy status can be formulated as follow:

$$\begin{aligned} \hat{r}_f &= \arg \min_{r_f} \|r_f\|_1 \\ \text{s.t.} \quad &\|y_t - \Phi F_N^{-1} r_f\|_2 \leq \varepsilon \\ &d[i] = 1 \quad \text{for } i \in S_o \\ &d[j] = 0 \quad \text{for } j \in S_u \end{aligned} \quad (9)$$

where d is the true occupancy status.

According to the relationship between occupancy status of each sub-band and the support of r_f as stated in (4), the optimization problem in (9) can be reformulated as follow:

$$\begin{aligned} \hat{r}_f &= \arg \min_{r_f} \|r_f\|_1 \\ \text{s.t.} \quad &\|y_t - \Phi F_N^{-1} r_f\|_2 \leq \varepsilon \\ &S_o \subset \text{supp}(r_f) \\ &S_u \not\subset \text{supp}(r_f) \end{aligned} \quad (10)$$

Unfortunately, the support set of r_f is a thresholding function of r_f which is unknown before reconstruction because it is scenario dependent. However, according to the definitions of support set as stated in (2) and (3), it is easy to see that signal values on its support set are relatively large while those outside are relatively small, even close to zero. Besides, for weighted l_1 -norm minimization problem (6) and (7) are special cases where the weights are equal to 1), it has been shown that large weights can be used to discourage nonzero entries in the recovered sparse signal while small weights encourage nonzero entries [23].

According to those analyses, a particular weighting strategy in weighted l_1 minimization reconstruction algorithm is proposed, in which small weights are imposed on the known occupied sub-bands which encourage relatively large entries in the recovered signal,

at the same time, large weights on the known unoccupied ones which encourage relative small entries. Problem stated in (10) then becomes

$$\begin{aligned} \hat{\mathbf{r}}_f &= \arg \min_{\mathbf{r}_f} \|\mathbf{W}\mathbf{r}_f\|_1 \\ \text{s.t.} \quad &\|y_t - \Phi\mathbf{F}_N^{-1}\mathbf{r}_f\|_2 \leq \varepsilon \end{aligned} \quad (11)$$

where the weight matrix $\mathbf{W} = \text{diag}(w_1, w_2, \dots, w_N)$ with

$$w_k = \begin{cases} \tau_l, & \text{if } k \in S_u \\ \tau_s, & \text{if } k \in S_o \text{ for } k=1, \dots, N \\ 1, & \text{if } k \notin S \end{cases} \quad (12)$$

and τ_l, τ_s denote pre-specified large and small constant, respectively. When spectrum estimate $\hat{\mathbf{r}}_f$ is obtained from (11), decision on spectrum occupancy status can be also made via thresholding in (8). It is noteworthy that the weighted formulation of l_1 minimization as stated in (11) is still a convex programming in which the problem remains simple and it effectively exploits the prior information on partially known occupancy status.

Apparently, problem (11) will be reduced to the traditional case without any prior information when $S_u = \emptyset, S_o = \emptyset$. In addition, problem (11) will be reduced to modified basis pursuit denoising (modified BPDN [14]) when $S_u = \emptyset$ and $\tau_s = 0$. In modified BPDN, partial support is assumed to be available, and the sparseness-inducing (l_1 -norm) term only contains entries outside the known support. In essence, the main difference between modified BPDN and (11) is that modified BPDN does not consider any additional knowledge outside the support, while (11) does.

V. SIMULATION RESULTS

In this section, simulation results are provided to illustrate performance of the proposed CWSS-PKOS approach with ideal and non-ideal prior information. First, the simulation setup and relevant performance metrics are described. Second, the performance of proposed approach with ideal prior information is evaluated by comparing with the traditional CWSS approach without any prior information as reference approach. Third, we consider cases in which some incorrect prior information exists in the known occupied part S_o and the known unoccupied part S_u respectively.

A. Simulation Setup and Performance Metrics

Consider a monitored wide band partitioned into $N=128$ equal-bandwidth sub-bands and 24 among all sub-bands are randomly occupied by LUs, that is $|\text{supp}(\mathbf{r}_f)| = 24$. Suppose that the amplitudes of elements in $\text{supp}(\mathbf{r}_f)$ are generated according to Rayleigh distribution. The received signal is corrupted by additive white Gaussian noise (AWGN). The signal to noise ratio (SNR) is defined as the ratio of the average received

signal to noise power over the entire wideband spectrum and is set to vary from -20dB to 30dB. For compressed sensing, the number of compressed samples M is set to range from 36 to 96. As to selecting parameter of weight matrix in (11), we set $\tau_l = 10^3$ to discourage nonzero entries in the reconstructed signal, and $\tau_s = 0$ to encourage nonzero entries which can also reduce the number of variables in objective function of (11) and then reduce the computational complexity of solving problem (11). For notational simplicity, we introduce N_o and N_u to denote the size of the known occupied part S_o and the known unoccupied part S_u respectively. That is, $N_o = |S_o|$ and $N_u = |S_u|$.

For the spectrum hole detection problem, performance metrics of interest are the probabilities of detection P_d and false alarm P_{fa} , which are evaluated by comparing the estimated occupancy status $\hat{\mathbf{d}}$ with the true occupancy status \mathbf{d} over all sub-bands, as follows:

$$\begin{aligned} P_d &= \frac{\|\mathbf{d} \& \hat{\mathbf{d}}\|_0}{\|\mathbf{d}\|_0} \\ P_{fa} &= \frac{\|(\sim \mathbf{d}) \& \hat{\mathbf{d}}\|_0}{\|\sim \mathbf{d}\|_0} \end{aligned}$$

where $\&$ and \sim denote bitwise AND and bitwise NOT respectively. The l_0 norm function here is used to calculate the number of nonzero elements.

B. CWSS-PKOS with Ideal Prior Information

In this subsection, we consider cases in which prior information is ideal, namely all components of S_o and S_u are correct. To show the improvement in detection performance obtained by incorporating partially known occupancy status and to verify the effectiveness of CWSS-PKOS in exploiting such prior information, variations of detection probability P_d versus SNR and compression ratio λ are plotted in Fig. 1 and 2 respectively. In the following figures, each point on the curve is the average of the detection probabilities over 800 Monte Carlo trials. In addition, the opportunities for DSA will lose if the desired false alarm probability is set to be too small while determining threshold for making the decision on spectrum occupancy status. Therefore, we apply the receiver operation characteristic (ROC) curve to illustrate the tradeoff between the probabilities of detection and false alarm. ROC curves for $P_{fa} \in [0.01, 0.11]$ are given in Fig. 3 since this regime of false alarm probability is of practical interest for achieving rational opportunistic throughput in CR. It has been shown in Section IV that the proposed CWSS-PKOS approach will be reduced to the traditional CWSS without any prior information when $S_u = \emptyset, S_o = \emptyset$. Therefore, in the following part, curves for

$N_o = 0, N_u = 0$ depict the detection performance of traditional CWSS.

As shown in Fig. 1, 2 and 3, detection probabilities of proposed CWSS-PKOS when partial occupancy status is available (curve 2, 3 and 4) are greater than that of traditional CWSS (curve 1). This suggests that incorporating partially known occupancy status, both the known occupied part and the known unoccupied part, indeed enables improvement in detection performance. Moreover, it is seen in Fig. 1 that compared to the improvement in high SNR regime, improvement in low SNR regime is much larger. This is because at high SNRs, such as 25dB and 30dB, the recovered signal obtained by solving (11), even when without any prior information, is quite closed to the original one.

In simulations, there are 24 occupied sub-bands and 104 unoccupied sub-bands, thus $N_o = 12$ and $N_u = 52$ represent 50% of the occupied and unoccupied sub-bands respectively. In Fig. 1, Fig. 2 and Fig. 3, we can see that there is a big gap between curve 2 and curve 3. It suggests that, for a same proportion of occupied and unoccupied sub-bands, performance improvement provided by the former is larger than that provided by the latter.

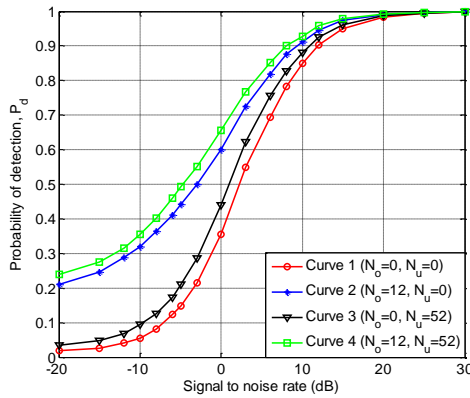


Figure 1. Detection probability versus SNR for various combinations of N_o and N_u ($\lambda = 0.5$, $P_{fa} = 0.01$)

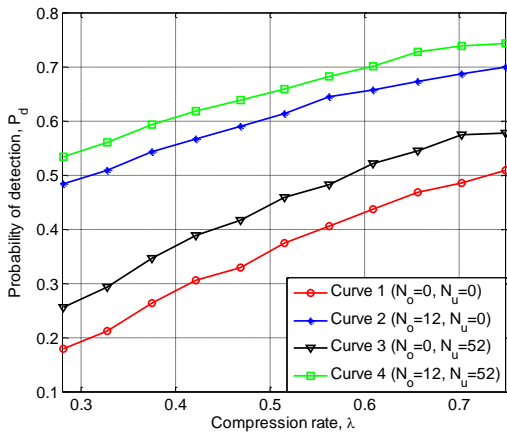


Figure 2. Detection probability versus compression ratio λ for various combinations of N_o and N_u (SNR=0dB, $P_{fa} = 0.01$)

Here we give an explanation for those improvements introduced by PKOS in qualitative analysis. It is well-known that the solution of traditional CS reconstruction is the one with the minimum number of nonzeros among infinite data-consistency candidates. When partial occupancy status is known in advance, the candidates will be restricted in a signal space smaller than that in traditional CS. Under the given simulation conditions stated in subsection V-A, traditional CS needs to search for solutions in $C(128,40)$ possible 24-dimensional subspaces, where $C(n,r)$ denotes the number of combinations of 'n' things selected 'r' at a time. If the occupied sub-bands are partially known such that 12 sub-bands are known to be occupied, we only need to search for solutions in $C(116,12)$ possible 12-dimensional subspaces. If 52 sub-bands are known to be unoccupied, we need to search for solutions in $C(76,24)$ possible 24-dimensional subspaces. Clearly, the search space is reduced as long as occupied or unoccupied sub-bands are partially known. For a same proportion of occupied and unoccupied sub-bands, the reduction is even more for the former since the number of unoccupied sub-bands is more than that of occupied ones. This reduction enables reduction in number of measurements required to achieve a given CS reconstruction quality or enables improvement in the quality of CS reconstruction given the same number of measurements, therefore gives birth to the improvement in detection performance of CWSS.

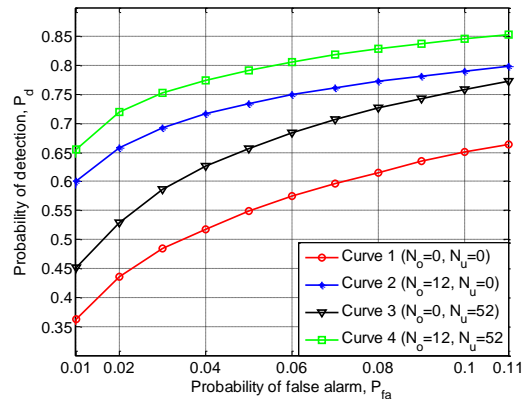


Figure 3. Receiver Operating Characteristic (ROC) curves for various combinations of N_o and N_u (SNR=0dB, compression rate $\lambda = 0.5$)

C. CWSS-PKOS with Non-Ideal Prior Information on S_o

In the procedure of acquiring partial occupancy status, especially from collaborative sensing nodes, incorrect prior information is unavoidable. In this subsection, we consider cases in which prior information on the known occupied part S_o is non-ideal, namely some sub-bands corresponding to S_o are unoccupied. To facilitate analysis, we assume $S_u = \emptyset$ in this subsection. Let $N_{o,c}$ and $N_{o,w}$ denote the size of correct and incorrect prior information on S_o , respectively. Thus, $N_o = N_{o,c} + N_{o,w}$.

Fig. 4 and 5 respectively illustrate the variations of probability of detection P_d versus SNR and compression ratio λ for various combinations of $N_{o,c}$ and $N_{o,w}$. It is shown in curve 2, 3, 4 and 5 that, as the size of incorrect prior information $N_{o,w}$ increases, the curves corresponding to detection performance are shifted to the bottom, which indicates a consistent degradation in detection performance when more incorrect prior information exist in S_o . In Fig. 4 and 5, it is noteworthy that the detection probabilities plotted in curve 2, 3 and 4 are greater than that plotted in curve 1 for all value of SNR and compression ratio λ , indicating an improvement in regard to the traditional CWSS provided the size of correct prior information is larger than that of incorrect prior information. This observation is corroborated by the comparison between curve 5 and curve 1 in Fig. 4 and 5. Comparing curve 1 and 5 in Fig. 4, it is interesting that, even when most of the sub-bands corresponding to S_o are unoccupied, the improvement in detection performance is still existed at relatively low SNR.

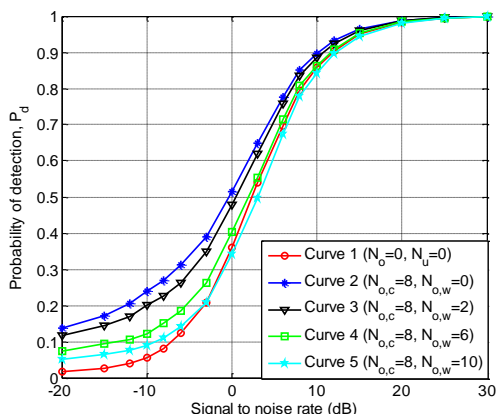


Figure 4. Detection probability versus SNR for various combinations of $N_{o,c}$ and $N_{o,w}$ ($\lambda=0.5$, $P_{fa}=0.01$ and $N_u=0$)

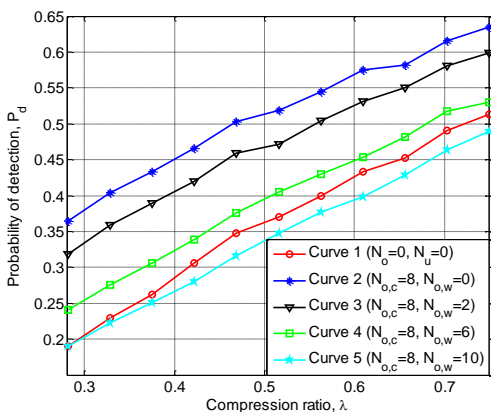


Figure 5. Detection probability versus λ for various combinations of $N_{o,c}$ and $N_{o,w}$ (SNR=0dB, $P_{fa}=0.01$ and $N_u=0$)

According to the observations stated above, it is safe to conclude that the proposed CWSS approach is robust to

errors in the known occupied part S_o . This suggests that, in order to contain more correct prior information in S_o , more greedy strategy can be adopted while obtaining prior information on S_o , as long as $N_{o,c}$ is sufficiently larger than $N_{o,w}$.

D. CWSS-PKOS with Non-Ideal Prior Information on S_u

In this subsection, we exemplify cases in which prior information on the known unoccupied part S_u is non-ideal, namely some sub-bands corresponding to S_u are occupied. Similar to the above subsection, we set $S_o = \emptyset$ in this subsection and use $N_{u,c}$ and $N_{u,w}$ to denote the size of correct and incorrect prior information on S_u , respectively.

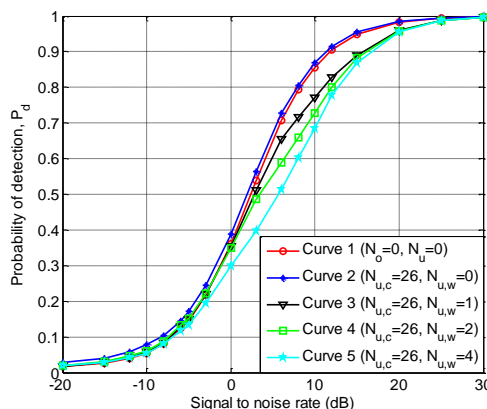


Figure 6. Detection probability versus SNR for various combinations of $N_{u,c}$ and $N_{u,w}$ ($\lambda=0.5$, $P_{fa}=0.01$ and $N_o=0$)

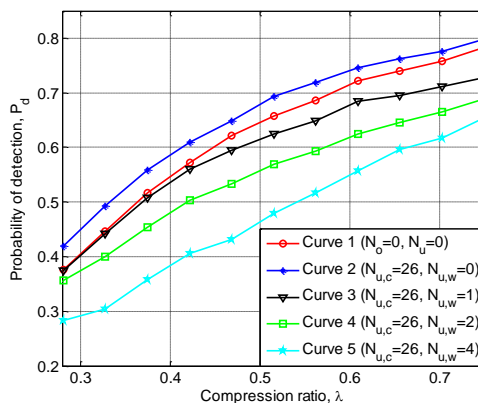


Figure 7. Detection probability versus λ for various combinations of $N_{u,c}$ and $N_{u,w}$ (SNR=5dB, $P_{fa}=0.01$ and $N_o=0$)

Fig. 6 shows variations of detection probability versus SNR for different combinations of $N_{u,c}$ and $N_{u,w}$. For a given SNR, the probability of detection decreases as the size of incorrect prior information $N_{u,w}$ increases, which indicates a consistent reduction in this quantity when more incorrect prior information exist in S_u . By

comparing cases $(N_{u,c} = 26, N_{u,w} = 1)$, $(N_{u,c} = 26, N_{u,w} = 2)$, $(N_{u,c} = 26, N_{u,w} = 4)$ with the case $(N_o = 0, N_u = 0)$, it can be observed that detection performance degrades obviously even when a small amount of errors exist in S_u , in comparison with that of traditional CWSS. Meanwhile this degeneration in detection performance occurs for all value of compression ratio, which is corroborated by variations of detection probability versus compression ratio for various combinations of $N_{u,c}$ and $N_{u,w}$ depicted in Fig. 7.

It can be seen from Fig. 6 and 7 that the detection performance of proposed approach is sensitive to errors in the known unoccupied part S_u . This suggests that, in order to avoid containing incorrect prior information in S_u , more conservative strategy should be adopted while obtaining S_u .

VI. CONCLUSIONS

We studied the problem of compressed wideband spectrum sensing when partial occupancy status is known. In order to incorporate partially known occupancy status, a new CWSS model is presented and then a particular weighting strategy in the reconstruction algorithm based on weighted l_1 minimization is proposed to solve the proposed model. Simulation results demonstrate that the introduction of partially known occupancy status, both the known occupied part and the known unoccupied part, enables improvement in detection performance, and the proposed CWSS-PKOS approach can exploit such prior information effectively. Moreover, performance improvement provided by partial occupied sub-bands is larger than that provided by the same proportion of the unoccupied sub-bands. It is also shown via simulations that the proposed approach is sensitive to errors in the known unoccupied part, however robust to errors in the known occupied part. These observations indicate that greedy and conservative strategy should be adopted to obtain prior information on the occupied and unoccupied sub-bands respectively.

ACKNOWLEDGMENT

This work was supported by the National Major Scientific and Technological Special Project (No. 2012ZX03006003-004).

REFERENCES

- [1] Q. Zhao and B. M. Sadler, "A Survey of Dynamic Spectrum Access," *Signal Processing Magazine, IEEE*, vol. 24, no. 3, pp. 79–89, 2007.
- [2] I. F. Akyildiz, F. Lo Brandon, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Communication*, vol. 2011, no. 4, pp. 40–62, 2011.
- [3] L. Lu, X. Zhou, U. Onunkwo, and G. Y. Li, "Ten years of research in spectrum sensing and sharing in cognitive radio," *Wireless Communications and Networking, EURASIP Journal on*, pp. 1–16, 2012.
- [4] J. N. Laska, W. F. Bradley, T. W. Rondeau, K. E. Nolan, and B. Vigoda, "Compressive sensing for dynamic spectrum access networks: Techniques and tradeoffs," in *2011 IEEE International Symposium on Dynamic Spectrum Access Networks DySPAN*, 2011, pp. 156–163.
- [5] R. G. Baraniuk, "Compressive Sensing [Lecture Notes]," *Signal Processing Magazine, IEEE*, vol. 24, no. 4, pp. 118–121, 2007.
- [6] E. J. Candes and M. B. Wakin, "An Introduction To Compressive Sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 21–30, 2008.
- [7] M. A. Davenport and M. F. Duarte, "Introduction to Compressed Sensing," *Electrical Engineering*, vol. 93, no. 3, pp. 1–68, 2011.
- [8] M. Marcus, J. Burtle, and N. Mcneil, "Federal communications commission spectrum policy task force report of the spectrum efficiency working group," *Spectrum*, vol. 1, no. 1, p. 37, 2002.
- [9] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Conference Record of the Thirty Eighth Asilomar Conference on Signals Systems and Computers 2004*, vol. 1, pp. 772–776.
- [10] Z. Tian, "Compressed Wideband Sensing in Cooperative Cognitive Radio Networks," in *IEEE Global Telecommunications Conference GLOBECOM*, 2008, no. 1, pp. 1–5.
- [11] F. Zeng, C. Li, and Z. Tian, "Distributed compressive spectrum sensing in cooperative multihop cognitive networks," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 1, pp. 37–48, 2011.
- [12] C. J. Miosso, R. von Borries, M. Arguez, L. Velazquez, C. Quintero, and C. M. Potes, "Compressive Sensing Reconstruction With Prior Information by Iteratively Reweighted Least-Squares," *Signal Processing, IEEE Transactions on*, vol. 57, no. 6, pp. 2424–2431, 2009.
- [13] N. Vaswani and W. Lu, "Modified-CS: Modifying Compressive Sensing for Problems With Partially Known Support," *Signal Processing, IEEE Transactions on*, vol. 58, no. 9, pp. 4595–4607, 2010.
- [14] W. Lu and N. Vaswani, "Modified Basis Pursuit Denoising (modified-BPDN) for noisy compressive sensing with partially known support," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 3926–3929.
- [15] C. Miosso, R. von Borries, and J. Pierluissi, "Compressive Sensing with Prior Information-Requirements and Probabilities of Reconstruction in l_1 -Minimization," *Signal Processing, IEEE Transactions on*, vol. 61, no. 9, pp. 2150–2164, 2013.
- [16] J. Scarlett, J. S. Evans, and S. Dey, "Compressed Sensing With Prior Information: Information-Theoretic Limits and Practical Decoders," *Signal Processing, IEEE Transactions on*, vol. 61, no. 2, pp. 427–439, 2013.
- [17] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-Based Compressive Sensing," *Information Theory, IEEE Transactions on*, vol. 56, no. 4, pp. 1982–2001, 2010.
- [18] Z. Quan, S. Cui, A. H. Sayed, and H. V. Poor, "Optimal Multiband Joint Detection for Spectrum Sensing in Cognitive Radio Networks," *Signal Processing, IEEE Transactions on*, vol. 57, no. 3, pp. 1128–1140, 2009.
- [19] S. Boyd, L. Vandenberghe, *Convex Optimization*. Cambridge University Press, New York, 2008.
- [20] M. Grant, S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming," <http://cvxr.com/cvx/>
- [21] J. F. Sturm, "Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11, no. 1, pp. 625–653, 1999.

- [22] E. C and J. Romberg, "l1-magic: Recovery of sparse signals via convex programming," 2005.
- [23] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing Sparsity by Reweighted l1 Minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877-905, 2008.



Zha Song was born in 1987. He received the B.S. degree in communication and information system and M.S. degrees in signal and information processing from the Electronic Engineering institute (EEI), Hefei, China, in 2007 and 2010, respectively. He is now a Ph. D student in National University of Defense Technology (NUDT), Changsha, China.

His current research interests include spectrum sensing, distributed compressed sensing.



Huang Jijun was born in 1970. He received the B.S. and M.S degrees in electromagnetic field and microwave technologies and the Ph.D. degree in electronic science and technology from the National University of Defense Technology (NUDT), Changsha, China, in 1993, 1997, and 2005, respectively. He is currently an Associate Professor in

NUDT.

His current research interests include spectrum management and electromagnetic compatibility.

Li Ning was born in 1983. He received the B.S. degree in communication and information system and M.S. degrees in military equipment from the Electronic Engineering institute (EEI), Hefei, China, in 2006 and 2010, respectively. He is now an Electronics Engineer in PLA Unit 77108, Chengdu, China.

An IP-Traceback-based Packet Filtering Scheme for Eliminating DDoS Attacks

Yulong Wang and Rui Sun

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Email: wyl@bupt.edu.cn, sunruixsh@gmail.com

Abstract—Distributed Denial-of-Service (DDoS) is still an important security challenge for computer networks. Filter-based DDoS defense is considered as an effective approach, since it can defend against both victim-resource-consumption attacks and link-congestion attacks. However, the high possibility of false positive and the huge consumption of router resources reduce the practicality of existing filter-based approaches. In order to solve this problem, we propose a new mechanism to efficiently eliminate the impact caused by DDoS attacks. We utilize the IP traceback results to obtain an attack graph that contains the candidate filtering routers. Taking the different filtering performance of the routers in the attack graph into consideration, we propose a filtering scheme to determine a small set of filtering routers that would increase filtering performance and reduce false positive. Simulation results based on real-world network topologies demonstrate that the proposed scheme can reduce the damage caused by DDoS attacks effectively and maintain the loss of normal traffic within an acceptable level.

Index Terms—DDoS Attack, Packet Filtering, IP Traceback

I. INTRODUCTION

DDoS attacks have been one of the most hazardous threats to Internet. These attacks, which generate enormous packets by directing thousands of comprised hosts, can easily exhaust the computing and network resources of a victim. As an example, in August 2013, China's Internet was hit by a major distributed denial of service (DDoS) attack that briefly disrupted and slowed access to sites in the.cn domain [1]. For another instance, a large scale DDoS attack was launched on March 4th, 2011, which targeted about 40 major websites [2] in South Korea, which means that DDoS attacks are still prevalent in the current Internet.

Many counter measures have been developed to defeat DDoS attacks. According to the location of deployment, these measures can be classified into three categories: victim-end protection schemes, source-end protection

schemes and intermediate-router protection schemes. A victim-end protection scheme can protect a victim from DDoS attack by blocking attack flows near the victim, but it is not able to defend flooding attacks aiming to consume network bandwidth. Source-end defense schemes are able to protect network bandwidth from being exhausted. However, in a large-scale DDOS attack, attack sources may include a large number of compromised hosts. The critical issue of this approach is how to deploy the schemes for the majority of end hosts since control and management of such schemes will be a huge burden [3].

Consequently, intermediate routers should be the most effective locations to defend against both victim resource attacks and bandwidth flooding attacks [4, 5, 6, 15, 16]. Therefore, we need an approach that installs filters on intermediate filtering routers to block undesired network flows. A filter is a rule or a command which can be installed on a router to manage a specified attack flow. A filtering router is an intermediate router with built-in filters. A good intermediate-router-based DDoS defense scheme should provide solutions to the following problems: 1) how to effectively identify attack paths, 2) how to determine the optimal filtering routers, 3) how to reduce the loss of normal traffic.

To address these challenges, we propose a new DDoS defense scheme. In our scheme, routers employ a traceback method to identify which routers would forward the undesired flow. Moreover, by analyzing the attack graph and the filtering cost effective, we propose a genetic algorithm for determining the locations of filtering routers.

We found that most existing algorithms [4, 5, 6, 15, 16] combine the filtering function with the attack path identifying function in a high coupling degree. In other words, these filtering methods are relying on specific attack path identifying methods. The heavy burden on the routers brought by these specific attack path identifying algorithms makes the ISP reluctant to deploy these filtering system on the internet. However our proposed scheme decouples filtering functions from attack path identifying functions, which gives ISP more freedom in selecting the attack path identifying method.

In the aspect of defense ability, advantages and contributions of our scheme focus on the following aspects.

Manuscript received September 26, 2013; revised December 15, 2013; accepted January 2, 2014.

This work was supported in part by the Youth Scientific Research and Innovation Plan of Beijing University of Posts and Telecommunications (2013RC1101) and the Important national science & technology specific projects: Next-generation broadband wireless mobilecommunications network (2012ZX03002008-002-03).

The proposed scheme not only protects the victim's resources, but also protects the bandwidth of the bottleneck network link from being exhausted.

Our scheme can reduce the loss of normal traffic since the scheme filters with a lower percentage.

The scheme takes filtering resource consumption into consideration and makes use of each filtering router in an effective manner.

The structure of the paper is arranged as follows: the second section describes the related works. The third section presents the overall architecture of the system model. The fourth section describes the filtering router set determination scheme in detail. Experimental results are presented in Section V. In Section VI we carry out a discussion. Section VII summarizes this paper.

II. RELATED WORK

Many research works on router-based DDoS defense had been done in recent years [7, 8, 9, 11, 12, 13, 14, 15, 16, 20].

Router-based Packet Filtering (RPF) by Park and Lee [7] is based on the principle that for each link in the core of the Internet, there is only a limited set of source addresses from which traffic on the link could have originated. If an unexpected source address appears in an IP packet on a link, then it is assumed that the source address has been spoofed, and hence the packet should be filtered. DPF [8], SAVE [9] and BASE [10] propose similar anti-spoofing mechanisms.

Capability-based defense approaches restrict the bandwidth of each sender. In SIFF [11], the end-host classifies network traffic into privileged traffic and unprivileged traffic, then selectively filters individual flows so as to protect privileged traffic from DDoS attack. Yau et al. [12] also propose a feedback control scheme on the router to throttle the attacking traffic with max-min fairness. This scheme can proactively rate-limit the attack traffic before it reaches the victim, and therefore forestalls the DDoS attack.

Since Savage et al. [13] proposed the packet marking method for IP traceback, several DDoS defense methods based on path identification are proposed. Generally, these methods utilize routers to mark packets as they travel through the network. When a packet reaches an end-host, the end-host can take correspondent actions using the marking in it. PI [14] verifies an attack path by using TTL values in router markings and filtering the packets with the same markings at end-hosts. Minh Sung and Jun Xu [15] use a modified IP traceback method that moves the filtering location to intermediate routers. Dongwon Seo and Heejo Lee [16] modified a probabilistic packet marking method to propagate filtering locations.

III. SYSTEM WORK

A. Scenario Assumption

In this section, we make some assumptions on the attack scenario so as to design our system properly.

DDoS attack affects both the victim and the network link: The victim is suffering from a DDoS attack with two

major impacts. Firstly, the victim has limited resources for processing the incoming packets. The victim's resources, such as CPU and memory, will be exhausted, and then the victim will be unable to serve for normal traffic. Secondly, consumption of network bandwidth will result in legitimate flows being blocked.

Filtering resource is contained while serving large numbers of victims in need of protection: In order to maintain an acceptable throughput, the number of filters installed on a router should be contained yet enough to protect the victims whose number maybe very large.

Attack is detected first: The IDS (Intrusion Detection System) on the victim side monitors traffic patterns and can identify attack traffic. Nowadays, many servers have installed Host-based IDS for such a purpose.

Network topology can be obtained: The victims (with HIDS installed) should be able to obtain a map of upstream routers. This assumption is relatively stronger and will be explained in detail in section VI.

B. System Goals

In this section, we describe the goals of our work.

Identifying attack paths: One feature of most DDoS attacks is the use of fork source IP addresses. For an intermediate-router-based filtering method, a victim should identify the path that the attack traffic traverses. But the stateless and non-authentication property of the routing network makes it difficult to locate the real sources and paths of network attacks. In our scheme, IP traceback methods are adopted to reconstruct the attack path. Intermediate routers are required to provide additional information for identifying attack paths.

Selecting optimal routers for installing filters: Filters should be installed on optimal routers in the network to obtain a good balance between filtering cost and packet transferring performance. There is a tradeoff between filtering cost and filtering efficiency. A good filter deployment scheme should provide a proper balance between victim protection, attack flow blocking and filtering resource conservation.

C. System Model

Our system is composed of three components: the Attack Path Reconstruction (APR) module, the Filtering-router Set Determination (FSD) module and the Scheduled Packet Filtering (SPF) module. The relationship of these modules with the actual physical devices is shown in Fig 1.

APR module. This model uses IP traceback schemes to reconstruct attack graphs. In our system, we prefer to use the PPM (Probabilistic Packet Marking) [13] method to achieve the goal because PPM is the most commonly used method for reconstructing DDoS attack paths. When packets are transferred in the network, each router needs to put a marking into certain IP header fields of the packets. Once identified as an attack packet, the marking of the packet will be used by the victim to reconstruct the attack path.

FSD module. This module runs on the victim and implements the following functions: (1) Analyzing the

attack graph consisting of all the detected attack paths. (2) Determining the set of routers that should install filters.

SPF module. This module is running on all the filtering-routers. The module mounts filters on the packet processing routine to block the specified packets. A self-adaptive filter management method is used here for filter rewinding.

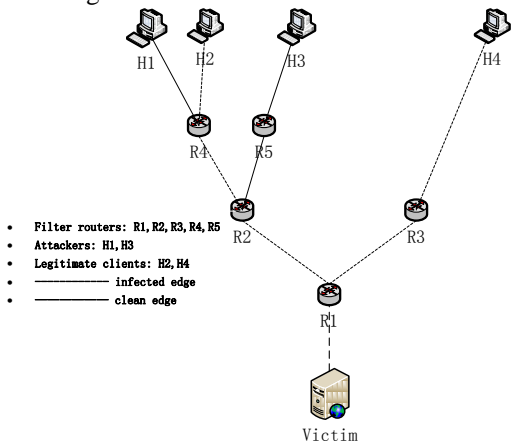


Figure 1. The locations of APR, FSD and SPF in a network

In a simplified attacking scenario shown in Fig. 1, H1 and H3 are attackers, while H2 and H4 are legitimate users. The APR modules and the SPF modules are deployed on the intermediate routers (R1, R2, R3, R4, R5). When attack occurs, the APR module will use IP traceback method to reconstruct the attack graph. After that, the FSD module will recommend a proper set of filtering routers to block the attack. Furthermore, the SPF module can be used to un-mount the filters when disappearance of attack flows is detected so as to avoid filtering legitimate flows.

IV. MODULE DESIGN

In this section, we describe the design of the system modules in detail.

A. Design of APR Module

APR module adopts the PPM method for identifying attack paths. PPM is a well-known approach for IP traceback. In our work, we adjust some important parameters of the PPM method to make it more suitable for our scheme.

Each network router marks its unique information into the IP headers of outgoing packets with probability p. These markings collectively allow the victim to reconstruct the attack tree, which consists of the network edges that the packets sent from the attackers have traversed. Fig. 2 shows the format of the marking.

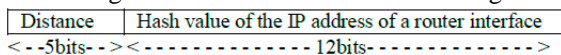


Figure 2. The format of the marking

When a packet arrives at a router, the router decides whether it will overwrite the current marking with probability p. If it decides to overwrite the marking, the distance field will be reset to zero. If the router decides to

keep the marking, the distance field will simply increase by one.

This 17-bit marking can be placed in two IP header fields: the identification field plus an unused bit next to the DF (Don't Fragment) bit. Because this space, called the marking field, is too small for holding IP addresses representing a network edge, we need to adopt an encoding method here. There are many existing IP traceback schemes that use different encoding methods, among which we choose the Advanced Marking Scheme (AMS) proposed by Song and Perrig [17]. The advantage of AMS is that it can provide fast and accurate reconstruction of attack paths even if there is a large number of attackers. AMS requires the assumption that victims should know the upstream routers, as we have proposed in section III A.

AMS uses eight different hash functions H_i ($1 \leq i \leq 8$) to encode network edges. When a packet goes through an edge e , i will be chosen randomly from 1 to 8 and $H_i(e)$ will be written in the IP header of the packet. When the victim begins to reconstruct the attack graph, whether an edge is on the attack path can be determined if and only if the victim has received at least k out of 8 values of $\{H_i(e) | 1 \leq i \leq 8\}$. k can be 6 to 8. A larger k will improve the accuracy but will need more time for attack graph reconstruction.

We use AMS as the fundamental traceback scheme of our work. In [17], simulation has been done to estimate the average number of packets, N , to be used for the victim to reconstruct an attack path. Results show that N is an approximate linear function of the length of the attack path, and it takes about 4000 packets to reconstruct a 30-hop attack path with marking probability $q=0.04$ and $k=7$. For a host attacking at 300kbps (i.e. sending 300 packets per second and the average size of a packet is 1 kilo-bits), it will take no more than 14 seconds to reconstruct the attack path, which is fast enough. Moreover, for our filtering scheme, we do not need to reconstruct such a long path. Our simulation shows that as long as most of the infected edges within 10 hops are reconstructed, our filtering scheme will achieve our goal.

B. Design of FSD Module

FSD module runs on the victim. Once the attack graph is reconstructed by the ARP module, the FSD module uses the attack graph to determine the set of appropriate routers and installs filters on them. This task is executed periodically after a DDoS attack has been detected.

The challenge here is how to determine the optimal filtering routers. If we simply select the nearest router to every attack source as a filtering router, we can block all attack flows. However, as we have emphasized in section II, in a large-scale DDoS attack, attack sources may consist of millions of compromised hosts, thus installing filters on all the routers will consume too much filtering and communication resources. We will benefit from carefully selecting a relatively small set of routers for filtering.

In this section we deduce the formulas to evaluate the cost and performance of filtering routers and put forward

a genetic-based algorithm to determine the set of filtering routers for defense.

1) Modeling

We propose an attack model to describe the current attack graph. Consider an attack whose target is a node, v , which refers to the victim. An attack source is referred to as the router that is nearest to the attack host. The set of attack sources will be referred to as A . An attack event can be represented by (A, v) . For each attack nodes a_i in A , $P(a_i, v)$ represents the attack path of a_i , which consists of a sorted list of router nodes from a_i to v . An attack graph can be denoted by $AG(A, v)$, which is a tree with v as the root and consists of all the attack paths from all attack sources.

For example, in Fig. 3, we demonstrate a simplified network in which a victim is under attack. There are five attack sources: a_1, a_2, a_3, a_4 and a_5 . The attack path from a_4 is $(a_4, r_4, r_8, r_{12}, r_{14}, \text{victim})$, and $AG(A, v)$ is presented with dotted lines.

S represents the set of filtering routers in the graph. Given a particular attack event (A, v) , S can be partitioned into two subsets: the positive nodes and the negative nodes. The positive nodes are on the attack graph of (A, v) while the negative nodes are not. For a given set of filtering routers S and the attackers A , we denote the collection of filter nodes associated with an attack event as $FG(A, v)$. The problem to solve in our scheme is how to choose a set of co-operative filtering nodes that is distributed in the network. We denote the selected set of co-operative filtering nodes as $BS(A, v)$ (i.e. the set of filtering routers for blocking attack flows). Obviously, any selected BS must be a subset of $FG(A, v)$.

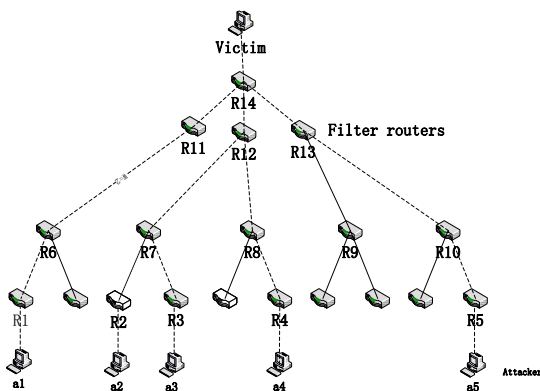


Figure 3. An attack graph example

2) Cost and Performance Evaluation

We observed that different filter nodes can play different roles in achieving the goal. For example, in Fig. 3, r_1 and r_{11} are both filtering routers on the same attack path $P(a_{11}, v)$. If we install filters on r_{11} , two links (i.e. $r_{11} - r_{14}$ and $r_{14} - \text{victim}$) will be protected from the attack flow, thus reducing the possibility of congestion on these links. However, if we install filters on r_1 instead, two more edges (i.e. $r_1 - r_6$ and $r_6 - r_{11}$) will be protected, thus strengthening the protection. For another example, r_2 and r_7 are both on the attack path $P(a_2, v)$. r_3 and r_7 are both on the attack path $P(a_3, v)$. If we want to block both attack flows with the least number of filtering routers, r_7 rather

than r_2 and r_3 should be a better choice. In order to choose the right BS for blocking attack flows, it is important to find a way to estimate the filtering performance of each filtering router on the attack graph. The following are two desirable characteristics that should be taken into consideration.

Attack Flow Blocked Percentage (AFBP)

This metric refers to the percentage of attack flows a filtering router can block. In Fig. 3, there are five attack flows detected in the attack graph and r_{12} can block three of them. Thus the AFBP of r_{12} is 60%. We use equation (1) to calculate the AFBP value of a filtering router.

$$AFBP = \frac{|\{P \mid r \in P(a_i, v)\}|}{|A|} \times 100\% \quad (1)$$

Link Protection Percentage (LPP)

In some situations, the attacker will use a lot of useless messages to exhaust transmission resources in order to force the victim host to lose the ability to provide services. Therefore, link bandwidth protection is another goal of our scheme. It is not easy to evaluate the effect of DDoS attacks on bandwidth. To achieve quantization, we denote a unit of bandwidth damage as one attack flow occupying one link. For example, in Fig. 3, the attacker a_1 causes five units of damage because the attack flow from a_1 has occupied five links. By gathering the bandwidth damage of all the attackers, the damage of an attack event (donate as D) can be calculated in whole. We use LPP to measure a filtering node's capability for protecting link resources, as shown in equation (2). $Dist(x, y, P)$ denotes the number of links between node x and node y on the attack path P . D denotes the original attacking damage before installing the filters.

$$LPP = \frac{\sum_{r_i \in P(a_i, v)} Dist(r_i, v, P(a_i, v))}{D} \quad (2)$$

It is worth noticing that once we are allowed to choose more than one filter nodes, different nodes may affect each other on filtering performance. For example, in Fig. 3, if we have chosen r_1 as a filtering router, r_6 would lose half of its original filtering performance because the attack flow would have been forestalled by r_1 . For this reason, the determination of BS is not a simple score-ranking problem. For evaluating the performance of BS , We should think of the nodes in BS as one atomic unit when estimating their total performance metrics. In equation (3), we define a new metric with BS as the parameter

$$AFBP(BS) = \frac{|\{p \mid p(a_i, v) \cap BS \neq \varphi \wedge a_i \in (A, v)\}|}{|A|} \quad (3)$$

While we can use equation (3) to evaluate the performance of a BS , our scheme still needs a method to estimate the cost of BS . Firstly, we should evaluate the cost of router resources including memory resources and computing resources that are used for filtering. This cost is proportional to the number of filtering routers used.

The number of nodes included in BS is denoted as $NF(BS)$, and this value can be used to evaluate the filtering resource cost. Secondly, the possible loss of legitimate flows should be calculated. When an attack flow is blocked by a filtering node, legitimate flows that take the same path towards the victim may also be filtered. The filtering of legitimate flows is considered as false positive. False positive reflects the adverse effect of BS on legitimate network flows. Since we do not know the distribution of legitimate flows, designing a target-oriented real-time measurement of false positive is impossible. However, in the Internet, routing algorithms tends to choose the shortest paths between the sources and the destinations as the routing paths. By gathering all the shortest paths leading from every end hosts to the victim, we can obtain the percentage of the paths blocked by BS . We denote this percentage as $PCPV$ (Percentage of Convergence of Paths to Victim), and use this value to evaluate the influence of false positive introduced by BS . For example, in Fig. 4, all of the ten shortest paths are depicted with dotted lines. If r_{11} and r_8 are chosen as the filtering routers, only four of the paths are blocked, with the remaining six still available for communication. That is to say, despite the influence of false positive, the victim still has the ability to provide services to a certain percentage of end hosts. That is where $PCPV$ makes sense. We use Algorithm 5 to calculate the $PCPV$ value. $SP(a, v)$ refers to the set of nodes on the shortest path leading from a to v . H refers to the set of end hosts on the attack graph. NSP refers to the number of shortest paths leading to the victim.

$$LPP(BS) = \frac{\sum_{n_i \in P(a_i, v) \cap BS} Dist(n_i, v, P(a_i, v))}{D} \quad (4)$$

$$PCPV(BS) = \frac{|\{n_i \in H \mid SP(n_i, v) \cap BS \neq \emptyset\}|}{NSP} \quad (5)$$

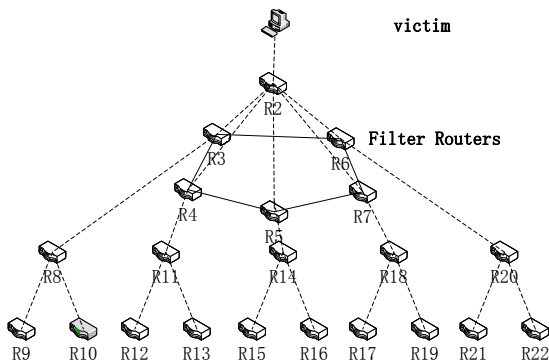


Figure 4. Shortest paths leading from the victim

To summarize, we evaluate the filtering performance with two metrics: $AFBP(BS)$ for victim protection and $LPP(BS)$ for link resource protection. $NF(BS)$ and $PCPV(BS)$ are metrics for evaluating filtering costs. We derive the ultimate mathematical formula for calculating

the filtering performance of BS in equation (6) and use $CP(BS)$ to denote the effectiveness of installing filters in BS . The formula allows for tradeoff between $AFBP$ and LPP , and the filtering cost metrics are contained by NF and $PCPV$. We will present an evaluation of (To summarize, we evaluate the filtering performance with two metrics: $AFBP(BS)$ for victim protection and $LPP(BS)$ for link resource protection. $NF(BS)$ and $PCPV(BS)$ are metrics for evaluating filtering costs. We derive the ultimate mathematical formula for calculating the filtering performance of BS in equation (6) and use $CP(BS)$ to denote the effectiveness of installing filters in BS . The formula allows for tradeoff between $AFBP$ and LPP , and the filtering cost metrics are contained by NF and $PCPV$. We will present an evaluation of (α, X, Y) in section V.

$$CP(BS) = \begin{cases} \alpha \times AFBP(BS) + (1-\alpha) \times LPP(BS) & NF(BS) < X, PCPV(BS) < Y \\ 0 & otherwise \end{cases} \quad (6)$$

3) BS Determination Algorithm

The ultimate goal is to determine the BS with the highest CP value. We have put forward the equation to evaluate BS , but still do not know how to determine BS . In this section, we will put forward a genetic-based algorithm to solve the problem.

Genetic algorithm introduction

A genetic algorithm is a heuristic search algorithm that mimics the process of natural evolution. The basic principles of the genetic algorithm were laid down by Holland [18], and have been proved useful in a variety of search and optimization problems. Genetic algorithm simulates the survival-of-the-fittest principle of nature. The principle provides an organizational reproductive framework: starting from an initial population, proceeding through some random selection, crossover and mutation operators from generation to generation, and converging to a group of best environment-adapting individuals.

The selection operator is applied to the current generation to pick up some individuals in probability P_s and copy them to an intermediate generation.

The crossover operator takes two individuals from the current generation, and recombines the two individuals with a probability P_c to create the individuals of the intermediate generation. The single point crossover method cuts the two individuals' encoding strings at a chosen position, and swaps the tail segments of the two strings.

The mutation operator chooses an individual from the current generation, and alters one bit in the individual's encoding string with a low probability P_m , then puts it in the intermediate generation.

Individuals of the intermediate generation will be evaluated with a fitness function that decides whether to put them in the next generation. After some rounds of iterations, the final generation is obtained, in which the fittest individuals will be kept. Finally, the algorithm will return the best individuals as the solution for the problem.

Algorithm definition

We find that in some special cases, the *BS* determination problem can be solved by efficient greedy-based algorithms. For example, we can evaluate every node in *BS* separately, and select a set of nodes with high *CP* values. But this kind of greedy-based method will probably end up with a local optimal solution because different nodes may affect each other's *CP* values. Genetic algorithm has strong global search capability. We can incorporate greedy methods into the initialization step of genetic algorithm. The algorithm will provide more accurate near-optimal solutions with higher speed.

Since every node in *BS* is selected from *FG*, we can encode each node in $FG(A, v)$ into a single bit, with 1 representing filters to be installed on this node and 0 representing filters not to be installed. In this way, each selection of *BS* can be encoded into a *N*-bit binary number, where *N* is the number of nodes in $FG(A, v)$. We use equation (6) to evaluate the level of fitness.

```

1: Encode the parameters and solution for the partitioning problem;
2: Initialize the first generation  $P_0$ ;
3: Calculate the fitness of each individual in  $P_0$ ;
4: Copy the individual with the highest fitness to the solution;
5: while ((the highest fitness of the current generation  $\leq$ 
fitness(solution) and TimeExpired = false) do
6: Select two individuals ( $g_1, g_2$ ) from the current generation;
7: Perform crossover on ( $g_1, g_2$ ) to produce a new individual  $g_i$  /*
start of crossover*/
8: if (fitness( $g_i$ )  $\leq$  max {fitness( $g_1$ ), fitness( $g_2$ )} then
9:  $\Delta C =$  fitness( $g_i$ ) - max {fitness( $g_1$ ), fitness( $g_2$ )}
10: if ( $-\Delta C / \min$ {fitness( $g_1$ ), fitness( $g_2$ )}  $\geq$  random(1,0)) then
11: Accept the crossover;
12: else
13: Reject the crossover;
14: end if
15: else
16: Accept the crossover;
17: end if /* end of crossover*/
18: Perform mutation on  $g_i$  to produce  $g_i'$ ; /* start of mutation*/
19: if (fitness( $g_i'$ )  $\leq$  fitness( $g_i$ )) then
20:  $\Delta C =$  (fitness( $g_i'$ ) - fitness( $g_i$ ));
21: if ( $-\Delta C / \text{fitness}(g_i) \geq$  random(1,0)) then
22: Accept the mutation;
23: else
24: Reject the mutation;
25: end if
26: else
27: Accept the mutation;
28: end if /* end of mutation*/
29: Calculate the fitness of each individual in current generation;
30: end while
31: return solution
    
```

Figure 5. Genetic algorithm for selecting filtering routers

The details of the algorithm is shown in Fig. 5. Steps 1 to 4 are initializations for parameters and the solution for the node set determination problem. Step 5 is used to check whether the termination condition of the iterations is met or not. The greedy-based crossover and mutation operations are performed in this iteration block to produce individuals of the next generation. Let us see the two operations in detail. Steps 7 to 17 are the crossover operations. The key idea is that when the crossover operation produces better individuals, the better individuals are accepted. Otherwise, we will accept newly generated individuals as candidates for the next generation. Steps 18 to 28 are mutation operations. Steps 29 to 32 update the solution and the generation number.

C. Design of SPF Module

In our scheme, the SPF module is running on filtering routers. Once the SPF module receives a filtering request from a victim, it installs a filter to block the attack flow.

Routers in the Internet have limited resources for filtering, while a router may receive many filtering requests from one or more victims. A filter should be weeded out once such attack flows tend to disappear after a certain period of time. The SPF module is used for such kind of filter management.

How does SPF determine whether a filter should be kept or weeded out? Two factors should be taken into consideration. Firstly, the installation time is an important factor. We should hold up for filters that are installed not long ago. Secondly, the frequency of a filter being used is also a critical factor. We should give an integrated measurement of both factors. Such measurement can be scored using equation (7).

$$SC(I) = T \cdot (tc - td) + F \cdot m \tag{7}$$

in which *T* denotes the weight of the filter installation time and *F* is the weight of the frequency by which filter *I* is used. *tc* denotes the current time and *td* denotes the filter installation time. *m* denotes how many times the filter is used.

The scores of filters should be computed and logged periodically. Filters whose scores are below a threshold should be weeded out.

V. PERFORMANCE EVALUATION

Simulations have been conducted on three network topologies to evaluate the effectiveness of the proposed scheme. The topology data is from CAIDA [19] and all the attack paths are routes starting from a single victim to multiple attackers. For performance evaluation purpose, we assume that each legitimate user sends packets at the rate of 1 unit per second and all the attackers attacks at the same rate.

TABLE I. PERFORMANCE METRICS AND CONTROL PARAMETERS

Performance Metrics	BDP(Bad Drop Percentage): the percentage of the DDoS traffic dropped
	LPP(Link Protected Percentage): the reduction of damage towards link resource
Control Parameters	<i>g</i> : the percentage of attackers among all the end hosts
	α : the impact factor α is set to 1 to protect the victim and 0 to protect the link resource
	<i>X</i> : <i>X</i> is determined by the victim's tolerance of false positive
	<i>Y</i> : <i>Y</i> represents the max number of filtering routers that a victim can use.

A. Metrics to Evaluate

Table 1 shows the performance metrics and control parameters used in our simulation. The bad traffic drop percentage (BDP) is the dropping rate of the attack traffic. The value of BDP can be obtained by calculating traffic statistics at the victim. Link Protected Percentage (LPP), which we have put forward in section VI, is used to evaluate the protection for link resources. Among the control parameters, *g* represents the percentage of

illegitimate incoming traffic, α , X and Y are impact factors in equation (6).

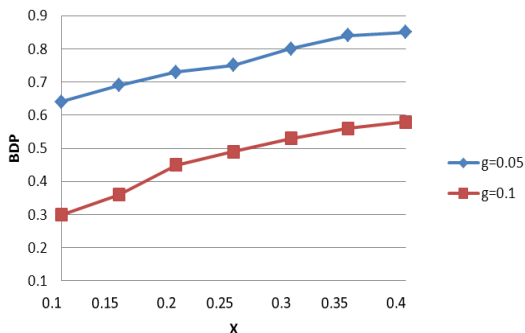


Figure 6. The influence of X on BDP

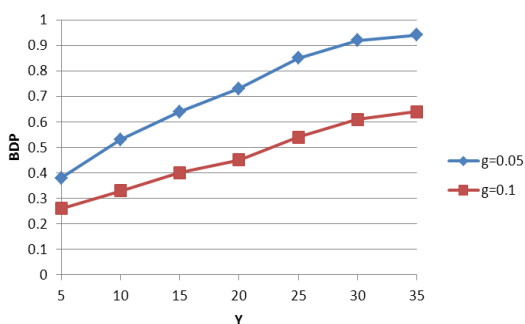


Figure 7. The influence of Y on BDP

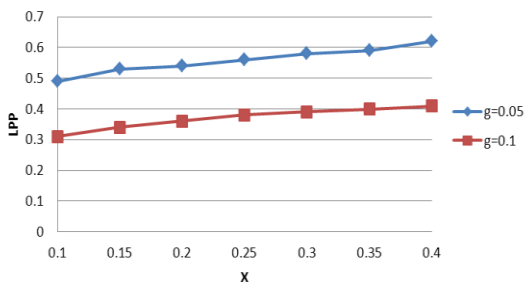


Figure 8. The influence of X on LPP

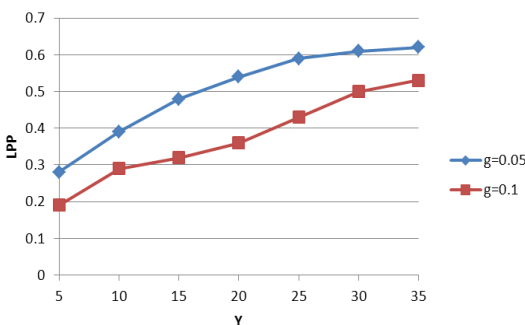


Figure 9. The Influence of Y on LPP

B. Simulation Results

Fig. 6 shows the BDP as a function of X. α is set to 1 to maximize the protection for the victim resource. Y is set to 20. There are two curves in the figure,

corresponding to two different values of g ($g = 0.05$ and $g = 0.1$). In other words, the number of attackers will be 5% and 10% of the total end hosts. The comparison of the two curves indicates that the scheme’s performance is related to the percentage of attackers. The scheme will work better under the scenario where the distribution of attacking sources is not very wide. Recall that X represents the impact factor determined by the victim’s tolerance of false positive. Both curves indicate that there will be an improvement on BDP when X increases. That is to say, there is a tradeoff between protection for victims and the severity of false positive. Fig. 7 shows how the BDP changes as the parameter Y varies. Y represents the max number of filtering routers a victim can use. The curves indicate that BDP increases when more routers are used for filtering. Fig. 8 and Fig. 9 reflect protection for link resources, and we can see that the curves have a similar changing trend.

VI. DISCUSSIONS

A. Mapping Upstream Routers

In previous sections, we declared that our scheme rely on the assumption that the victim has a map of the upstream routers. In this subsection, we will demonstrate that this assumption is reasonable and practical.

Many tools can be used for mapping. CAIDA’s skitter is the most persevering measurement project, which can map routers from the victim to a large set of selected destinations. Rocketfuel [21] is another available choice. It is easy to get a map of upstream routers using these tools. We can also use itrace [22] in our application. By collecting itrace packets we can construct the upstream map.

It is difficult to obtain an updated accurate Internet topology. However, since our goal is not to get the exact attackers, such a map does not have to be perfect. Moreover, the attack graph itself is a good tool for mapping upstream routers.

B. Potential Applications of the Proposed Scheme

A significant hurdle for DDoS defense may be the lack of viable economic incentives. Installing filters in a domain consumes valuable router resources and reduces the overall routing performance. Moreover, its beneficiaries are likely to be other domains rather than the domain performing filtering. With our scheme, the economic model can be built as follows: victims can be charged for the filtering services provided by an ISP. The payment may depend on the number of filtering routers being used and the number of attack packets to be filtered.

Differentiated qualities in services can also be provided by our scheme, as the impact factors in equation (6) can be adjusted according to the payment. For example, if the victim has paid for a higher service quality, more filters can be used to improve the filtering effect.

VII. CONCLUSION

This paper presents an IP-traceback-based packet filtering scheme against DDoS attack. It utilizes the

attack graph obtained through IP traceback to evaluate the cost and the performance of filtering routers and deploy filters accordingly on the proper routers. The simulation results demonstrate that the scheme is very effective in protecting the victim's resources as well as the link resources, yet keeping filtering resource consumption and loss of normal traffic within a proper limit.

REFERENCES

- [1] Michael Kan, "Major DDoS attacks. cn domain; disrupts Internet in China," [Online]. Available: http://www.computerworld.com/s/article/9241899/Major_DDoS_attacks_cn_domain_disrupts_Internet_in_China. [Accessed: 10 Aug. 2013]
- [2] "Government websites hit by new DDoS attack," [Online]. Available: http://www.koreatimes.co.kr/www/news/nation/2011/03/117_82505.html. [Accessed: 15 May 2013]
- [3] T. Peng, C. Leckie and K. Ramamohanarao. "Survey of network-based defense mechanisms countering the DoS and DDoS problems", *ACM Comput. Surv.* vol. 39, no. 1, 2007, doi: 10.1145/1216370.1216373
- [4] J Lee, G de Veciana. "Scalable network-layer defense against internet bandwidth flooding attacks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 4, pp. 1284–1297, 2009. doi: 10.1109/TNET.2008.2007431
- [5] J Lee, G de Veciana. "Scalable multicast based filtering and tracing framework for defeating distributed DoS attack", *International Journal Of Network Management*, vol. 14, 2004.
- [6] K. J. Argyraki and D. R. Cheriton, "Active internet traffic filtering: Realtime response to denial-of-service attacks," in *USENIX Annual Technical Conference, General Track*, pp. 135–148, 2005.
- [7] Heejo Lee, "On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets" *Computer Communication Review*, vol. 31, no. 4, pp. 15–26, 2001. doi: 10.1145/964723.383061
- [8] K. Park and H. Lee, "On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law internets," in *Proc. of SIGCOMM 2001*, pp. 15–26, 2001. doi: 10.1145/964723.383061
- [9] Jun Li, Jelena Mirkovic and Mengqiu Wang, SAVE: Source Address Validity Enforcement Protocol, in *Proc. of INFOCOM 2002*, pp. 1557 - 1566, 2002. doi: 10.1016/j.comnet.2007.09.024
- [10] H. Lee, M. Kwon, G. Hasker and A. Perrig, "BASE: an incrementally deployable mechanism for viable IP spoofing prevention," in *Proc. of ASIACCS 2007*, pp. 20–31, 2007.
- [11] A. Yaar, A. Perrig and D. Song, "SIFF: A stateless internet flow filter to mitigate DDoS flooding attacks," in *IEEE Symposium on Security and Privacy*, pp. 130–143, 2004.
- [12] YAU, LUI AND LIANG. "Defending against distributed denial-of-service attacks with max-min fair server-centric router throttles". In *Proc. of the IEEE International Workshop on Quality of Service (IWQoS)*, pp. 35–44, 2002. doi: 10.1109/TNET.2004.842221
- [13] S. Savage, D. Wetherall, A. Karlin, and T. Anderson, "Practical Network Support for IP Traceback," in *Proc. of ACM SIGCOMM*, pp. 295–396, 2000. doi: 10.1145/347057.347560
- [14] A. Yaar, A. Perrig, and D. X. Song, "Pi: A path identification mechanism to defend against DDoS attack," in *IEEE Symposium on Security and Privacy*, pp. 93–109, 2003
- [15] Minh Sung, Jun Xu, "IP Traceback-based Intelligent Packet Filtering: A Novel Technique for Defending Against Internet DDoS Attacks" *IEEE Transactions on Parallel and Distributed Systems*, vol. 14, no. 9, pp. 861–872, 2003.
- [16] D. Seo, H. Lee, A. Perrig. "PFS: Probabilistic Filter Scheduling Against Distributed Denial-of-Service Attacks", in *Proc. of the 2011 IEEE 36th Conference on Local Computer Networks*, pp. 9-17, 2011.
- [17] D. X. Song, A. Perrig. Advanced and authenticated marking schemes for IP traceback, in *Proc. of INFOCOM 2001*, vol. 2, pp. 878-886, 2001
- [18] Genetic Algorithm. [Online]. Available: http://en.wikipedia.org/wiki/Genetic_algorithms. [Accessed: 6 Jun 2013]
- [19] "Cooperative association for internet data analysis," [Online]. Available: <http://www.caida.org>[Accessed: 15 Jun 2013]
- [20] X. Liu, X. Yang, and Y. Lu, "To filter or to authorize: network-layer DoS defense against multimillion-node botnets," in *Proc. of SIGCOMM 2008*, pp. 195–206, 2008
- [21] Rocketfuel. [Online]. Available: <http://www.cs.washington.edu/research/networking/rocketfuel/>. [Accessed: 2 Jul 2013]
- [22] ICMP Traceback (itrace), [Online] Available: <http://datatracker.ietf.org/wg/itrace/charter/> [Accessed: 6 Aug 2013]

Yulong Wang received his Ph.D. degree in computer science from the Beijing University of Posts and Telecommunications, China, in 2010. He is currently a lecturer at the Beijing University of Posts and Telecommunications. His research interests include network security and cloud computing.

Rui Sun is currently a MS candidate at the Beijing University of Posts and Telecommunications, China. His research interests include computer security and network.

Multi-Object Optimization Based RV Selection Algorithm for VCN

Rong Chai, Bin Yang, Li Cai, Xizhe Yang, and Qianbin Chen

Key Lab of Mobile Communication Technology, Chongqing University of Posts and Telecommunications, Chongqing, P. R. China

Email: chairong@cqupt.edu.cn, 876914625@qq.com

Abstract—Vehicular communication network (VCN) has recently received considerable attention both from academia and industry. In VCN, vehicles are expected to be capable of communicating with other vehicles as well as stationary infrastructures, i.e., the access points (APs) of wireless access networks. In the case that the direct connection between a source vehicle (SV) and APs is inaccessible, relay vehicles (RVs) can be applied for supporting multi-hop connection between SVs and APs. In this paper, a multi-object based RV selection algorithm for VCN is proposed, which jointly considers the characteristics of physical channel, link status between SVs and RVs, the bandwidth and delay characteristics of RVs and user service requirements. The utility functions of both SVs and RVs are modeled and a multi-object optimization problem is formulated. Applying ideal point method, the problem can be solved and the optimal SV-RV pairs can be obtained. Simulation results demonstrate that compared to previous algorithms, the proposed algorithm offers better performance in terms of user throughput, successful transmission rate and average transmission delay.

Index Terms—Multi-Object; Relay Selection; Utility Function; VCN

I. INTRODUCTION

Vehicular communication network (VCN) has recently received considerable attention both from academia and industry [1-3]. In VCN, three vehicular communication modes are expected to be supported, i.e., vehicle to vehicle (V2V) communication, vehicle to infrastructure (V2I) communication, and hybrid vehicular (HV) communication. Figure 1 shows an example of VCN model.

For both V2I and HV communications, the direct transmission link between a vehicle, referred to as source vehicle (SV) hereafter, and the access points (APs) of wireless access networks might become inaccessible due to the reasons such as the mobility of vehicles, the limited coverage area of APs and the fading characteristics of physical channels, resulting in the unavailability of communication services. To solve this problem, relay vehicles (RVs) which are capable of forwarding data packets between APs and SVs, can be applied. In the case that multiple candidate RVs are available, the problem of optimal RV selection has to be considered.

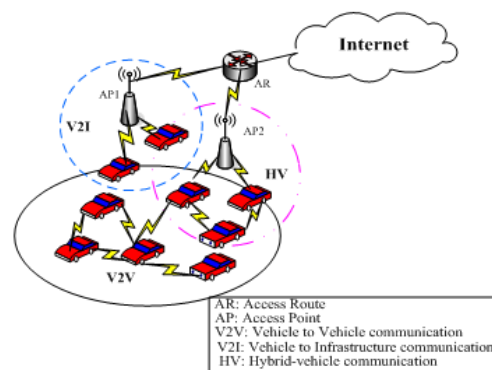


Figure 1. Vehicular communication network mode

In practical VCN scenario, a group of users, for instance, passengers in a bus, may seek to access RVs at almost the same time when the bus moves out of the coverage area of all adjacent APs. In this case, the problem that multiple SVs select RVs arises.

The problem of RV selection for VCN has been studied in previous literatures. While the signal strength of physical channel and the position and velocity information of vehicles are considered in previous works, a number of factors affecting transmission performance of SVs and RVs, such as channel propagation model, channel accessing scheme, available bandwidth of RVs, etc., have failed to be considered extensively. Furthermore, previous works only address the problem of RV selection for one SV, the extension to multiple SVs is still an open issue.

In this paper, we propose a RV selection method for multiple SVs in V2I and HV communication mode, which jointly considers the characteristics of physical link, channel accessing scheme, and the available bandwidth of RVs. The utility functions are modeled for both SVs and RVs and the optimal RV selection scheme which achieves the joint utility optimization of SVs and RVs is presented.

The rest of paper is organized as follows. The related works are discussed in Section II and the system model is described in Section III. The proposed scheme is presented in Section IV. The ideal point method is applied to solve the formulated optimization problem in Section V. Numerical results are given in Section VI. Finally, we conclude the paper in Section VII.

II. RELATED WORKS

In recent years, various relay selection algorithms have been proposed for VCN [4-12]. These algorithms can be classified into three categories, i.e., single attribute based RV selection algorithm, multi-attribute based RV selection algorithm and opportunistic RV selection algorithm.

A. Single Attribute Based RV Selection Algorithms

Single attribute based RV selection algorithms choose one metric as the only selection criterion, and select the candidate relay with the metric being the largest/smallest. For instance, the metric of link quality is chosen as the metric for RV selection in [4] and [5]. More specifically, instantaneous link quality (ILQ) is stressed and a relay selection method which selects the candidate RV with the best ILQ is proposed in [4]. To avoid the drawback of excessive signaling overhead resulted from link quality measurement, [5] proposes an average link quality (ALQ) based relay selection algorithm, which offers the highest throughput at the destination.

B. Multi-Attribute Based RV Selection Algorithm

While the effect of link quality is taken into account in relay selection, many other factors which may also affect the performance of relay selection failed to be considered in [4] and [5]. Multi-attribute based RV selection algorithms jointly consider the factors affecting data forwarding from the relays in designing relay selection criterions.

In [6], the authors propose a cross-layer RV selection algorithm for VCNs. A cost function is defined based on multiple factors, such as the geographic location and velocity of candidate relays, and physical layer channel conditions, the candidate relay with the largest cost is chosen as the target relay. A cross-layer mobile relay selection scheme is proposed in [7] which jointly considers the factors including the status of the links, the bandwidth and delay features of the candidate relays, and the quality of service (QoS) requirements of users. A simple additive weighting (SAW) method is applied to evaluate the performance of the candidate relays and the one with the best weighted value is chosen as the target relay.

To stress the dead spot and out of coverage problem in VCN, a multi-hop relay selection scheme is proposed based on three metrics, i.e., the received signal strength (RSS) from UMTS, route life time (RLT) and available relay capacity to extend the coverage area and decrease the number of handoff in [8]. In [9], physical layer channel condition characterized by signal-to-noise ratio (SNR), geographical locations and velocities of vehicles are jointly taken into account for optimal RV selection.

C. Opportunistic RV Selection Algorithm

Unlike single attribute and multi-attribute based RV selection algorithms, in which a deterministic relay is selected, opportunistic RV selection algorithm assigns various priorities to candidate relays according to certain selection criterion, and the candidate relay with high priority has high probability of being selected as the

target relay. A robust distance-based relay selection scheme for multi-hop broadcast of emergency notification messages is proposed in [10] and the candidate node with longer directional distance to the sender is of higher probability for being selected as the target relay. In [11], the authors propose a novel opportunistic relay protocol for VCN which exploits multiuser diversity and effectively copes with the dynamic fading channel. The relay candidate that has the highest average channel quality is given the highest priority of accessing the channel and becoming the target relay. The authors in [12] propose an ExOR scheme in which the source node sends a possible relay list in the header, and the potential relays having received the message send back an ACK message, based on which the source node can then select the best relay.

While some relay selection algorithms have been proposed for VCN, the factors of affecting the performance of RV selection and data forwarding have not been considered extensively. Furthermore, most research works consider selecting one RV for one SV, however, in some particular application scenarios, multiple RVs may need to be selected for multiple SVs. In this paper, a RV selection method for multiple SVs is proposed for VCN, the utility functions of both SVs and RVs are defined based on the factors including the characteristics of physical link, channel accessing scheme, and available bandwidth of RVs, etc., and the multi-objective optimization problem is solved to obtain the optimal RV selection strategy.

III. SYSTEM MODEL

In this section, the network scenario and channel model considered in this paper are described.

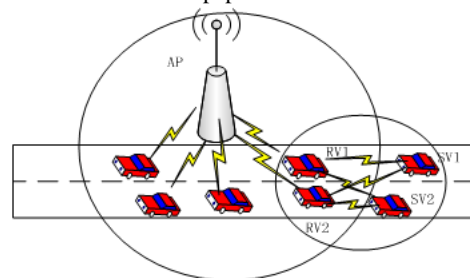


Figure 2. Network model

A. Network Scenario

In this paper, we consider a road segment of which one AP is deployed. Vehicles within the coverage area of the AP are allowed to communicate with the AP directly. Those SVs which are out of the coverage area of the AP have to seek the help from adjacent RVs for data forwarding between the SVs and the AP. In this paper, we focus on the scenario that multiple SVs need to select multiple RVs for data forwarding. Considering the infeasibility and low transmission efficiency that may be resulted from one RV forwarding data for multiple SVs, we assume that each RV can only accept the relay request from one SV and each SV can only choose one RV for data forwarding.

Figure 2 shows an example of network model considered in this paper, which consists of one AP, multiple RVs and multiple SVs. As SV1 and SV2 are out of the coverage area of the AP, they need to choose RV1 or RV2 for accessing the AP.

B. Channel Model

It has been shown that Nakagami fading channel can be applied to model the communication channel between SVs and RVs, and that between RVs and APs [13, 14]. Specifically, the communication channel between a RV and an AP can be modeled as Nakagami fading channel with the channel characteristics h_1 following the probability distribution function (pdf):

$$f(h_1) = \frac{2m^m}{\Omega(d)^m \Gamma(m)} h_1^{2m-1} \exp\left(-\frac{m}{\Omega(d)} h_1^2\right) \quad (1)$$

where m represents the Nakagami fading parameter ($m \geq 1/2$), $\Gamma(\cdot)$ denotes the Gamma function [15], d denotes the transmission distance, $\Omega(d)$ denotes the power loss due to transmission distance d , and can be expressed as:

$$\Omega(d) = \frac{P_t G_t G_r h_t^2 h_r^2}{L} d^{-\theta} \quad (2)$$

where, P_t is the transmission power, G_t and G_r are the antenna gains of the transmitter and receiver, respectively, h_t and h_r are the antenna heights of the transmitter and receiver, respectively, θ is the path loss exponent, and L is the system loss.

The channel between a SV and a RV can be modeled as a cascaded Nakagami fading channel with the number of cascade being 2 and the channel pdf as follows:

$$f(h_2) = \frac{2}{h_2 \Gamma(m_1) \Gamma(m_2)} G_{0,2}^{2,0} \left[\frac{m_1 m_2 h_2^2}{\Omega_1 \Omega_2} \middle| \begin{matrix} m_1, m_2 \end{matrix} \right] \quad (3)$$

where $G_{0,2}^{2,0}(\cdot|\cdot)$ denotes Meijer G-function, $\Omega_l = E[h_l^2]$ and $m_l = \Omega_l^2 / E(h_l - \Omega_l)^2 \geq 1/2$, $l=1,2$.

IV. PROPOSED RV SELECTION ALGORITHM

A. Basic Idea

In this paper, we consider the application scenario that multiple SVs of which the direct links to APs are inaccessible need to select optimal RVs. For these SVs, data forwarding through RVs provides an efficient manner for performing communication, thus certain data transmission revenue can be obtained. On the other hand, to receive forwarding services from RVs, SVs need to pay a certain amount of service fees to the corresponding RVs. To choose different RVs for data forwarding, one SV may receive different revenue and undertake different forwarding costs as well. Therefore, each SV tends to choose the optimal RV from which relatively high revenue can be obtained while low cost is required.

The candidate RVs, which are qualified to forward data for SVs, may choose to accept or reject the

forwarding requests from the SVs. On one hand, through offering data forwarding for the SVs, the RVs are capable of receiving a certain amount of data forwarding fees, on the other hand, the RVs have to undertake certain data forwarding costs, for instance, some bandwidth and transmission power have to be consumed. As choosing different SVs for offering data relaying, different revenue and costs might be resulted, the RVs tend to select the optimal SV from which the high forwarding revenue can be obtained while a little cost is required.

As data forwarding revenues and costs are both complicated quantities associated with multiple factors, the exact mathematical modeling is prohibited. In this paper, the concept of utility function [16] is introduced to characterize the revenue and costs of RV selection for both SVs and RVs, and the problem of utility optimization for both SVs and RVs is formulated.

B. Utility Function Modeling of SVs

Assuming the i th SV selects the j th RV for data forwarding, the utility function of the SV can be modeled as follows:

$$U_{ij}^S = W_{ij}^S - C_{ij}^S, 1 \leq i \leq M, 1 \leq j \leq N \quad (4)$$

where W_{ij}^S denotes the revenue the i th SV receives through applying the forwarding scheme from the j th RV. It can be expected that the better data forwarding performance, the higher revenue the SVs can receive. Hence, W_{ij}^S can be formulated as a function of the transmission rate and the availability of the link between the i th SV and the j th RV, the access probability to the j th RV and the available bandwidth of the RV, i.e.,

$$W_{ij}^S = \begin{cases} F_{ij} \alpha_i R_{ij} (1 - P_{ij}^C), & RET_{ij} \geq T_{ij} \\ F_{ij} \alpha_i R_{ij} (1 - P_{ij}^C) \frac{RET_{ij}}{T_{ij}}, & RET_{ij} < T_{ij} \end{cases} \quad (5)$$

where F_{ij} denotes the available bandwidth index of the j th RV, and can be expressed as:

$$F_{ij} = \begin{cases} 1, & B_j^a \geq B_i^r \\ 0, & B_j^a \leq B_i^r \end{cases} \quad (6)$$

where B_j^a and B_i^r denote the available bandwidth of the j th RV and the required bandwidth of the i th SV, respectively. The available bandwidth of the j th RV can be calculated as [17]:

$$B_j^a = \frac{k \times S_B \times 8}{T} \quad (7)$$

where, k denotes the number of packets sent and received by the vehicle within time period T , S_B denotes the size of packets in byte.

α_i in (5) denotes the unit rate revenue factor of the i th SV, R_{ij} represents the data rate of the link between the i th SV and the j th RV and can be expressed as:

$$R_{ij} = B_j^a \log(1 + \psi_{ij}) \quad (8)$$

where ψ_{ij} denotes the average SNR of the link, which can be expressed as:

$$\psi_{ij} = \frac{E_s}{N_0} E[h_{ij}^2] \quad (9)$$

where h_{ij} denotes the channel gain from the i th SV to the j th RV, E_s denotes the average energy of the transmitted symbol and N_0 is the single-sided power spectral density of additive white Gaussian noise (AWGN).

P_j^c in (5) denotes the probability of collision when multiple SVs/RVs choose to access the j th RV simultaneously, and can be expressed as [18]:

$$P_j^c = \tau[1 - (1 - \tau)^{m_j - 1}] \quad (10)$$

where m_j is the number of adjacent vehicles of the j th RV, τ is the message transmission probability of each vehicle in the considered slot time, and can be calculated by jointly solving the following nonlinear equations:

$$\tau = \frac{2(1 - 2p)}{(1 - 2p)(CW_{\min} + 1) + pCW_{\min}(1 - (2p)^m)} \quad (11)$$

and

$$p = 1 - (1 - \tau)^{m_j - 1} \quad (12)$$

where m is the maximum number of retransmissions and CW_{\min} denotes the minimum contention window.

RET_{ij} in (5) denotes the link duration between the i th SV and the j th RV, and can be calculated as follows. Assuming that the coordinates of the i th SV and the j th RV are (x_i^S, y_i^S, z_i^S) , (x_j^R, y_j^R, z_j^R) , respectively, r_j denotes the maximum transmission range of the j th vehicle, d_{ij} denotes the distance between the i th SV and the j th RV, v_i^S, v_j^R denote the velocity of the i th SV and the j th RV, θ_i^S denotes the angle between the connection line of the i th SV and the j th RV, and the moving direction of the i th SV, θ_j^R denotes the angle between the connection line of the j th RV and the i th SV, and the moving direction of the j th RV. It can be proved that RET_{ij} can be expressed as:

$$RET_{ij} = \frac{-(ab + ce) + \sqrt{(a^2 + c^2)r_j^2 - (ae - bc)^2}}{a^2 + c^2} \quad (13)$$

where $a = v_j^R \cos\theta_j^R - v_i^S \cos\theta_i^S$, $b = y_j^R - y_i^S$, $e = x_j^R - x_i^S$, $c = v_j^R \sin\theta_j^R - v_i^S \sin\theta_i^S$.

T_{ij} in (5) denotes the required transmission time from the i th SV to the AP via the j th RV and can be expressed as:

$$T_{ij} = 2T_{OH} + \frac{S_{PL} + S_{MAC}}{R_j} + \frac{S_{PL} + S_{MAC}}{R_{jA}} + T_{SIFS} + T_{BO} \quad (14)$$

where $T_{OH} = T_{DIFS} + 3T_{SIFS} + T_{RTS} + T_{CTS} + T_{ACK} + 2T_{PLCP}$, T_{DIFS} and T_{SIFS} are respectively, the duration of distributed inter-frame space and short inter-frame space defined in IEEE 802.11 [19], T_{PLCP} denotes the duration of physical layer convergence procedure, T_{RTS} denotes the duration of an RTS frame, T_{CTS} denotes the duration of a

CTS frame, and T_{ACK} denotes the duration of an acknowledgement frame, S_{PL} and S_{MAC} denote the size of an MAC payload and an MAC header, respectively, R_{jA} represent the data transmission rate of the link between the j th RV and the AP, T_{BO} denotes the average backoff time.

In (4), C_{ij}^S denotes the cost the i th SV has to undertake for seeking for data forwarding from the j th RV, and can be modeled as the service fee that the SV has to afford. In this paper, we assume that the service fee of the SV is charged based on the bandwidth resource it spends, and can be calculated as the product of the amount of user bandwidth and unit service fee, i.e.,

$$C_{ij}^S = \beta_j F_{ij} B_i^r \quad (15)$$

where β_j denotes the unit bandwidth price factor for data forwarding of the j th RV.

C. Utility Function Modeling of RVs

Assuming the j th RV offers data forwarding service for the i th SV, the utility function of the j th RV can be expressed as follows:

$$U_{ij}^R = W_{ij}^R - C_{ij}^R, 1 \leq i \leq M, 1 \leq j \leq N \quad (16)$$

where W_{ij}^R represents the revenue received by the j th RV through offering data forwarding service to the i th SV. Assuming that the revenue a RV received is proportional to the bandwidth resource it offers, furthermore, for higher probability of successful transmission, better data forwarding performance can be obtained, thus higher revenue can be received. Hence, W_{ij}^R can be modeled as:

$$W_{R_{i,j}} = F_{ij} P_{ij}^S \beta_j B_i^r \quad (17)$$

In (16), C_{ij}^R denotes the data forwarding cost the j th RV undertakes when forwarding data for the i th SV. In this paper, we apply Sigmoid function which was originally introduced in Machine Learning theory [16] for quantifying the cost the RV undertakes. C_{ij}^R can be modeled as:

$$C_{ij}^R = \frac{c_j F_{ij}}{1 + e^{\theta_j(\varphi_j - P_{jAP}^S P_{ij}^S B_i^r)}} \quad (18)$$

where θ_j , φ_j are both parameters characterizing the steepness and the inflection point of the cost curve of the j th RV, c_j denotes the unit cost factor of the j th RV, P_{ij}^S denotes the successful transmission probability of the link between the i th SV and the j th RV, similarly, P_{jAP}^S denotes the successful transmission probability of the link between the j th RV and the AP. The successful transmission probability can be evaluated as the probability that the SNR at the receiving node is larger than a given threshold. According to the channel model defined in Section III, P_{ij}^S and P_{jAP}^S can be calculated respectively as follows:

$$P_{jAP}^S = \Pr\left(\frac{E_s}{N_0} h_{jAP}^2 > \psi_{th1}\right) = 1 - \frac{\gamma(m, \frac{mN_0}{\Omega E_s} \psi_{th1})}{\Gamma(m)} \quad (19)$$

$$P_{ij}^S = \Pr\left(\frac{E_s}{N_0} h_{ij}^2 > \psi_{th2}\right) = 1 - \frac{G_{1,3}^{2,1}\left[\frac{m_1 m_2 N_0 \psi_{th2}}{\Omega_1 \Omega_2 E_s} \Big|_{m_1, m_2, 0}\right]}{\Gamma(m_1) \Gamma(m_2)} \quad (20)$$

where h_{jAP} denote the channel gain from the j th RV to the AP, ψ_{th1} and ψ_{th2} denote the given SNR thresholds and γ represents incomplete Gamma function.

D. Optimization Problem Modeling

As it is assumed that one SV can only select one RV for forwarding data, the SV evaluates the utility function resulted from choosing various RVs and tends to select the one corresponding to the maximal utility. Similarly, assuming each candidate RV can only forward data for one SV or choose no SV for data forwarding, the RV evaluates the utility function resulted from choosing various SVs for data forwarding and tends to select the SV leading to the maximal utility. Therefore, the optimal RV selection problem can be modeled as a multi-object optimization problem:

$$\begin{aligned} \max \quad & U_i^S = \sum_{j=1}^N x_{i,j} U_{ij}^S, \quad i=1, 2, \dots, M \\ \max \quad & U_j^R = \sum_{i=1}^M x_{i,j} U_{ij}^R, \quad j=1, 2, \dots, N \\ \text{s.t.} \quad & x_{i,j} \in \{0, 1\}, \sum_{i=1}^M x_{i,j} \leq 1, \sum_{j=1}^N x_{i,j} \leq 1. \end{aligned} \quad (21)$$

where M and N denote, respectively the number of SVs and RVs.

V. IDEAL POINT METHOD BASED OPTIMIZATION SOLUTION

The problem formulated in (21) is a multi-object optimization problem, the optimal solution of which is in general difficult to obtain. In this section, the ideal point method [20] is applied to solve the optimization problem.

The basic idea of ideal point method is that for each objective function, the optimal solution, referred to as ideal solutions can be obtained independently without considering the joint constraints and the feasibility of the solutions, then the joint solutions subject to given constraints can be calculated by minimizing the distance between the feasible solutions and the ideal solutions.

Collection all the objective functions formulated in (21), we define:

$$f_i(X) = \sum_{j=1}^N x_{i,j} U_{ij}^S, \quad i=1, 2, \dots, M \quad (22)$$

$$f_{M+j}(X) = \sum_{i=1}^M x_{i,j} U_{ij}^R, \quad j=1, 2, \dots, N \quad (23)$$

Denoting the optimal solution to each objective problem by x_l^* , $l=1, 2, \dots, M+N$, and assuming x_l^* exists in the range of D :

$$D = \left\{ x_{i,j} \in \{0, 1\}, \sum_{i=1}^M x_{i,j} \leq 1, \sum_{j=1}^N x_{i,j} \leq 1, 1 \leq i \leq M, 1 \leq j \leq N \right\} \quad (24)$$

The corresponding objective functions can be expressed as:

$$f_l^* = f_l(x_l^*) = \max_{x \in D} f_l(x), l=1, 2, \dots, M+N \quad (25)$$

Rewriting (22) and (23) in vector form, we obtain:

$$F(x) = (f_1(x), f_2(x), \dots, f_{M+N}(x))^T \quad (26)$$

Define $F^*(x) = (f_1^*(x), f_2^*(x), \dots, f_{M+N}^*(x))^T$ as the ideal point in the space R^{M+N} of the vector objective function $F(x)$, then

(1) If $x_1^* = x_2^* = \dots = x_{M+N}^*$, that is, $f_l(x_l^*) \geq f_l(x)$, for $l=1, 2, \dots, M+N$, then x_l^* is the optimal solution of the optimization problem.

(2) In the case that $x_1^*, x_2^*, \dots, x_{M+N}^*$ are not equal, namely, the ideal point is infeasible in the set D , the optimal solution can be obtained by evaluating the distance between the ideal solutions and the feasible solutions, and choosing the solution corresponding to the minimal distance. To examine the distance between the solutions, we define the p th order norm in the objective space R^{M+N} as:

$$u(F) = \|F - F^*\|_p = \left[\sum_{l=1}^{M+N} (f_l - f_l^*)^p \right]^{\frac{1}{p}}, 1 \leq p \leq \infty \quad (27)$$

The original multi-objective optimization problem can be converted into a single object optimization problem:

$$\min_{x \in D} u(F) = \min_{x \in D} \left[\sum_{l=1}^{M+N} (f_l(x) - f_l^*)^p \right]^{\frac{1}{p}} \quad (28)$$

which can then be solved to obtain the optimal solution of the original optimization problem.

VI. NUMERICAL RESULTS

In this section, the performance of the proposed RV selection algorithm is examined and compared with previous algorithms, including the algorithms proposed in [8] and [9]. In the simulation, we consider a straight road of 1Km with one AP being deployed in the middle of the road. A number of vehicles are randomly located and move along the road with the velocity randomly chosen from 80Km/h to 120Km/h. The number of total vehicles varies from 20 to 60, with the numbers of SVs and RVs being equal. The detailed parameters used in the simulation are summarized in Table I.

Figure 3 shows the throughput of all the vehicles versus the number of vehicles. It can be seen that for all the three algorithms, the total throughput increases with the increase of the number of vehicles. This is because the connectivity of the network improves accordingly. Comparing the total throughput resulted from three algorithms, it can be seen that the proposed scheme outperforms the algorithms proposed in [8] and [9]. The reason is that the proposed algorithm takes into account

both available bandwidth of RVs and the link quality between SVs and RVs in selecting the optimal RV.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Simulation area	1KM
Number of AP	1
Position of AP	500m
Height of AP	20m
Transmission range of AP	300m
Number of vehicles	20-60
Transmission range of vehicles	200m
Velocities of vehicles	80-120Km/h
Packet length	1000bits
Size of control messages	60 units

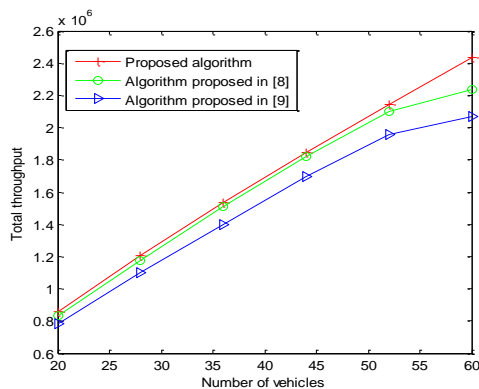


Figure 3. Overall throughput

Figure 4 plots the average transmission delay of various algorithms versus the number of vehicles. It can be seen from the figure that for all the three algorithms, as the number of vehicles increases, the average transmission delay increases, this is mainly because the collision probability increases resulting in large channel accessing time. Comparing the average transmission delay resulted from three algorithms, it can be seen that the proposed algorithm offers much lower transmission delay compared to the other two algorithms. The reason is that the factors affecting transmission delay, i.e., packet collision and data transmission rate are both considered in the proposed algorithm.

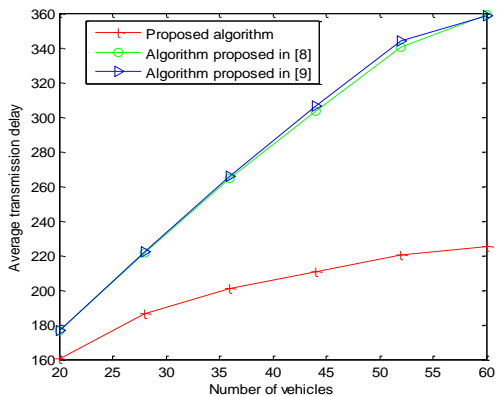


Figure 4. Average transmission delay

Figure 5 shows the successful transmission rate versus the numbers of vehicles obtained from various algorithms.

It can be seen from the figure that the successful transmission probability increases first and then decreases. This is mainly because when the number of vehicles is small, the connectivity of the vehicles may be limited, thus resulting in low probability of successful transmission. When the number of vehicles increases, the connectivity of the vehicles improves, resulting in larger probability of successful transmission. However, on the other hand, the probability of collision increases with the increase of the number of vehicles. As the effects of data collision dominates the performance of data transmission for a large number of vehicles, thus resulting in the decrease of successful transmission probability. It can also be seen from the figure that the proposed algorithm offers higher successful transmission rate compared to two other algorithms. That is because the factors affecting successful transmission including physical channel characteristics, access collision and the bandwidth of candidate RVs are taken into account in the proposed algorithm, whereas two other algorithms fail to consider extensively.

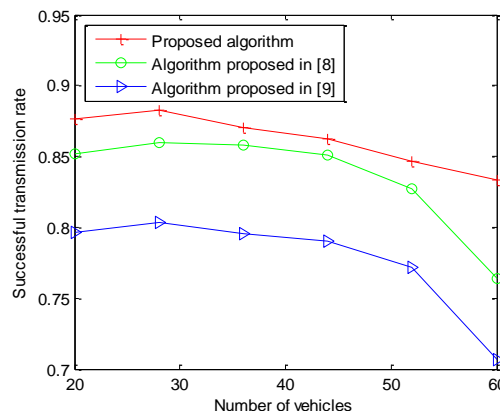


Figure 5. Successful transmission rate

VII. CONCLUSION

In this paper, a multi-object utility optimization based RV selection algorithm is proposed for VCN. The utility functions of both SVs and RVs are modeled respectively by choosing the metrics, i.e., available RV bandwidth, collision probability, packet successfully received probability, link capacity and stability, and different utility functions are designed for SVs and RVs. To maximize the utility of each SVs and RVs, a multi-object optimization problem is formulated and solved based on ideal point method. Numerical results demonstrate that compared to previous algorithms, the proposed algorithm offers better performance in terms of throughput, transmission delay and successful transmission rate.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (61102063), National Science and Technology Specific Project of China (2011ZX03005-004-02), the special fund of Chongqing key laboratory (CSTC) and the project of Chongqing Municipal Education Commission (Kjzh11206). The authors would

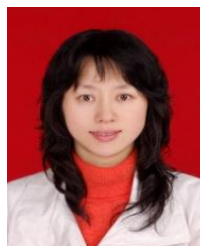
like to thank Dr. Niki Pissinou and the reviewers for their valuable comments.

REFERENCES

- [1] Z. Su, P. Y. Ren, and Y. Chen, "Consistency control to manage dynamic contents over vehicular communication networks," *IEEE Globecom 2011*, pp. 1-5, Dec. 2011.
- [2] M. L. Sichitiu and M. Kihl, "Inter-vehicle communication systems: a survey," *IEEE Commun. Surveys Tuts.*, vol. 10, pp. 88-105, Oct. 2008.
- [3] M. Jerbi, and S. M. Senouci, "Characterizing multi-hop communication in vehicular networks," *IEEE WCNC 2008*, pp. 3309-3313, Apr. 2008.
- [4] D. Lal, A. Manjeshwar, F. Herrmann, E. Uysal-Biyikoglu, and A. Keshavarzian, "Measurement and characterization of link quality metrics in energy constrained wireless sensor networks," *IEEE Globecom 2003*, pp. 446-452, Dec. 2003.
- [5] B. Zhao, and M. C. Valenti, "Practical relay networks: a generalization of hybrid-ARQ," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 7 -18, 2005.
- [6] J. Camp, and E. Knightly, "Modulation rate adaptation in urban and vehicular environments: cross-layer implementation and experimental evaluation," *IEEE/ACM Trans. Netw.*, vol. 18, no. 6, pp. 1949-1962, Dec. 2010.
- [7] H. Li, R. Chai, H. Hu, and Q. Chen, "An improved cross-layer RSU/mobile relay selection scheme in VANET," *IEEE Chinacom 2012*, pp. 481-486, Aug. 2012.
- [8] M. A. Alawi, R. A. Saeed, and A. A. Hassan, "Cluster-based multi-hop vehicular communication with multi-metric optimization," *International Conference on Computer and Communication Engineering (ICCCCE 2012)*, pp. 22-27, Jul. 2012.
- [9] Y. Bi, L. Cai, X. Shen, and H. Zhao, "A cross layer broadcast protocol for multihop emergency message dissemination in inter-vehicle communication," *IEEE ICC 2010*, vol. 10, pp. 88-105, Oct. 2010.
- [10] X. Ma, J. Zhang, X. Yin, and K. S. Trivedi, "Design and analysis of a robust broadcast scheme for VANET safety-related services," *IEEE Trans. Veh. Technol.*, Vol. 61, no. 1, pp. 46-61, Jan. 2012.
- [11] J. Yoo, B. S. C. Choi, and M. Gerla, "An opportunistic relay protocol for vehicular road-side access with fading channels," *IEEE ICNP 2010*, pp. 233-242, Oct. 2010.
- [12] S. Biswas, and R. Morris, "ExOR: opportunistic multi-hop routing for wireless networks," *SIGCOMM, 2005*.
- [13] A. S. Akki, F. Haber, "A statistical model of mobile-to-mobile land communication channel," *IEEE Trans. Veh. Technol.*, vol. 25, no. 1, pp. 2-7, 1986.
- [14] G. K. Karagiannidis, N. C. Sagias, and P. T. Mathiopoulos, "N*Nakagami: A novel stochastic model for cascaded fading channels," *IEEE Trans. Commun.*, vol. 55, pp. 1453-1458, Aug. 2007.
- [15] I. S. Gradshteyn, and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. New York: Academic, 2000.
- [16] T. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
- [17] R. D. Renesse, M. Ghassemian, V. Friderikos, and A. H. Aghvami, "Adaptive admission control for ad hoc and sensor networks providing quality of service", *Technical Report, King's College London*, 2005.
- [18] G. Bianchi, "Performance analysis of the IEEE 802. 11 distributed coordination function," *IEEE J. Select. Areas of Commun.*, vol. 18, pp. 535-547, Mar. 2000.

[19] IEEE 802. 11, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, 1999.

[20] G. Eichfelder, *Adaptive Scalarization Methods in Multi-objective Optimization*, Springer-Verlag, 2008.



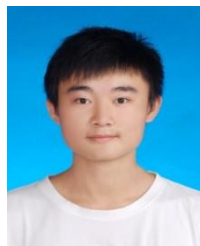
Rong Chai received her Ph.D. degree in electrical engineering from McMaster University in 2008. From April 2008, she joined Chongqing University of Posts and Technology as an associate professor. Her research interest is in wireless communication and network theory and has published over 40 papers in the area.



Bin Yang is a master student of Chongqing University of Posts and Technology. His research interest is in wireless communication and network theory, focusing on relay selection in vehicular communication network.



Li Cai is an undergraduate student of Chongqing University of Posts and Technology. Her research interest is in wireless communication and network theory, focusing on relay selection in vehicular communication network.



Xizhe Yang is an undergraduate student of Chongqing University of Posts and Technology. His research interest is in wireless communication and network theory, focusing on relay selection in vehicular communication network.



Qianbin Chen received his Ph.D. degree in electrical engineering from University of Electronic Science and Technology in 2006. He joined Chongqing University of Posts and Technology as a faculty member from 1988. He became a Member of IEEE in 2007. His research interest is in wireless communication, network theory and multi-media technology, and has published over 100 papers in the area.

The Effect of MAC Parameters on Energy Efficiency and Delay in Wireless Sensor Networks

Zhihua Li^a, Bin Lian^a, Zhongcheng Wei^b, Liang Xue^a, Jijun Zhao^a

^a School of Information and Electrical Engineering, Hebei University of Engineering, Handan, 056038, China
Email: yl_sandy@sina.com, lianbin620@163.com, liangxue@hebeu.edu.cn, zhaojjun@hebeu.edu.cn

^b School of information and communication engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China
Email: godframe@bupt.edu.cn

Abstract—As the most important indices to evaluate the performance of wireless sensor networks, energy efficiency and delay have been achieved more and more attention by researchers. In order to build the mathematical relationship between network parameters and performances, this paper starts from the three basic elements of communication and takes account backoff mechanism in sending period, bit error rate in transmission period and collision rate in receiving period. Using theoretical analysis and simulation confirmation, we can make a conclusion that packet payload length, bit error rate, node number, contention window size and the maximum retransmission times allowed have influences on packet drop rate, and then affect network performances.

Index Terms—wireless sensor networks, energy efficiency, backoff mechanism, collision rate, packet drop rate, packet payload length

I. INTRODUCTION

WIRELESS/ sensor networks (WSNs) have been utilized in considerable applications as both a connectively infrastructure and a distributed data generation network due to their self-organized and ubiquitous nature [1], [2]. Energy saving is regarded as the primary goal of WSNs studies because of the limited resources of sensor nodes. Emerging applications of WSNs require QoS guarantees, and the most important metric of QoS is transmission delay [3]. So energy efficiency and delay are two crucial metrics to evaluate the whole network performance and quality.

The factors lead to network resources consumption in WSNs mainly exist in three modules installed in a sensor node: perception module, processing module and communication module. Conflicts and errors are occurred in communication module, so the resources economization problem is solved by scheduling the communication module previously [4], [5]. Recent studies indicate that

some parameters optimization strategies can effectively achieve network performance improvements [6]–[8], but the premise is that the relationship of network parameters and performance had been figured out. To our knowledge, there is no a quite comprehensive energy efficiency and delay analytical model. The existed network performances analyzed are only suitable for one type of transmission mechanisms or the packet drop models are not precise.

In order to study the effect of MAC parameters on energy efficiency and delay in WSNs, this paper builds a more accurate performance analytical model. Different from other performance analytical models, it starts from the three basic elements of communication i.e. transmitting node, wireless channel and receiving node to analyze the critical factors including 1) collision probability, 2) packet error rate, and 3) packet drop rate. We adopt the retransmission mechanism, which can considerably guarantee reliable transmission under the circumstance of packet drop. A cyclic redundancy check (CRC) is preferred to check whether the received packet is wrong or not instead of any error correction coding so as to reduce the overhead. The main contributions of this paper are listed as follows:

- 1) Energy efficiency and delay performance analytical model is constructed. This paper analyzes and computes energy efficiency and delay in no retransmission and retransmission situations respectively. Analytical model is proposed by distinguishing bit error rate, collision rate and packet drop rate. We also consider the backoff mechanism, the maximum retransmission number limitation as well as the control overhead.
- 2) In the premise of CSMA/CA transmission mechanism, the input parameter set of network performance analytical model is also separately calculated in no retransmission and retransmission mechanisms. The relationship between network performance and bit error rate, node number, contention window size, maximum retransmission times allowed is further studied respectively.

The remainder of the paper is organized as follows:

Manuscript received August 21, 2013; accepted and revised November 4, 2013;

This work was supported by the National Natural Science Foundation of China (61304131), Department of Education Project of Hebei Province (No. Q2012045, Q2012088) and Science Technology Research and Development Fund of Handan (No. 1121103137).

Corresponding author email: lianbin620@163.com

Section II discusses the related works; IEEE802.15.4 MAC backoff mechanism is introduced in section III; section IV presents the analytical models and detailed computation of energy efficiency and delay performance; section V studies the critical factors including collision probability, packet error rate and packet drop rate; section VI is the simulation result; this paper is concluded in section VII.

II. RELATED WORKS

In the year of 2003, *I.F.Akyildiz* proposed energy efficiency in literature [9], which represents the energy consumed in useful information of a packet divided by the whole energy in one successful packet transmission. However, this model of energy efficiency is not accurate because that the effect of collision probability on performance is neglected. *MehmetC.Vuran* built a framework on throughput, energy per useful bit, and resource utilization in respect of network parameters in 2008 [10]. The computation of packet drop rate in this framework takes into account the packet error rate and collision probability, but it neglected the influence brought by retransmission situations. *Wang, Y.* proposed the delay distribution to meet a specific deadline for QoS-based communication in WSNs in 2012 [3], but it only captured the effects of wireless channel errors also neglected the collision rate in CSMA/CA mechanism.

Because a lot of researchers neglect the packet drop probability due to the reason of collisions, *A.Zakhori* gave a local estimation of collision probabilities with experimental method in 2011 [11]. They distinguished between packet loss due to channel error and collision and divided the collisions into two types: direct collisions, and staggered collisions. This is very helpful for constructing the network performance analysis model. In the IPSN conference 2012, *MarcoZimmerling* etc. presented pTunes [8], a framework for adaptation of MAC parameters. However, it just gave an abstractive system, the mathematical relation between network performance and MAC parameters is not presented clearly in this paper.

In summary, The MAC operating parameters bear great influence on the system performance. Energy efficiency and delay performance analytical models are worth to be constructed and studied.

III. PRELIMINARY WORKS

IEEE802.15.4 MAC uses CSMA/CA to access the finite channels. Backoff mechanism is the core of CSMA/CA algorithm, and it can avoid collisions obviously. The random number for the backoff counter C_R is chosen in the interval $(0, CW-1)$, where CW is the contention window size. The value of CW depends on the number of failed transmissions of a packet. At the first transmission attempt $CW=CW_{min}$, which is the minimum contention window. After each retransmission due to a collision, CW is doubled up to a maximum value, where CW_{max} is the largest contention window size. Once the CW reaches

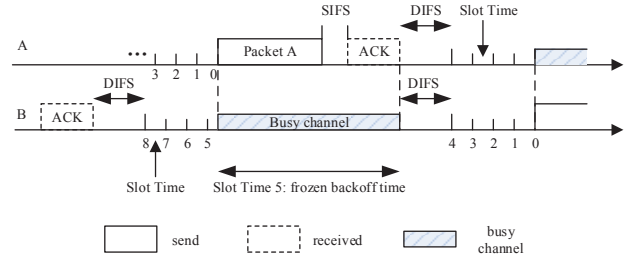


Fig. 1. Description of CSMA/CA backoff algorithm

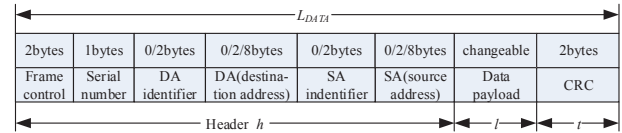


Fig. 2. Data frame format in MAC layer

CW_{max} , it will remain at the value of CW_{max} until it is reset [12], [13].

Figure 1 describes the contention relationship of node A and B. These two nodes can hear the state of each other and contend to access the same channel. If channel state is idle when the backoff counter of A reaches 0, that means A accesses the channel successfully. At the same time B detects that the channel state is busy and freezes its backoff counter. When the transmission of A has finished, channel state will go back to be idle. B detects the idle time equal to DIFS and recovers its backoff counter. When its backoff counter reaches 0, B can use the common channel. From Figure 1, we know that collision may occur only when two (or more) packets are transmitted within the same slot time [14].

IV. ENERGY EFFICIENCY AND DELAY ANALYSIS

A. Energy efficiency analysis

Control overhead, packet drop rate and the number of retransmissions are the critical factors that affect the transmission energy efficiency of a packet. We develop the energy consumption Eq. (1) based on literature [9] and [15].

$$E_{comm} = E_{TX} + E_{RX} \tag{1}$$

E_{TX} represents the energy consumption of sending a packet from the source node A; and E_{RX} represents the energy consumption of receiving a packet in destination node C. Figure 2 shows the common format of MAC data packet. We assume that the frame header length is h bit, frame load length is l bit, and frame tail length is t bit.

$$E_{TX} = \frac{P_{T-elec}}{R} \cdot L_{DATA} + \frac{P_t}{R} \cdot L_{DATA} + \frac{P_{R-elec}}{R} \cdot L_{ACK} + P_{T-start}t_{T-start} + P_{R-start}t_{R-start} \tag{2}$$

$$E_{RX} = \frac{P_{R-elec}}{R} \cdot L_{DATA} + \frac{P_{T-elec}}{R} \cdot L_{ACK} + \frac{P_t}{R} \cdot L_{ACK} + P_{T-start}t_{T-start} + P_{R-start}t_{R-start} \tag{3}$$

P_{T-elec}/P_{R-elec} : Running power consumed in the transmitter/receiver electronics.

P_t : Output transmitting power.

$P_{T-start}/P_{R-start}$: Startup power consumed in transmitting/receiving components equipped on nodes.

$t_{T-start}/t_{R-start}$: Transmitter/receiver start-up time, they can also be explained as the transition time from sleep mode or idle mode to transmit/receive modes.

R : Transmission rate.

We insert Eq.(2) and Eq.(3) into Eq.(1). Then another Eq. (4) on packet energy consumption is obtained.

$$E_{comm} = \left(\frac{P_{T-elec} + P_{R-elec} + P_t}{R} \right) \cdot (L_{DATA} + L_{ACK}) + 2 \cdot (P_{T-start}t_{T-start} + P_{R-start}t_{R-start})$$

$$= k_1 \cdot (L_{DATA} + L_{ACK}) + 2 \cdot k_2 \quad (4)$$

$$k_1 = \frac{(P_{T-elec} + P_t) + P_{R-elec}}{R} \quad (5)$$

$$k_2 = P_{T-start}t_{T-start} + P_{R-start}t_{R-start} \quad (6)$$

k_1 is the energy consumption by transmitting one bit information; k_2 is the start-up energy consumption by transmitting one packet. The energy efficiency is the energy consumed on useful information percentage in the total energy consumption for transmitting one packet successfully. Eq.(7) is the energy efficiency without considering retransmission.

$$\eta_{comm} = \frac{k_1 \cdot l}{E_{comm}} \cdot (1 - p_{drop}) \quad (7)$$

p_{drop} is the probability that fail to transmit a packet.

Eq.(4) and Eq.(7) represent energy consumption and energy efficiency under the circumstance that the incorrect packet will be dropped directly and retransmissions are not allowed. However, as to the retransmission mechanism, the packet drop rate changes from p_{drop} to p'_{drop} , because it is not only associated with the failed transmission probability, but also need to consider the limitation of maximum retransmission number allowed, presented by m . The expressions of energy consumption and energy efficiency in retransmission mechanism will be changed to the following two Eq.(11) and Eq.(12). The calculation process is as following.

$$E_{comm}' = E_{TX}' + E_{RX}' \quad (8)$$

$$E_{TX}' = N \cdot \frac{P_{T-elec}}{R} \cdot L_{DATA} + N \cdot \frac{P_t}{R} \cdot L_{DATA} + N \cdot P_{T-start} \cdot t_{T-start} + E_{TX} \quad (9)$$

$$E_{RX}' = N \cdot \frac{P_{R-elec}}{R} \cdot L_{DATA} + N \cdot P_{R-start} \cdot t_{R-start} + E_{RX} \quad (10)$$

$$E_{comm}' = N \cdot k_1 \cdot L_{DATA} + N \cdot k_2 + E_{comm}$$

$$= (N + 1) \cdot k_1 \cdot L_{DATA} + k_1 \cdot L_{ACK} + (N + 2) \cdot k_2 \quad (11)$$

$$\eta_{comm}' = \frac{k_1 \cdot l}{E_{comm}'} \cdot (1 - p_{drop}') \quad (12)$$

N is the average retransmissions times, p_{drop}' represents the packet drop rate with retransmission.

B. Delay analysis

m is the maximum retransmissions number allowed. According to Figure 1, time delay for transmitting one packet t_{comm} mainly contains channel access time, transmission time and transmitting/receiving components start-up time, showed in Eq. (13).

$$t_{comm} = DIFS + t_{backoff} + t_{DATA} + SIFS + t_{ACK} + t_{T-start} \quad (13)$$

$$t_{backoff} = C_R \times t_{slot}, (0 \leq C_R \leq CW) \quad (14)$$

$$t_{DATA} + t_{ACK} = \frac{L_{DATA} + L_{ACK}}{R} \quad (15)$$

$$t_{T-start} = C \quad (16)$$

$t_{backoff}$ is the backoff time; t_{slot} is a slot time. C_R is the backoff counter in no retransmission condition. The value of C_R is between 0 and $CW-1$. For example, channel access contention window CW equals to 25, so C_R would be any integer among $[0, 24]$. $t_{DATA}+t_{ACK}$ is the data and ACK packets transmission time; the star-up time of transmitting and receiving components $t_{T-start}$ equals to a constant and may almost be neglected compared with $t_{backoff}$ and $t_{DATA}+t_{ACK}$.

To a certain extent, retransmission is able to guarantee the packet transmission reliability, but it adds the latency. We assume that the value of CW is a constant number between CW_{min} and CW_{max} , that is to say the value of CW has no relation to retransmission. But C_R is a variable still in the range of $[0, CW-1]$. $C_R^{(i)}$ presents the value of the backoff counter in i th retransmission stage. This assumption simplifies computation but still can guarantee the rationality of backoff mechanism. Eq.(17) is the time delay to transmit one packet by using the retransmission mechanism.

$$t_{comm}' = t_{ACK} + \sum_{i=0}^N \left(DIFS + C_R^{(i)} \times t_{slot} + t_{DATA} + SIFS + t_{T-start} \right) \quad (17)$$

It is obvious that the energy efficiency and delay are the functions of packet length, packet drop rate, retransmission number and other parameters through the observation of formulae (12) and (17). Parameters are divided into two different types: performance parameters, such as packet drop rate p_{drop}' , and state parameters, like $l, h, t, L_{ACK}, n, m, R, k_1, k_2$. Performance parameters can be calculated by more simple parameters, and state parameters can be directly given constant values. Network performances

are directly influenced by state parameters, which do not change by any protocol or transmission mechanism. Performance parameters have different calculation methods according to the type of transmission mechanisms.

V. PARAMETERS ANALYSIS AND COMPUTATION

The task of Medium access control (MAC) layer is cutting long data information collected by transmitting node into short packets and transmits them to receiving node. Wireless communications suffers from many time-varying phenomena including signal attenuation, channel fading due to multi-path propagation, and interference caused by other transmissions at overlapping frequencies [16], [17].

Bit errors and conflicts are two main factors that can bring about the packet drop rate [11], [18]. In retransmission mechanism, the influence factors of packet drop rate also include the limitation of retransmission number. Nodes cannot distinguish between packet loss due to channel error and collisions because the symptoms are the same. But the solutions to the two problems are different. The channel error can be overcome by using link adaptation or forward error correction at the application layer, collision avoidance is usually achieved by means of Binary Exponential Backoff scheme [19]. At first, we will analyze the packet error rate caused by bit error rate.

A. Packet error rate

Energy efficiency and delay performance are related to the nondeterministic impacts of the wireless channel and the collisions in nodes competition. It is difficult to maintain a good performance continuously as a result of wireless channel intrinsic error-prone and changeable characteristics [20], [21]. Bit error rate BER is the probability that one bit goes awry in data transmission process, and the packet length we have already known is L_{DATA} bit. We can finally calculate the probability of failure in a packet transmission, that is packet error rate PER .

$$PER = 1 - (1 - BER)^{L_{DATA}} \quad (18)$$

We consider the multiple fading, so the Rayleigh fading channel is chosen as channel model. According to [9], bit error rate is related to pass loss exponent β and transmission distance d and BER is between 7×10^{-4} and 3×10^{-3} when $\beta=3.5$ and d is 20-30 meters.

B. Collision rate

If the backoff counter of node A has already decreased to zero, accessing collisions would happen if another node tries to transmit data packets at the same time. The backoff mechanism is introduced in detail as shown in section III. From Figure 1, we know that collision rate p_C is the probability that at least one of the $n-1$ remaining nodes transmits in a time slot [22], where τ is the transmission probability, n is the number of nodes.

$$p_C = 1 - (1 - \tau)^{n-1} \quad (19)$$

A discrete-time Markov chain in [12] calculates the transmission probability τ . It is related to collision probability p_C , maximum contention window CW_{max} , minimum contention window CW_{min} , and maximum retransmission number allowed m . It is assume that the contention window CW is a constant, which does not change with retransmission times. So transmission probability is only dependent on CW [14], Eq.(20) is the computation equation.

$$\tau = \frac{2}{CW + 1} \quad (20)$$

C. Packet drop rate

Through the analysis above, we can compute the packet drop rate p_{drop} , which is caused by channel errors or collisions detected in receiving nodes.

$$p_{drop} = 1 - (1 - PER)(1 - p_C) \quad (21)$$

We drop the packet which has wrong bits or conflicts with others immediately at receiving node. If ACK has not been received by transmitting node in a predefined time, the initial packet saved in the buffer of transmission node will be retransmitted in the limitation number of retransmissions. On one hand, this mechanism increases the packet average latency, energy consumption. On the other hand, it reduces the packet drop rate of networks. It can guarantee the continuity and reliability of transmission information. Packet drop rate without retransmission p_{drop} has been obtained in Eq.(21). In retransmission mechanism, packet drop rate is the probability that the retransmission number in reality beyond the maximum retransmission number. m is the limitation retransmissions times, and the packet drop rate in retransmission mechanism p'_{drop} is calculated by Eq.(22).

$$p'_{drop} = p_{drop}^{m+1} \quad (22)$$

Packet average retransmission number is dependent on packet drop rate without retransmission p_{drop} and the maximum retransmission number m . We can get packet average retransmission number N by calculating the expectation of p_{drop} .

$$N = \sum_{i=1}^m (i \cdot p_{drop}^i \cdot (1 - p_{drop})) \quad (23)$$

To sum up, packet drop rate is caused by two reasons: errors and collisions. So we focused on studying that which parameters are responsible for packet error rate PER and collision rate p_C in the beginning of this section. According to Eq.(18), packet error rate PER is the function of BER and data packet length L_{DATA} . Bit error rate BER is influenced by environment of networks, its value is between 7×10^{-4} and 3×10^{-3} when the channel model is Rayleigh fading. Eq. (19) provides the collision rate which is affected by contention window size CW and node number n , so packet drop rate is affected by BER , L_{DATA} , CW , and n . In retransmission mechanism, the factors that affect packet drop rate also include the maximum retransmission number allowed m .

TABLE I
NETWORK PARAMETERS

Parameters	Value	Parameters	Value
k_1	1.85uJ/bit	R	20Kbps
k_2	24.86uJ	BER	0.0007
h	224bit	$DIFS$	50us
t	16bit	$SIFS$	20us
L_{ACK}	112bits	n	20
CW	256	m	3
t_{slot}	20us		

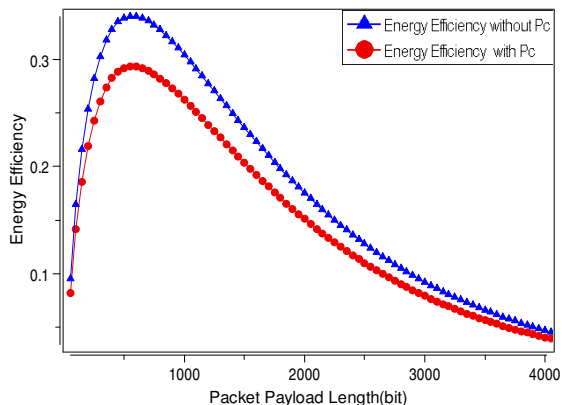


Fig. 3. Comparison of Energy Efficiency with and without collision

VI. SIMULATION RESULTS

In this part, we model the wireless sensor networks as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of edges representing communication links. The graph includes n nodes randomly deployed in a unit square. Energy efficiency and delay have been calculated, and the network parameters have been analyzed. According to the results, we know that BER, l, CW, n and m are parameters that have influence on performance metrics. At first, we focus on the change of energy efficiency and delay with respect of l . A set of typical parameters are listed in Table I.

3 is the comparison of energy efficiency with and without considering collisions. The blue curve is the energy efficiency only affected by bit error rate with variable packet length. In reality, the curve of energy efficiency is lower than the blue one especially when packet length is about 550 bits. The red curve is the description of energy efficiency that considers collision rate. There is an optimal packet payload length that makes energy efficiency maximal. When the packet payload length is shorter than the optimal value, a big part of total energy is consumed in control overhead. So the percentage of energy consumed in transmitting useful data is getting smaller. Energy efficiency declines. On the contrary, if packet payload length is longer than optimal value, we can see that longer packet payload length leads to bigger packet drop rate from formulae (18) and (21). So energy efficiency goes down.

Different transmission mechanisms have great influences on network performances, which we have introduced in section IV. Energy efficiency and delay without and with retransmission are given in Figure 4 and Figure

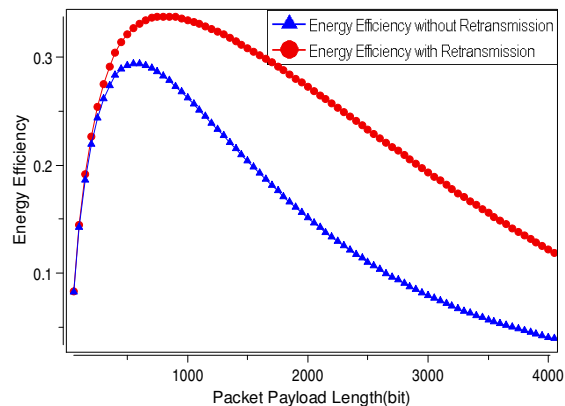


Fig. 4. Comparison of energy efficiency with and without retransmission

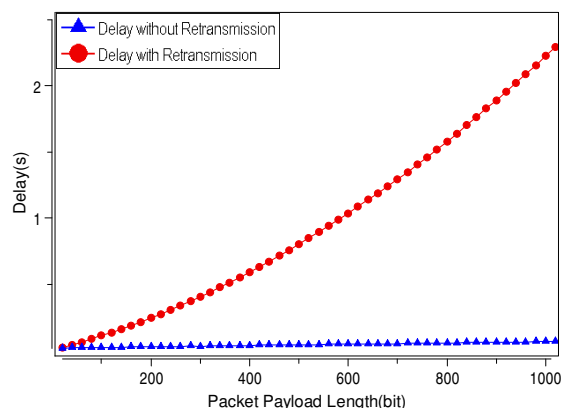


Fig. 5. Comparison of delay with and without retransmission

5 respectively.

Retransmission mechanism increases the average energy consumed by transmitting a packet according to Eq.(11), but it decreases the packet drop rate obviously, as showed in Eq.(22). We can see from Figure 4 that the curves of energy efficiency in the two mechanisms are almost same when packet length is smaller than 500 bits, but when packet payload length exceeds 500 bits, retransmission mechanism is more effective to guarantee network energy efficiency performance.

Figure 5 is the delay comparison in no retransmission and retransmission conditions. In retransmission mechanism, the average transmission number of a packet is more than 1. As packet length increases, packet drop rate will increase, thus the retransmission number will increase. Every retransmission needs to contend the channel again, and backoff time becomes increase. So delay in retransmission mechanism is greater than that in no retransmission mechanism, and it increases quickly along with the packet payload length.

In table I, the bit error rate BER , node number n , contention widow size CW , and maximum retransmission number allowed m are given constant values. But BER affects packet error rate according to Eq.(18), n and CW affect the collision rate according to Eq.(19) and Eq.(20), and m is related to the packet drop rate and the average

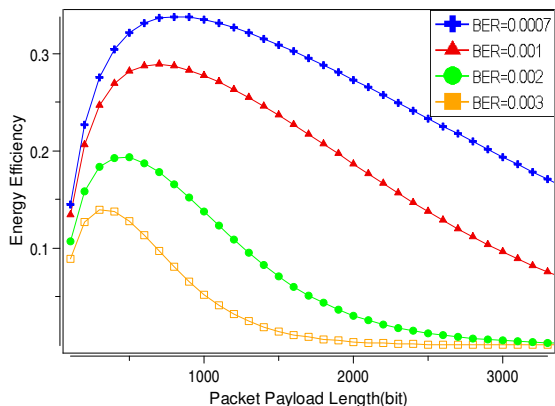


Fig. 6. Energy efficiency with different bit error rate *BER*

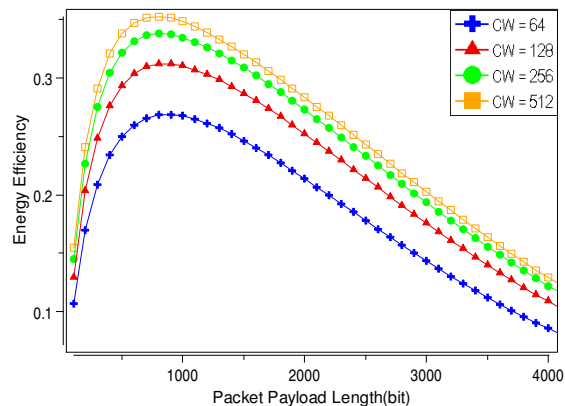


Fig. 8. Energy efficiency with different contention window size *CW*

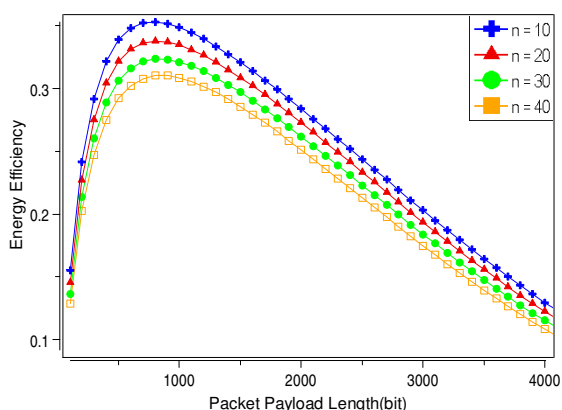


Fig. 7. Energy efficiency with different node number *n*

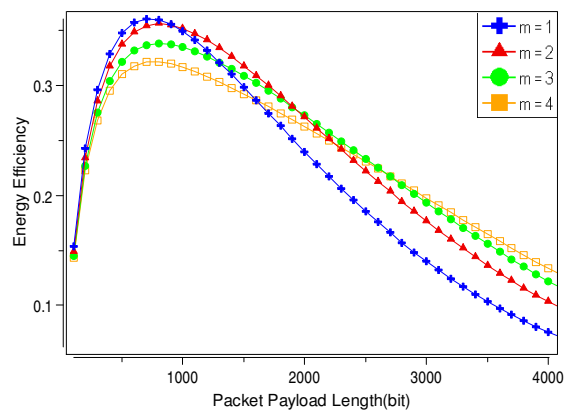


Fig. 9. Energy efficiency with different *m*

retransmission number according to Eq.(22) and Eq.(23). So the influences of these four parameters on energy efficiency and delay are deserved to discuss deeply.

Figure 6 describes the change of energy efficiency with respect of packet payload length when bit error rate is 0.0007, 0.001, 0.002 and 0.003 respectively. Packet drop rate goes up along with the increase of bit error rate, so energy efficiency becomes worse. The optimal packet payload length, which makes the energy efficiency maximal, becomes small with the increase of *BER*. As to delay, packet drop rate goes up, retransmission number will get bigger, and delay will increase with the increase of *BER*.

Figure 7 shows that energy efficiency deteriorates when the value of contention node number is chosen as 10, 20, 30 and 40 successively. Because larger node number causes higher packet collision rate and packet drop rate as shown in Eq.(19) and Eq.(21), which further leads to the decrease of energy efficiency and the increase of delay.

Figure 8 indicates that the value of contention window size increases, the probability of collision gets lower. The energy efficiency is getting better. But the large backoff time gives rise to the increase of delay with the increase of *CW*.

In Figure 9, the value of maximum retransmission number allowed is given by 1, 2, 3 and 4. When the packet payload is relatively small, larger *m* leads to the lower

energy efficiency. That is because energy consumption plays the most important role in energy efficiency at this time, when *m* becomes larger, the retransmission number will become larger, and energy consumption will become larger. So at this time energy efficiency will decrease with the increase of *m*. When the packet payload length becomes longer and longer until exceeds a boundary, packet drop rate will play the most important role in energy efficiency. When *m* becomes larger, packet drop rate will be lower. At this time, energy efficiency will increase with the increase of *m*.

VII. CONCLUSIONS

In the premise of CSMA/CA, we built a considerable comprehensive energy efficiency and delay performance model. It took into account the contention collisions and bit error rate to calculate the packet drop rate. Control overhead and retransmission mechanism are also taken into consideration. Retransmission mechanism is better for energy efficiency, but increases delay. Packet error rate and collision rate are two reasons that cause packet drop rate. In retransmission mechanism, packet drop rate is also influenced by retransmission number. At last, we explored the relation between energy efficiency and some key parameters, such as packet payload length, node number, bit error rate, contention window and maximum retransmission number. In future work, we plan to extend our

network performance analytical model by adding hidden terminal problem and design a variable packet payload length adaptation algorithm to optimize the performance metrics.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for their valuable comments and suggestions to improve the presentation of this paper.

REFERENCES

- [1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.
- [2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.
- [3] Y. Wang, M. C. Vuran, and S. Goddard, "Cross-layer analysis of the end-to-end delay distribution in wireless sensor networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 1, pp. 305–318, 2012.
- [4] M. A. Yigitel, O. D. Incel, and C. Ersoy, "Qos-aware mac protocols for wireless sensor networks: A survey," *Computer Networks*, vol. 55, no. 8, pp. 1982–2004, 2011.
- [5] Y. Sun, S. Du, O. Gurewitz, and D. B. Johnson, "Dw-mac: a low latency, energy efficient demand-wakeup mac protocol for wireless sensor networks," in *Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2008, pp. 53–62.
- [6] C.-L. Chen, C.-Y. Yu, C.-C. Su, M.-F. Horng, and Y.-H. Kuo, "Packet length adaptation for energy-proportional routing in clustered sensor networks," in *Emerging Directions in Embedded and Ubiquitous Computing*. Springer, 2006, pp. 32–42.
- [7] X. Wu, H. R. Sadjadpour, and J. Garcia-Luna-Aceves, "From link dynamics to path lifetime and packet-length optimization in manets," *Wireless Networks*, vol. 15, no. 5, pp. 637–650, 2009.
- [8] M. Zimmerling, F. Ferrari, L. Mottola, T. Voigt, and L. Thiele, "ptunes: Runtime parameter adaptation for low-power mac protocols," in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*. ACM, 2012, pp. 173–184.
- [9] Y. Sankarasubramaniam, I. F. Akyildiz, and S. McLaughlin, "Energy efficiency based packet size optimization in wireless sensor networks," in *Sensor Network Protocols and Applications, 2003. Proceedings of the First IEEE. 2003 IEEE International Workshop on*. IEEE, 2003, pp. 1–8.
- [10] M. C. Vuran and I. F. Akyildiz, "Cross-layer packet size optimization for wireless terrestrial, underwater, and underground sensor networks," in *INFOCOM 2008. The 27th Conference on Computer Communications*. IEEE, 2008, pp. 226–230.
- [11] M. Christine, M. N. Krishnan, S. Ng, E. Haghani, and A. Zakhor, "Local estimation of collision probabilities in 802.11 w lans: An experimental study," in *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*. IEEE, 2011, pp. 43–48.
- [12] P. Chatzimisios, A. C. Boucouvalas, and V. Vitsas, "Ieee 802.11 packet delay-a finite retry limit analysis," in *Global Telecommunications Conference, 2003. GLOBECOM'03. IEEE*, vol. 2. IEEE, 2003, pp. 950–954.
- [13] Y. Zhang and F. Shu, "Packet size optimization for goodput and energy efficiency enhancement in slotted ieee 802.15. 4 networks," in *Wireless Communications and Networking Conference, 2009. WCNC 2009. IEEE*. IEEE, 2009, pp. 1–6.
- [14] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 3, pp. 535–547, 2000.
- [15] M. C. Domingo, "Packet size optimization for improving the energy efficiency in body sensor networks," *ETRI Journal*, vol. 33, no. 3, pp. 299–309, 2011.
- [16] M. Vutukuru, H. Balakrishnan, and K. Jamieson, "Cross-layer wireless bit rate adaptation," in *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4. ACM, 2009, pp. 3–14.
- [17] F. Hermans, O. Rensfelt, T. Voigt, E. Ngai, L.-Å. Nordén, and P. Gunningberg, "Sonic: classifying interference in 802.15. 4 sensor networks," in *Proceedings of the 12th international conference on Information processing in sensor networks*. ACM, 2013, pp. 55–66.
- [18] S.-C. Wang and A. Helmy, "Beware: Background traffic-aware rate adaptation for ieee 802.11," *IEEE/ACM Transactions on Networking (TON)*, vol. 19, no. 4, pp. 1164–1177, 2011.
- [19] M. N. Krishnan, E. Haghani, and A. Zakhor, "Packet length adaptation in w lans with hidden nodes and time-varying channels," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*. IEEE, 2011, pp. 1–6.
- [20] W. Dong, X. Liu, C. Chen, Y. He, G. Chen, Y. Liu, and J. Bu, "D-pltc: Dynamic packet length control in wireless sensor networks," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.
- [21] G. W. Challen, J. Waterman, and M. Welsh, "Idea: Integrated distributed energy awareness for wireless sensor networks," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 35–48.
- [22] B. Jang and M. L. Sicitu, "Ieee 802.11 saturation throughput analysis in the presence of hidden terminals," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 2, pp. 557–570, 2012.

Zhihua Li (1978-), female, born in Handan, Hebei province, China. She received the Ph.D. degree from Hebei University of Technology, Tianjin, China. Now she is an associate professor and works at Hebei University of Engineering. Her main research interests include Wireless Sensor Network and Intelligent Electric Apparatus.

Bin Lian(1988-), female, born in Handan, Hebei province, China. She received a B.S. degree in Electronic and Information Engineering from Hebei University of economics and business, China, in 2011, and now is continuing his studies at Hebei University of Engineering for M.S. degree in Wireless Sensor Network. Her research interests include energy conservation and the MAC layer designing in wireless sensor networks.

Zhongcheng Wei(1987-), male, born in Shangqiu, Henan province, China. He received a M.S. degree from Hebei University of Engineering, and now he is continuing his studies at Beijing University of Posts and Telecommunications for Ph.D. degree. His research interests include body sensor networks and data gathering in wireless sensor networks.

Liang Xue(1982-), male, born in Handan, Hebei province, China. He received the Ph.D. degree from Yanshan University, China. Now he is a lecturer and works at Hebei University of Engineering. His main research interests include Wireless Sensor Network and Wireless Cognitive Networks.

Jijun Zhao(1970-), male, born in Handan, Hebei province, China. He received the Ph.D. degree from Beijing University of Posts and Telecommunications and accomplished postdoctoral research in ZTE. Now he is a professor in Hebei University of Engineering. His main research interests include Wireless Sensor Network and Broadband Communication Network.

A Multicast Routing Algorithm for Datagram Service in Delta LEO Satellite Constellation Networks

Yanpeng Ma, Xiaofeng Wang, Jinshu Su, Chunqing Wu, Wanrong Yu, and Baokang Zhao
 College of Computer, National University of Defense Technology, Changsha, China
 Email: {yanpengma, xf_wang, sjs, wuchunqing, wlyu, bkzhao}@nudt.edu.cn

Abstract—Satellites can broadcast datagram over wide areas, therefore, the satellite network has congenital advantages to implement multicast service. LEO satellite has the property of efficient bandwidth usage, lower propagation delay and lower power consumption in the user terminals and satellites. Therefore, the constellation network composed by LEO satellites is an essential part of future satellite communication networks. In this paper, we propose a virtual center based multicast (VCMulticast) routing algorithm for LEO satellite constellation network. The algorithm uses the geographic center information of group users to route multicast datagrams, with less memory, computer power and signaling overhead. We evaluate the delay and performance of our algorithm by means of simulations in the OPENET simulator. The results indicate that the delay of the proposed multicast method exceeds the minimum propagation by at most 29.1% on the average, which is a quite acceptable achievement, considering the resource overhead reduction that can be introduced by our proposal.

Index Terms—Multicast Routing; Low Earth Orbit (LEO); Satellite Constellation Networks

I. INTRODUCTION

Multicast is the action of delivering information to a group of nodes at different locations simultaneously with efficient strategies. It delivers messages through each link of the network only once, and creates multiple message copies only at the diverging links. Since satellites can broadcast datagrams over a very wide area, it is becoming a hot research topic to implement multicast service over satellites networks¹.

As indicated in [1], few of the existing multicast routing algorithms can be directly implemented for satellite network. The reason is that they need periodic exchanges of some types of message to create and maintain the trees for multicast routing algorithm, which are impractical due to the time-variety characteristic of satellite constellation network topology. Hence, several related work attempt to hide the dynamic characteristic of networks before the design of routing algorithm. It can be achieved by two ways: Dynamic Virtual Topology (DVT)

[2] and Virtual Node (VN) [3].

DVT divides the system cycle (orbit period of satellite multiplied by the self-rotation period of the earth) into fixed time slices. In each of this slice, the network topology is considered to be stable. Changes of the topology arise only at the beginning of the time slice. The network routing table can be calculated offline and stored into the satellite memory. This can reduce the computational overhead of each satellite node. However, it also increases the storage requirement of satellite nodes. Moreover, once the network node or link fails, this kind of routing algorithm will fail to cope with the failure. Hence, this method is mainly employed in the research of unicast routing algorithm.

VN tries to divide the physical earth surface into an invariant logical topology composed by virtual nodes. The virtual node is embodied by a satellite all the time, so that the network composed of virtual nodes is embodied by the satellite constellation all the time. Through this way, the satellite's mobility can be hidden. Since the topology is invariant, the routing algorithm can use the mature routing algorithms which are designed for the topological invariant network. There have been several multicast routing algorithms proposed in [1] [4-6] that are based on VN method. However, as mentioned in [7], the mapping between the VNs and the real satellite nodes is not as simple as it seems. On some special conditions, the mapping between satellite and virtual node will cause a problem that the mobile node cannot communicate with the satellite node. Thus, further work is required to improve the VN method.

The only algorithm that does not hide the dynamic topology is proposed in [9]. It has to acquire the information of all the multicast group members before constructing the core based multicast tree [19], and this result in a higher signaling and memory overhead. In Ref [9], a simple vector algorithm is suggested, which selects the core satellite based on aggregating the locations of group members. Although this strategy can select a core satellite which will be nearest to the center of group members in the network, it may perform poorly in many cases since satellites moves very fast relative to the earth surface. Because the relationship between the satellites and group members change ceaselessly, The core satellite of network is ceaselessly changing and it will incur

1. In this paper, we use satellite network, LEO satellite network and LEO satellite constellation network interchangeably.

considerable network overhead of signaling and processing.

In this paper, we propose a virtual center based multicast (VCMulticast) algorithm for the Delta type [9] satellite constellation without hiding the dynamic of topology of satellite network. The VCMulticast aims at reducing the system signaling and memory overhead, and increasing the system scalability. The main contributions of VCMulticast are:

First, by utilize relative static character “position center” of the group members to route multicast datagram, our work does not need to hide the time-variety character of the satellite network topology. This avoids the map operation between the logical invariant network and the time-variety network, which can greatly reduce the signaling overhead.

Second, we divide the satellite network into clusters according to orbit planes. Each satellite only acquires all the information of group members that their access satellites are within the same cluster. For the other clusters, only the latitude centers (*LC*) of group members information is required. The number of planes and number of satellites per-plane is around ten. Hence, the memory overhead of our method is very low.

And third, we employ a geographic based unicast like algorithm to route inter-plane multicast datagram. Through this way, we improve the scalability of the algorithm since it does not need to store the huge amount of group member information to forward multicast datagram.

The VCMulticast can greatly reduce the signaling and memory overhead for the system, and increase the system scalability. Experimental results show that the delay of the proposed multicast method exceeds the minimum propagation by at most 29.1% on the average.

This paper is organized as follows. In Section II, we present the system architecture as well as some interesting relative invariant properties of the mobile node of Delta LEO satellites networks. In Section III, we describe VCMulticast scheme. In Section IV we evaluate the performance of VCMulticast scheme through simulation, and in Section V we conclude our research.

II. SYSTEM ARCHITECTURE

The system architecture of satellite network is shown in Fig. 1. It consists of two layers: Terrestrial Layer and LEO Satellite Layer.

A. Terrestrial layer

The terrestrial layer is consisted of two segments, i.e., the user segment and the ground segment. The user segment includes all the mobile nodes can be classified according to the mobility, size and transmission rate (e.g., vehicles, street walkers). They are considered as “multicast group members” in our work. The ground segment includes gateways (GWs, i.e., FESs), Network Control Center (NCC), and Satellite Control Center (SCC). GWs connect the terrestrial part of Internet with the satellite network. NCC is responsible for managing user access and storing user profiles of all mobile nodes. SCC monitors the performance of the satellite

constellation and controls a satellite’s position in the sky [8].

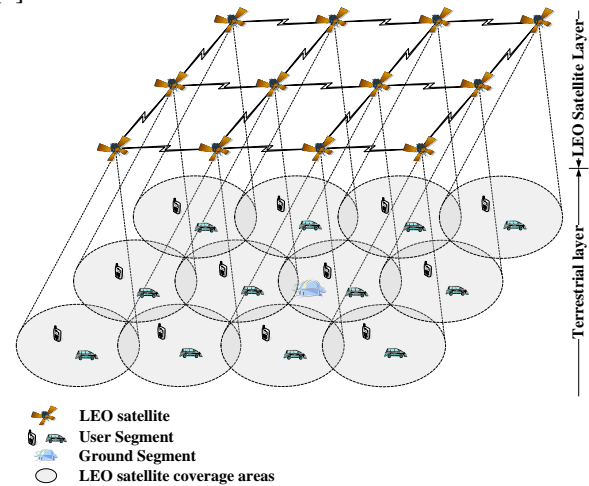


Figure 1. System architecture

B. LEO Satellite Layer

The LEO satellite layer includes a set of LEO satellites with on-board processing capabilities. The LEO satellites act as routers which carry out the functions of multicast routing and forwarding. Our work focus on the inclined Delta satellite constellation [9] with inter-satellite links (ISLs), and it can be easily extended to the Star satellite constellation [9]. As Fig. 2 depicts, the Delta satellite constellation employs satellites with inclined circular orbits. This type of satellite constellation can set up a satellite network with a number of ISLs that can be maintained permanently with the requirements of positioning, acquisition and tracking [11]. In this work, we set our research basis to be in the NeLS [13] constellation with ISL connections. The logical satellite connection of this constellation is illustrated in Fig. 3. Besides the bidirectional communication capability with mobile nodes, we also assume that each satellite has bidirectional wireless links with the neighboring LEO satellites.

The mobile node of Delta LEO satellite network has specific relative invariant properties is used to construct our multicast routing algorithm, which are summarized as follows:

(1) Surface property of access satellite

Satellites in the Delta constellation can be separated into an ascending surface and a descending surface [9]. All satellites move from south-west to north-east, composing a mesh-like surface called ascending surface. The descending surface includes satellites moving from north-west to south-east. If the access satellites of two mobile nodes belong to different surfaces, the traffic must travel over the highest latitudes via ISLs, and the length of communication path will be longer than the case in which the access satellites belong to the same surface. To reduce the delay jitter between the consecutive datagrams, mobile nodes are required do its best to access satellites in the same surface when the satellite handover occurs.

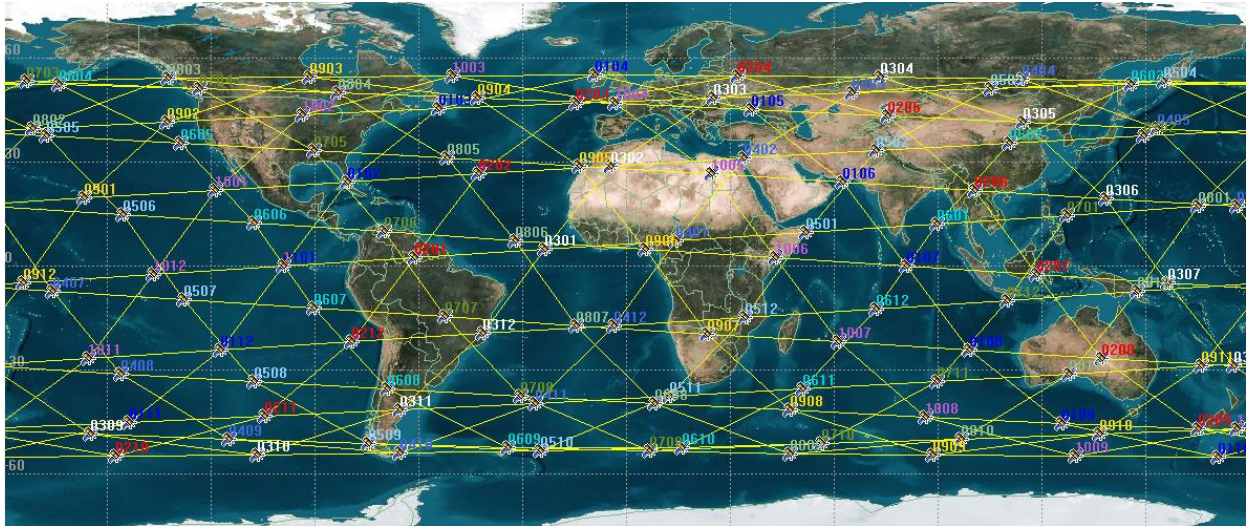


Figure 2. Delta constellation (NeLS, drawn by STK [10])

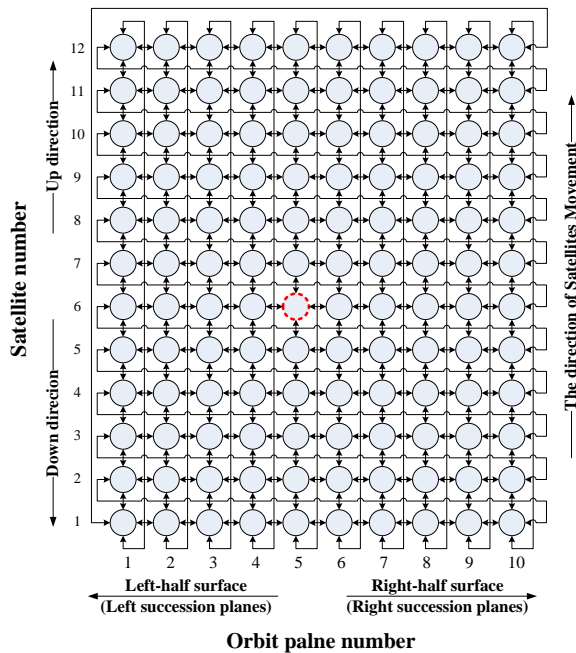


Figure 3. Logical connections for satellite ISLs [12]

(2) The types of access area of mobile nodes

Normal Area: in low latitude region of earth (about 64° to -64° in our constellation for research), satellite network provides double surface coverage [9]. The mobile node just has to connect with the satellite only in one (ascending or descending) surface to maintain the connection with satellite network.

Special Area: in high latitude region towers the limits of north or south coverage of satellite network (about 64° to 70° of north and south in our constellation for research), the mobile node that only connects with satellite in one surface couldn't be guaranteed to keep on connecting with the satellite network. This is because the satellite handover is dictated by available coverage. However, mobile node can keep connected with satellite network if the accessing satellite can be selected freely.

For normal mobile nodes such as vehicles and street walkers, this property is not frequently changed.

(3) Orbit selection of mobile node:

The orbital period of LEO satellite is not greater than 128 minutes. Therefore, the satellite rotates around earth for at least twelve rounds and the earth rotates once on its axis every 24 hours. We can easily deduce that the satellite handover of a mobile node takes place between the satellites that either belong to the same plane or belong to two neighboring planes. As shown in Fig. 4, even in the special areas the satellite handover mainly takes place between two neighbor planes. We also notice that this property depends on the satisfaction of property 1. In the normal area, if a mobile node connects access satellite without considering the surface of satellite, this property will not be guaranteed.

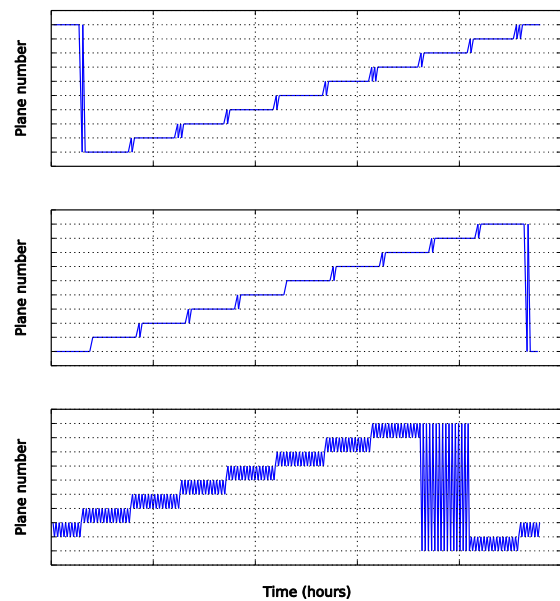


Figure 4. The plane selection of inter-satellite handover

III. THE VCMULTICAST SCHEME DESCRIPTION

VCMulticast is a source-based [14] scheme, but no real multicast tree is maintained. Thus, it does not require the

flood and prune process or short path algorithm to construct the multicast trees. The satellite node has to do its best to forward the datagram to the center of the group members that have not received the datagram. This forwarding is based on geographic unicast schemes, such as Compass routing [15]. The geographic unicast scheme requires no efforts on the establishment and maintenance of routes. It has low signaling and processing overhead, in addition to their fast response to dynamics of network topology [16]. We assume that the “void” problem in geographic unicast routing is solved, because the malfunction satellite will be replaced by the on-orbit backup satellites (about seven on-orbit backup satellites in Iridium system [17]).

In this paper, we make the following assumptions:

- (1) Each satellite and each mobile node knows its own position. Such information can be obtained through equipped GPS receiver for finding their location.
- (2) Each satellite knows the positions of its direct-connected satellites. This information can be piggybacked by periodical *Hello* messages (status advertisement message) between neighboring satellites.
- (3) We employ soft satellite handover method. In this type handover method, link that connects to the old satellite will not be released until the next connection to the new satellite is established.

A. VCMulticast Overview

When a source mobile node wants to send a multicast datagram to a multicast group, it will send the datagram to its connected satellite. Once the satellite receives this datagram, it will generate next hops for two types of forward: the inter-plane forward and the intra-plane forward. The aim of first forward type is to send datagram to the group members that connect satellites in other orbit planes. The aim of second forward type is to send datagram to the group members that connect satellites in the same plane as current satellite.

To generate the next hops for inter-plane forward, current satellite creates two “virtual centers” (VC_L and VC_R) of group members to access satellites in left and right “half network surface”. VC is the position center of group members and the detailed definition will be given in Section III.B. As shown in Fig. 3, there are 12 satellites in orbit plane 5 and the current satellite number is 6, then the “left” half surface is composed by the satellites in plane number 1, 2, 3 and 4 while the “right” half surface is composed by the satellites in plane number 6, 7, 8, 9 and 10. Then it employs the geographic based unicast routing to generate next hop of multicast datagram to VC_L and VC_R if it exists.

Once the next hop of inter-plane forward is generated, the current satellite will search the list of group member information from current plane to find whether the current node, *up* direction satellites (intra-plane neighbor satellite in the same direction of movement of current satellite), *down* direction satellites (intra-plane neighbor satellite in the inverse direction of movement of current satellite) have other accessing group members or not. This result is then used to generate the next hops for intra-plane forward.

Next, the current satellite replicates the datagram for each next hop that need forward multicast datagram. In order to do that, the current satellite appends a header consisting of a list of “succession planes” for inter-plane forward datagram. If we consider orbit plane 5 as the current plane and satellite 6 as the current satellite, as depicted in Fig. 3, then the left “succession planes” are composed by plane 1, 2, 3 and 4, and the right “succession planes” are composed by plane 6, 7, 8, 9 and 10. If next hop of intra-plane and inter-plane forwarding is to the same satellite, then only one datagram needs to be replicated. To inform the other satellites that the datagram the forward type (inter-plan or intra-plan) and the successive plane direction (left or right) of the network surface, we append two flags to the datagram header: the forward flag (F_{flag}) and the inter-plane direction flag (D_{flag}) (shown in Fig. 5). The F_{flag} indicates forward type of datagram, and the D_{flag} indicates which direction of successive planes that datagram will be forward to. A special case is that when the F_{flag} is set to be inter-plane but D_{flag} is not set ($D_{flag}=NULL$). This means that the next hops of the two inter-plan forward directions are the same, and then the D_{flag} flag has to be determined in the next hop. The datagrams transmitting procedure is summarized by the pseudo-code in Algorithm 1(See Appendix A).

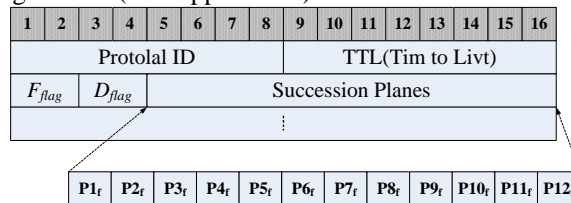


Figure 5. Packet header of the VCMulticast scheme

When a satellite receives a multicast datagram from other satellite, it firstly retrieves the F_{flag} and D_{flag} values from the datagram header. The current satellite node will search the group member information list to find whether the current node, *up* direction satellites, and *down* direction satellites have other accessed group members or not. Then satellite generates next hops for intra-plane multicast datagram. If the intra-direction of next-hop satellite is the same as the *sender* that sends datagram to the current satellite node, the direction of next hop is then set to be NULL. If the forward type is inter-plane, the current satellite generates the specific VC from the value D_{flag} and plane list in the datagram header. Then the satellite will employ the geographic based unicast routing to generate next hop for the inter-plane multicast datagram. After these processes, the current satellite node replicates the datagram for each next hop that need the forward multicast datagram. For the inter-plane forward, the current satellite node appends the remaining succession plane list into the header of inter-plane forward datagram. If the next hop of the intra-plane and the inter-plane are the same, then only one datagram will be replicated for that next hop and the F_{flag} of datagram will be set to be inter-plane type. The procedure execution after receiving datagrams is summarized by the pseudo-code in Algorithm 2 (See Appendix B).

Fig. 6 gives an example of how VCMulticast scheme is executed. When the satellite S_1 receives a multicast datagram from the source node M_S , there are five group members distributed in the service areas of the satellite network. The left planes contain only one group member, thus a datagram is sent to transmitted to the left virtual center of group members (dotted circle with label VC_L in the figure, this point is calculate by the directly connected satellite of source node) without been duplicated. The right planes have three group members accessed, thus an inter-plane datagram is transmitted to the right virtual center of group members (dotted circle with label VC_R in the figure, this point is calculate by the directly connected satellite of source node). For the routing path from M_S to M_2 , an intra-plane datagram is transmitted through the intra-plane $S_1-S_2-S_3-S_4$. When the datagram reaches a node with different directions between the next hops of inter-plane and intra-plane, the node will split off datagram to each of the multicast direction accordingly.

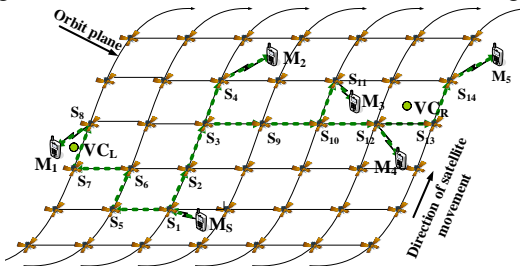


Figure 6. A datagram is send from node M_S to all group members

B. Latitude Center (LC) and Virtual Center (VC)

LC is the “gravity center” of the latitudes of group members that connect to the satellites in the same orbit plane. There are two type LC s in each plane that can be calculate by

$$LC = \left(\sum_{i=1}^{N_G} GM(i).lat / N_G \right) \quad (1)$$

where N_G is the number of group members and GM represent group member. We call the first type LC as *up LC* (LC_u) and it is calculate from the group members situated in the normal access area that access satellites are in the ascending satellite network surface or in the two special areas. We call the second type of LC as *down LC* (LC_d) and it is calculate from group members of situated in normal access area that access satellites are in the descending surface or in the two special areas. As shown in Fig. 7, the LC_u is generated from the latitude of node 1-5 (in the normal area) and node 7-9, 14 and 15 (in the special areas). The LC_d is generated from the latitude of the node 10-13 (in normal area) and the node 7-9, 14 and 15 (in special areas).

To save the network bandwidth, only one surface is used to send and forward datagram for one multicast group. The selection of using which surface can be arbitrary predefined or decided by the source node of multicast group.

VC is a virtual geographic point of group members that can be calculated by

$$VC_{lat} = \sum_{i=1}^{N_p} LC(i) / N_p \quad (2)$$

$$VC_{lon} = (CS_{lon} + A_{I-p} \times N_p) / N_p \quad (3)$$

where N_p and CS_{lon} are the number of left or right direction successive planes and longitude of current satellite node. A_{I-p} is the spacing between two neighbor planes.

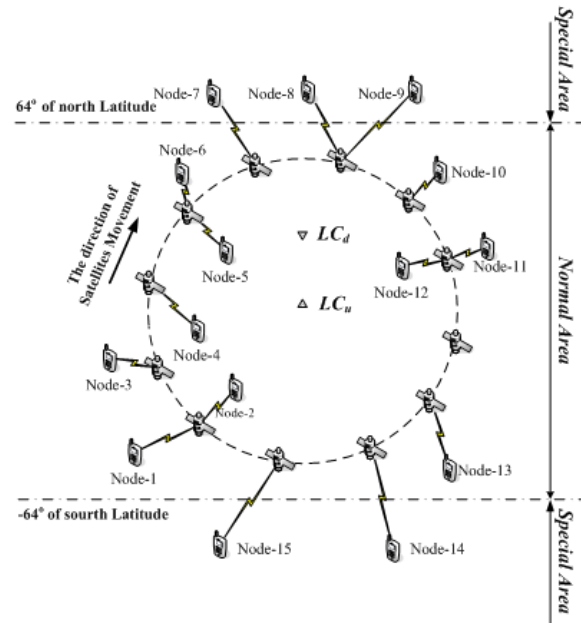


Figure 7. LC s of an orbit plane

C. Group Management

a) Group member join

Group member join happens under two conditions: (1) when an incoming group member wants to join a multicast session, it sends a *Join_Request* message to the session hosting satellite; or (2) when the current hosting satellite is going out of sight, the group member has to send a *Join_Request* message to the new hosting satellite. After receiving the *Join_Request* message, the satellite replies a *Join_Confirm* message to the group member, updates the profile list of group member (L_{GU}) of current satellite, and advertises the profile of this new member to the rest of satellites which is in the same plane by sending *Join_Information* message to *up* and *down* satellite neighbors.

The other satellites in the same plane will update its L_{GU} after receiving the *Join_Information* message, and forward this message to others until all the satellites in the plane have received this message.

b) Group member leave

Group member leave happens under two conditions: (1) when a group member wants to leave a multicast session, it sends a *Leave_Request* message to the session hosting satellite; (2) when the current hosting satellite is going out of sight, the group member has to send a *Leave_Request* message to the current hosting satellite.

The satellite will reply a *Leave_Confirm* message to the group user after receiving the *Leave_Request* message.

From the relative invariant property 3, we can get that the inter-satellite handover takes place either in one plane or between two neighboring planes. To reduce the process and signaling overhead, the satellite does not immediately clear the member profile in L_{GU} after sending the *Leave_Confirm* message. If the current and new access satellite is in the same plane, only the *Join_Information* message will be sent to the rest satellites of the current plane, and the access satellite prosperity of the profile of this member in L_{GU} will be modified to be the new access satellite. If current satellite and new satellite is not in the same plane, the *Leave_Information* message will be sent, and at the same time the *member status flag* and a *timer* of member profile in L_{GU} are set. The *Member status flag* indicates the group member is left current plane or not, and the *timer* stores the deadline when the profile of this group member is to be removed. The other satellites in the same plane will set the *member status flag* and the *timer* for the group member profile after receiving the *Leave_Information* message.

In this way, when a group user leaves its current access satellite, the LCs of the current plane does not change immediately. Since the *Member status flag* property of L_{GU} item is irrelevant to (1), it is not necessary to immediately recompute the LCs after the group member leaves. This is especially important to the *special area*, in which group user must endure ceaselessly inter-plane satellite handover.

D. Cluster and Cluster Header

We assume that the satellites in one plane form a cluster. Each cluster selects a satellite as the cluster header. The cluster header is mainly to flood the LCs of current plane to the satellites in the rest planes via inter-plane ISLs.

The cluster header recomputes the latitude centers of plane under the following two conditions:

First, if the *member status flag* of a group member indicates that the member is left and the *timer* is expired and the information entry of this member is cleared.

Second, if a group member connects to a satellite plane for the first time (in other words, the group member has never connected with any satellite in the plane before).

To reduce the signaling and process overhead, the advertisement of LCs of the current plane is neither periodical nor imitatively after LC changed. The cluster header will generate LC update messages and send it to the rest satellites in other planes only when the change of LC satisfies the following condition:

$$\begin{cases} |LC_{new} - LC_{old}| \geq \alpha & (LC_{old} \neq NULL) \\ LC_{new} \neq NULL & (LC_{old} = NULL) \end{cases} \quad (4)$$

where α is coverage of satellite constellation related threshold, LC_{new} and LC_{old} are new generated latitude center and last advertised latitude center. For NeLS constellation, α ranges from 0° to 140° . The value of α affects the degree of multicast tree degenerates. A little

value of α will decrease the degree of tree degenerates but will lead to a large value of LC advertisement. While a large value will increase degree of tree degenerates but it will lead to little values of LC advertisement. If α equals 0° , it means that whenever the LC changes, the advertisement will be sent. If α equals 140° , the advertisement of LC will never change.

IV. PERFORMANCE EVALUATION

In order to assess the performance of VCMulticast scheme, we perform three groups of experiment.

(1) We investigate the end-to-end propagation delay between the source and each multicast group member for the unicast-connection case, and for the multicast trees case which is generated by our routing algorithm. We also inspect the bandwidth savings by using our multicast scheme instead of individual unicast connections. Three user topology distributions are considered in our experiments, namely, uniform-1, uniform-2 and non-uniform multicast group member distributions. The details of these distributions will be given in the next subsection.

(2) We analyze the effect of the dynamic multicast group membership on the multicast tree length.

(3) We compare our multicast scheme with MOSPF [18], and CBT [19] schemes. The experiments cover the aforementioned three group member distributions as well.

In all experiments, NeLS constellation with 120 satellites is utilized. These 120 satellites are distributed uniformly among 10 planes. Table I lists the parameters of the NeLS constellation. We use the Satellite Tool Kit (STK) simulator to generate satellite orbit data, and use the OPNET [20] simulator to construct satellite constellation network. All experiments presented in this section simulate a 24-hour operation and cover a multicast group size ranging from 5 to 100.

To simplify the discussion, we set threshold value α to be 1° . And the cluster header satellite in our discussion is assumed to be situated in the ascending surface, and its latitude is smaller than 0° and the latitude of its *up* direction satellite is greater than 0° .

TABLE I. ORBIT PARAMETERS FOR NELLS CONSTELLATION

Orbital parameters	Value
Number of orbital planes	10
Number of satellites per orbital plane	12
Orbital altitude	1200 km
Orbital inclination	55°
Eccentricity	0 (circular orbit)
Orbital period	109 min 25 s
Difference angle of ascending node between adjacent orbit planes	36°
Inter-orbital phasing	3°
Minimum elevation angle from user terminal (single-satellite coverage)	20°
Minimum elevation angle from user terminal (dual-satellite coverage)	13°
Intra-plane ISL distance	3922 km
Inter-plane ISL distance	3062–4909 km

A. Generation of Multicast Groups

The performance of our new multicast routing scheme is assessed by three different group member distributions,

namely, two uniform cases and a non-uniform case. These three distributions are defined as follows:

uniform-1: the sender as well as the group members are distributed randomly in terrestrial areas in coverage of satellite network;

uniform-2: the sender as well as the group members are distributed randomly in the coverage of satellite network.

non-uniform: the sender as well as the group members are distributed according to the global traffic demands.

The two uniform distributions are utilized to evaluate the efficiency and fairness of the proposed scheme for the scenarios that users are distributed in the terrestrial area and all users (terrestrial and marine), respectively.

The non-uniform case is utilized to evaluate the efficiency of our proposed scheme when encountering high member density. The non-uniform distribution we utilize in this work is based on a real traffic demand model presented in [21]. In this model, the plan sphere is divided into 288 cells with 24 bands along the longitude and 12 along the latitude. The traffic expectation of each cell low-to-high is indicated by an intensity level ranging from 0 to 8 (shown in Fig. 8). The probability p that a group member belongs to a specific cell i then can be defined as:

$$p(i) = \frac{C_{IL}(i)}{\sum_{i=1}^{288} C_{IL}(i)} \quad (5)$$

where C_{IL} is the intensity level of cell.

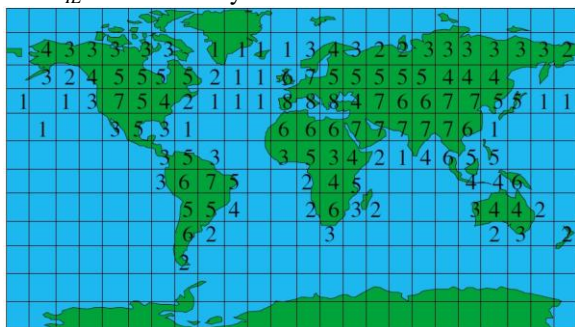


Figure 8. World's traffic demand model matrix [21]

B. Performance of VCMulticast Scheme

VCMulticast does not maintain a real multicast tree. The “multicast tree” in VCMulticast is generated by combining all routing paths from the source node to all the destination nodes together. Each satellite use geographic based unicast algorithm to find next hop towards the VCs of the rest group members that access satellites in succession planes. Thus, the resulting paths of geographic based unicast algorithm do not necessarily have the minimum propagation delays. In the first experiment, we compare the propagation delays on the multicast tree created by our multicast algorithm with the minimum propagation delay paths determined by the Dijkstra's shortest path algorithm [22] (DJK) from source to each destination. The results are evaluated by

percentage increase in propagation delays and shown in Fig. 9.

As shown in Fig. 9, for all multicast group member distributions, the differences between our proposal and the minimum propagation delay path range from 19 to 29.1 percent. And the deviation from the minimum propagation delay path increases as the group size increases up to 10. After this point, as the group size increases, the percentage increment in the end-to-end propagation delay starts decreasing. When the group size is small, the ratio of shared hops is smaller. Therefore, the end-to-end propagation delay is larger than the one of the minimum propagation delay path. When there are more group members, the ratio of shared hops begin to increase and the path-length deviation begin to decrease. In addition, the performance of uniform-2 distribution is better than the other two distributions. The reason is that the group member density of uniform-2 is uniformly distributed in all service areas of satellite network (including terrestrial and marine areas), thus the ratio of sharing the hops by different source to destination pairs is smaller than the other two distributions. The uniform-1 distribution and non-uniform distribution might lead to a situation that the group members aggregate in certain areas, thus the ratio of shared hops by different source to destination pairs become higher resulting in a bigger end-to-end communication delay difference. Furthermore, as depicted in Fig. 9, the number of group members has little effect on the end-to-end communication delay differences. Because the tree generated by our protocol is highly depended on the LCs of each plane, which is sensitive to the position of group member center. The group member center is related to the member distribution, thus the end-to-end communication delay difference has a dependence on the group member distribution.

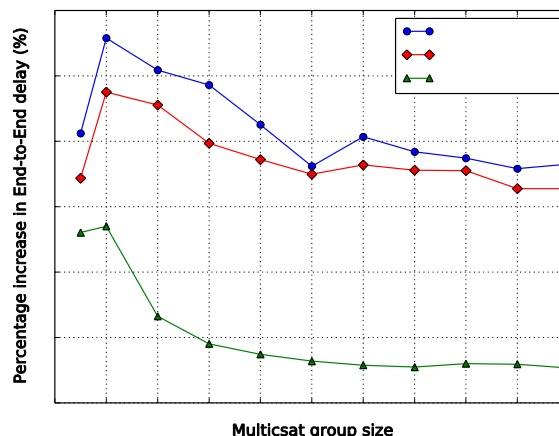


Figure 9. Percentage increment in propagation delay

In the second experiment, we evaluate the bandwidth saving for two scenarios by using our multicast scheme and sending independent datagrams to each destination, respectively. For comparison, we calculate the length of the multicast trees for the former scenario and the sum of the individual short path lengths to all destinations for the later scenario. The results are shown in Fig. 10. The

multicast/unicast path length ratio can be interpreted as follows. If the ratio is 1, then the sum of the unicast path lengths to all destinations and the length of the multicast tree are the same, i.e., there is no link sharing at all. The closer this ratio approaches to zero, the higher probability the link will be shared.

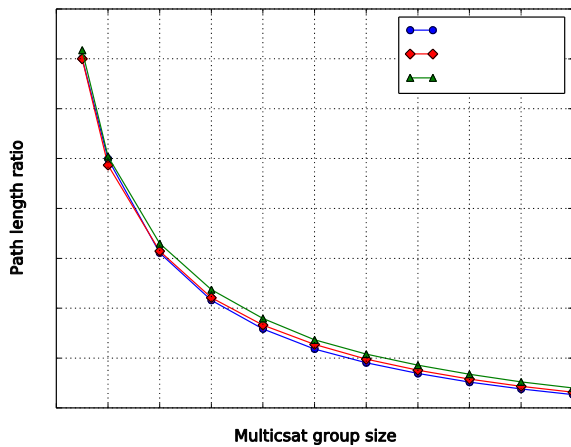


Figure 10. Multicast/unicast path length ratio

Fig. 10 indicates that link sharing increases as the group size grows for all types of group member distributions. The more group members there are, the higher probability there is that an outgoing link is shared by more than a single destination. For group size ranging from 5 to 40, the curves decline significantly. This means that adding a new member increases link sharing when there are few group members. Since link sharing depends on the location of the group members, it grows as the group member density grows. However, the difference in link sharing between different group member distributions decreases as the multicast group size increases.

C. Effect of Dynamic Group Membership

The addition and removal of multicast group members causes deviations from the original structure of the multicast tree. The changes to the group membership result in the changing of LC s of related orbit plane. As described in Section III.D, the cluster header initiates the LC update procedure when the deviation of current LC_{new} and last advertised LC_{old} exceeds the threshold value α .

In order to show the effects on propagation delay of the rest group members in member addition and removal, we perform a set of experiments with all the three aforementioned member distributions. The exact simulation process is: first, we generate a multicast group member distribution; then, we increase the number of addition and removal of multicast group members from 1 to 10 and record the changing of length of the tree. This tree is generated by combining all routing paths from the source node to all the destination nodes that don't dynamically join or leave. The length difference between the trees before and after the group membership changing is expressed as the percentage deviation with respect to the length of tree before addition and removal of multicast group members. The experiment is repeated for

group sizes 5, 10, 20, 30, 40 and 50. For each group size, the average length difference of the three group member distributions is calculated and the results are depicted in Fig. 11.

As shown in Fig. 11, as the group size grows from 10 to 50, the tree deviates little from its original structure. The addition and removal of the tree membership in small-size multicast groups has bigger influence than in large size multicast groups. For example, the increase of the tree length is only about 0.12% on average for a group of 50 members. The only exception here is for a group with only 5 members, the maximum increase of the tree length is about 2.1%. The reason is that any group membership change (addition or removal) is significant enough to the member distribution due to the small group size. As mentioned in Section IV.B, the member distribution is critical to the multicast tree construction in our proposal. Another result we can get is that the deviation is not greatly affected by the number of the addition and removal of group members. This is because we use the "gravity center" VC of group members to route multicast datagram. If the "gravity center" is not changed significantly, there will be no great change in the tree length.

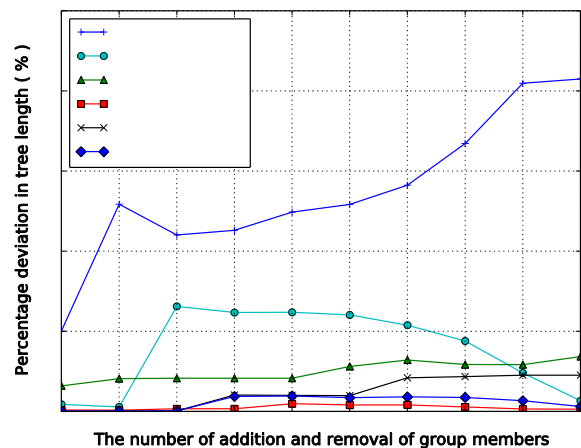


Figure 11. Effect of dynamic group membership

D. Comparisons with Other Multicast Schemes

As outlined in Section I, few of the existing multicast routing algorithms can be directly implemented for LEO satellite network. There are two basic sets of multicast schemes: Source-based tree and Core-based tree multicast schemes [14]. In this section, we compare our multicast scheme with MOSPF, and a version of CBT schemes. MOSPF is a Source-based tree protocol and employ a shortest-path tree. For the Core-based scheme, a CBT version in Ref. [9] is utilized for comparison. To the best knowledge of the authors, Ref. [9] is the only multicast algorithm that does not hide the dynamic property of satellite network topology.

a) MOSPF: MOSPF multicasts datagrams over the shortest-path tree within an Autonomous System (AS). MOSPF extends OSPF [23] by adding a new type of link-state advertisement (LSA), called the group membership LSA. In MOSPF, a router uses IGMP [24] to keep track

of group membership information on its attached network, and distributes this information by flooding the group membership LSA throughout the AS. When a router receives a multicast datagram, it computes a shortest-path tree rooted at the source of the datagram and forwards the datagram accordingly.

In our simulations, for MOSPF, we consider the entire satellite network as a single OSPF area [23] and we compute the shortest path tree for each multicast datagram independently due to the ISL and the mapping between satellite and connected group member keep on changing. The bandwidth demands of our multicast scheme and MOSPF are compared by the tree-length difference of these two types for different multicast group member distributions.

MOSPF tries to establish shortest paths between source and destinations, and links are shared only if they belong to multiple shortest paths. In VCMulticast, on the other hand, the datagram will be duplicated for each direction only when the next-hops of inter-plane and intra-plane are different. In Fig. 12, we demonstrate the bandwidth savings by our scheme relative to MOSPF. The results presented here show the tree length differences in percentages. When the group size is below 20, the hop sharing ratios of our scheme and MOSPF are both small, and the performance of MOSPF scheme is slightly better than our scheme. As group size increases, the difference of these two schemes decreases due to the increasing hop-sharing ratio. However, MOSPF is not feasible in real implementations because it will generate a great process overhead (timing complexity is $O(N^2)$, where N is the number of satellite). Note our scheme employment geographic based unicast scheme requires no effort on the establishment and maintenance of multicast tree and routes.

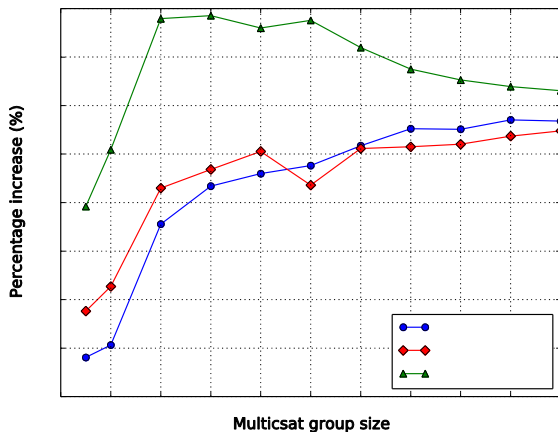


Figure 12. Percentage increment in tree length of MOSPF relative to our scheme

In the case of group member distribution “uniform-2”, when the group size ranges from 20 to 100, the tree-length generated by MOSPF exceeds the tree-length of our scheme at rates from 3.7% to 0.6%. The reason is that as discussed in Section IV.B, for this distribution, the ratio of hop sharing of uniform-2 is smaller than the other two distributions. Therefore the tree-length generated by our scheme is smaller than the one generated by MOSPF.

b) Core-Based Tree Scheme: In this section, we compare our multicast routing scheme with a CBT algorithm [9]. In the CBT scheme, the multicast datagrams are sent to a designated node called the “core”, which relays these datagrams to other multicast group members. The multicast datagrams are routed from the source to the core as unicast datagrams. The multicast tree from core to all group members is accomplished via shortest-path tree. The selection of the core is an important issue and affects the performance of the CBT scheme. In Ref [9], a simple vector algorithm is suggested, which selects the core based on aggregating the locations of group members. Although this strategy can select a core satellite which will be nearest to the center of group members in the network, it may perform poorly in many cases since satellites moves very fast relative to the earth surface. The core satellite of network is ceaselessly changing and it will incur considerable network overhead of signaling and processing. In our experiments, we compute the core for each multicast datagram independently (timing complexity is $O(N)$), generate the shortest path from the source to the core node and generate the shortest-path tree from core to all group members (as MOSPF the timing complexity is $O(N^2)$). The procedure for core calculation is as follows [9]:

- (1) Sum all the vectors for the group to get a resultant vector

$$c = [\sum_k x_k, \sum_k y_k, \sum_k z_k] \tag{6}$$

where k is number of group members and (x, y, z) is geocentric coordinate position of group member.

- (2) Convert the direction of c into spherical coordinates (lat_c, lon_c) to determine the position of the core satellite. Search a satellite that currently nearest to this position and nominate this satellite as the core node of the multicast tree.

In our simulations, we compare the lengths of the trees generated by our algorithm and the CBT scheme. We also compare the propagation delays of the datagrams routed on the minimum propagation delay paths and trees generated by the CBT scheme. These tests are performed for uniform and non-uniform member distributions.

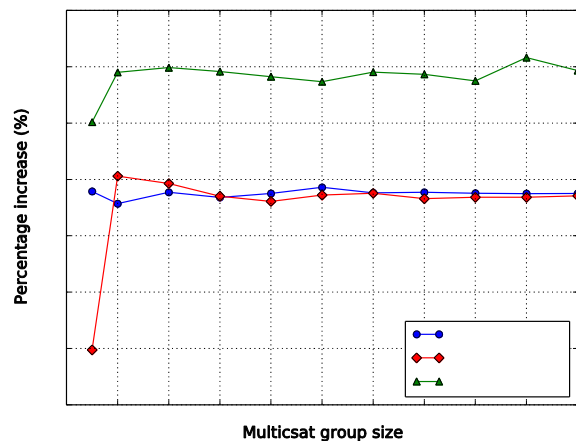


Figure 13. Percentage increment in tree length for CBT

In Fig. 13, the percentage increase in tree length of our multicast scheme relative to the CBT scheme is depicted. As discussed in Section IV.B, the uniform-1 distribution and non-uniform distribution might lead to a situation that the group members aggregate in certain areas. And the unicast connection between the source and core play an important role in the total tree length increment, thus the tree length is almost the same with the tree created by our multicast scheme. Because the CBT scheme employs a shortest path tree from the core to each destination, this will introduce a great processing overhead. As a result, the performance of CBT is just slightly better than VCMulticast under the uniform-1 distribution and non-uniform distribution. However, for the uniform-2 distribution, the tree length of CBT scheme is relatively larger, which exceeds the tree length of our scheme at rates from 5% to 10%. The reason is that for the uniform-2 distribution, the hop sharing ratio is smaller than the ratios in the other two distributions. And the unicast connection between the source and core play an important role in the total tree length increment, therefore, tree-length generated by our scheme is smaller than the one generated by CBT. Furthermore, if we consider a hybrid complex network which is consisted of an equal-weighted combination of these three user distribution, then our scheme outperforms the CBT scheme since the average tree-length of these three distributions in our scheme is shorter than the one in the CBT case.

We also present experimental results to show the delay performance of the CBT protocol. In Fig. 14, the percentage difference of the propagation delay of the CBTs and the minimum propagation delay paths between the source and each destination are depicted. For all multicast group member distributions, the minimum propagation delay path is 62% to 105% longer than the minimum propagation delay paths. Note that the increase in propagation delay of our scheme is from 19% to 29.1% for the same range of group size, which is shown in Fig. 9. our proposed method reduces more than 50% propagation delay.

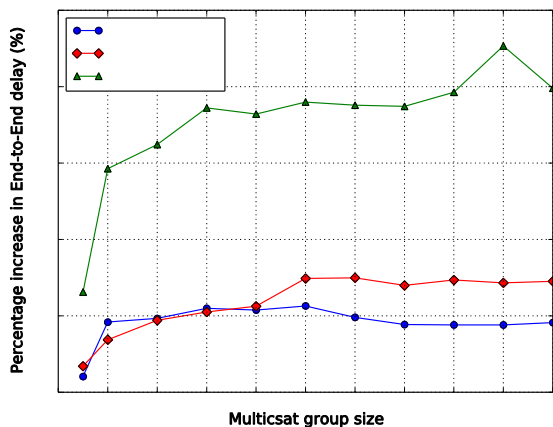


Figure 14. Percentage increment in propagation delay for CBT

V. CONCLUSIONS

In this work, we proposed a new multicast routing scheme for datagram traffic in LEO satellite constellation

networks. The new scheme is a source-based like algorithm, but no real multicast tree is maintained. Thus, it does not undergo flood and prune process or employ DJK to construct the multicast trees. Each satellite node in the satellite constellation network only needs to maintain a small amount of information to forward multicast datagram. The simulation results show that the multicast trees provide delays exceeding the average minimum propagation delay by at most 29.1 %. Multicast trees are multiple times shorter than the sum of unicast paths. The performance of MOSPF and CBT might be slightly better than our scheme under certain user distributions (i.e., non-uniform and uniform-1), however, the resource consumption of our scheme is less than the counterparts for all circumstances. We also present simulation results regarding the dynamic group membership and show that the dynamic group membership scales well with the increasing multicast group size.

APPENDIX A Algorithm 1

```

Algorithm 1. Multicast datagram Send
Require: datagram received from in-port MAC layer
Ensure: datagram inserted to out-port MAC queue
1: Generate VCL and VCR
2: Employ geographic routing to generate inter-plane next-hop set A
according to VCL and VCR
3: Get group member list N from current orbit plane
4: if N is not NULL
5:   Generate intra-plane next-hop set B by searching N
6: end if
7: for next-hop h in A do
8:   Duplicate a new datagram P
9:   if h is the same for VCL and VCR
10:    Dflag ← NULL
11:   else
12:     if h towards VCL
13:       Dflag ← left
14:     else
15:       Dflag ← right
16:     end if
17:     Append succession plane list to P
18:   end if
19:   Insert P to the out-port MAC queue
20: end for
21: for next-hop h in B but not in A do
22:   if h is a mobile node connected to the current
satellite
23:     Int.sat_to_land ++;
24:   else
25:     Duplicate a new datagram P
26:     Fflag ← intra-plane
27:     Insert P to the out-port MAC queue
28:   end if
29: end for
30: if Int.sat_to_land > 0
31: Duplicate a new datagram P
32: Insert P to the MAC queue of satellite to land port
33: end if
    
```

APPENDIX B Algorithm 2

```

Algorithm 2. Multicast Datagram Forward
Require: datagram received from in-port MAC layer
Ensure: datagram inserted to out-port MAC queue
1: Get Fflag and Dflag from datagram header
2: Get group list N from current orbit plane
3: if N is not NULL
4:   Generate intra-plane next-hop set A by searching N
5: end if
6: if Fflag = inter-plane
7:   if Dflag = NULL
8:     Generate VCL and VCR
    
```



```

9:   Employ geographic routing to generate inter-plane next-hop
set B according to VCL and VCR
10:  else
11:   Generate VCD of succession planes indicated by Dflag
12:   Employ geographic routing to generate inter-plane next-hop
HC according to VCD
13:  end if
14: end if
15: for next-hop h in B and HC do
16:   Duplicate a new datagram P
17:   Fflag ← inter-plane
18:   if h is the same for VCL and VCR
19:     Dflag ← NULL
20:   else
21:     if h towards VCL
22:       Dflag ← left
23:     else
24:       Dflag ← right
25:     end if
26:     Append remain succession plane list to P
27:   end if
28:   Insert P to the out-port MAC queue
29: end for
30: for next-hop h in A but not in B and HC do
31:   if h is a mobile node connected to the current
satellite
32:     Int.sat_to_land ++;
33:   else
34:     Duplicate a new datagram P
35:     Fflag ← intra-plane
36:     Insert P to the out-port MAC queue
37:   end if
38: end for
39: if Int.sat_to_land > 0
40: Duplicate a new datagram P
41: Insert P to the MAC queue of satellite to land port
42: end if

```

ACKNOWLEDGMENT

The authors are with the College of Computer, National University of Defense Technology, Changsha, China.

This research was supported in part by the CHINA National Science Foundation grants No. 61070199, 61103182, 61202488 and 61103194; the National High Technology Research and Development Program of China (863 Program) No. 2011AA01A103, 2012AA01A506 and 2013AA013505; the Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province: "network technology"; Changjiang Scholars and Innovative Research Team in University (No.IRT 1012) and Hunan Province Natural Science Foundation of China (11JJ7003).

REFERENCES

- [1] E. Ekici, I. F. Akyildiz, and M. D. Bender, "A Multicast Routing Algorithm for LEO Satellite IP Networks," *IEEE/ACM Trans. Net.*, vol. 10, no. 2, pp. 183–92, Apr. 2002.
- [2] M. Werner, "A dynamic routing concept for ATM-based satellite personal communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1636–1648, 1997.
- [3] R. Mauger and C. Rosenberg, "QoS guarantees for multimedia services on a TDMA-based satellite network," *IEEE Communications Magazine*, vol. 35, no. 7, pp. 56 – 65, 1997.
- [4] L. Kai, C. Lianzhen, and Z. Jun, Efficient, "Multicast Routing for LEO Satellite IP Networks," *IEEE 70th Vehicular Technology Conference Fall (VTC 2009-Fall)*, 2009.
- [5] D. N. Yang and W. Liao, "On Multicast Routing Using Rectilinear Steiner Trees for LEO Satellite Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2560–2569, 2008.
- [6] C. Lianzhen, Z. Jun, L. Kai, and S. Xuegui. "A multiple-cores shared-tree multicast routing algorithm for LEO satellite IP networks," *IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications (MAPE 2005)*, 2005.
- [7] T. H. Henderson and R. H. Katz, "On Distributed and Geographic-Based Datagram Routing for LEO Satellite Networks," *Proc. GLOBECOM, San Francisco, CA, Dec. 2000*.
- [8] R. E. Sheriff and Y. F. Hu, "Mobile Satellite Systems," in *Mobile Satellite Communication Networks*, ed. University of Bradford, UK: John Wiley & Sons, Ltd, pp. 43–83, 2001.
- [9] L. Wood, "Internetworking with satellite constellations," *Centre for Communication Systems Research, School of Electronics, Computing and Mathematics, Guildford, United Kingdom, University of Surrey. Doctor of Philosophy: 230*. 2001.
- [10] [http://www. agi. com/products/by-product-type/applications/stk/](http://www.agi.com/products/by-product-type/applications/stk/).
- [11] W. Markus, F. Jochen, F. Wauquiez, and G. Maral. "Topological design, routing and capacity dimensioning for ISL networks in broadband LEO satellite systems," *International Journal of Satellite Communications*, vol. 19, no. 6, pp. 499–527, 2001.
- [12] R. Suzuki and Y. Yasuda, "Study on ISL network structure in LEO satellite communication systems," *Acta Astronautica*, vol. 61, pp. 648–658, 2007.
- [13] R. Suzuki, I. Nishiyama, S. Motoyoshi, E. Morikawa, Y. Yasuda, "Current status of NeLS project: R&D of global multimedia mobile satellite communications," in: *The 20th International Communications Satellite Systems Conference and Exhibit, AIAA-2002-993, Montreal, Canada, May 12–15, 2002*.
- [14] L. Wood, A. Clerget, I. Andrikopoulos, G. Pavlou, and W. Dabbous, "IP routing issues in satellite constellation networks," *International Journal of Satellite Communications*, vol. 19, pp. 69–92, 2001.
- [15] E. Kranakis, H. Singh, and J. Urrutia, "Compass routing on geometric networks," in *Proc. 11th Canadian Conf. Computational Geometry, Vancouver, BC, Canada, pp. 51–54, Aug. 1999*.
- [16] T. Camp, "Location Information Services in Mobile Ad Hoc Networks," in *Handbook of Algorithms for Wireless Networking and Mobile Computing. University of Ottawa: Taylor & Francis Group, LLC, pp. 319–341 2006*.
- [17] [http://en. wikipedia. org/wiki/ Iridium_satellite_constellation. htm](http://en.wikipedia.org/wiki/Iridium_satellite_constellation.htm).
- [18] J. Moy, "Multicast Extensions to OSPF", *RFC 1584*, March 1994.
- [19] A. Ballardie, "Core Based Trees (CBT version 2) Multicast Routing", *RFC 2189*, Sep 1997.
- [20] [http://www. opnet. com/](http://www.opnet.com/).
- [21] A. Ferreira, J. Galtier, and P. Penna. "Topological design, routing and hand-over in satellite networks". In I. Stojmenovic, editor, *Handbook of Wireless Networks and Mobile Computing*, pp. 473–507. John Wiley and Sons, 2002.

- [22] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, "Introduction to Algorithms," *MIT Press*, pp. 527-532, 1990.
- [23] J. May, "OSPF Version 2," *RFC 2328*, Apr 1998.
- [24] S. Deering, "Host extensions for IP multicasting," *RFC 1112*, Aug. 1989.

Bandwidth Consumption Efficiency Using Collective Rejoin in Hierarchical Peer-To-Peer

Sri Wahjuni, A.A.Putri Ratna, and Kalamullah Ramli

Department of Electrical Engineering, Faculty of Engineering, University of Indonesia, Depok, Indonesia

Email: {sri.wahjuni, anak.agung, kalamullah.ramli}@ui.ac.id

Abstract—Having the advantage of scalability and efficiency in performing a successful lookup query, a hierarchically structured P2P is a promising architecture for heterogeneous networks. However, in addition to potentially decrease the system performance, the presence of a superpeer failure event also forces normal peers under its influence to disconnect from the system. Optimally, the disconnected normal peers should perform rejoin, otherwise they will lose the granted access to the required service. Currently, the most common rejoin algorithm is based on the flat Chord maintenance algorithm, which is an individual rejoin. Addressing the high bandwidth consumption in individual rejoin, we propose a new approach, termed the collective rejoin algorithm. The analytical results, as well as simulation outputs, show that the rejoin process using the collective rejoin algorithm produces less traffic load than the individual rejoin algorithm.

Index Terms—Collective Rejoin; Churn; Heterogeneous Network; Superpeer Failure

I. INTRODUCTION

According to Kangasharju [1] the P2P architectures are categorized as unstructured and structured. From the unstructured architecture, the most well-known centralized architecture is Napster [2], while the one from fully distributed architecture is Gnutella [3]. The compromised architecture between the centralized and fully distributed structures is known as a hybrid architecture, such as Kazaa [4]. The natural benefits of the unstructured P2P architecture are simplicity and stability even though they suffer from a high traffic load as the impact of the use of flooding search method. The later generation of P2P is structured architecture, which is based on a distributed hash table (DHT) for indexing both the peer and the shared-object. Some examples are CAN [5], Chord [6], Pastry [7], and Tapestry [8]. The use of DHT addresses the efficiency and scalability problems, which are the main problems in unstructured architectures. However, structured P2P suffers from the dynamics of peers, which is a natural characteristic of heterogeneous networks.

The scalability of Chord is related to its indexes distribution approach, that shares the indexes among the nodes. The maximum nodes needed to retrieve to find the object is $O(\log n)$, for an overlay network formed by n peers. Chord implements consistent hashing to define an

identifier of a node and the object key of the shared objects/services. The identifier of a node is obtained from the hash value of the node identity (using a function such as SHA-1), and the key identifier is obtained from the object identity (such as the file's name). The identifiers are ordered in a modulo 2^m ring size for m -bit identifier. The key k is assigned to the first node that has the same value or follows the identifier k in the identifier space. This node is the successor node of key k , called successor (k). Suppose the identifiers are represented in a ring numbered of 0 to 2^m-1 , then the successor(k) is the first node (succeed k) in the clockwise direction. Each node has a finger table and a successor list.

As the device's capability in accessing networked application grows, the hierarchical P2P architecture [9-13] is a suitable architecture for handling the heterogeneity of the peer participants. Garces-Erice et al. [9] proposed the generic framework of the two-level hierarchical architecture for DHT-based P2P. Higher capability devices (the ones with longer connection times) form the first level of the hierarchy, and connect to each other in a ring-like structure (such as Chord), while the lower ones are grouped into clusters that form the second level. The peers in clusters communicate with their upper level peers through various schemes and communicate with other peers in other clusters through the gateway peer. The superpeer concept, where a peer with a higher capacity handles more responsibilities than others, was used in the architecture proposed by Montresor [10]. Pandey et al. [11] implemented the Chord-based structure for the top layer of the hierarchical P2P as well as for the lower layer, while Peng et al. [12] implemented unstructured architecture for the lower layer. A formal analysis to compare various lower level peers organizations as proposed by Garces-Erice et al. [9] was presented by Zoels et al. [13]. The authors concluded that a simple star-like design is optimal for two-level hierarchical P2P. Based on this star-like design [13], the works of [14] and [15] are performed on finding the optimal proportion of the superpeers and the regular peers to reach a better performance.

Although the hierarchical P2P architecture accommodates the heterogeneity of networks, the diversity of the peers capabilities may cause a higher probability of churn. Churn is a dynamic condition of the P2P network when a node joins or leaves the system [16]. The higher frequency of the churn event (churn rate) will

potentially damage the stability of the overlay. This condition is a critical one in a Chord-based P2P, since the validity of its finger table is an important point in supporting a successful lookup query [6]. When a superpeer fails, besides degrading the system performance, this condition also forces the normal peers to be disconnected from the system. In an online group-based application such as chatting or voice conferencing, the disconnected normal peers must reestablish the connection in order to remain joined to the service. According to Wahjuni et al. [17], the join request for a Chord ring produces a significant amount of latencies, which in turn degrade the successful lookup query rate. Therefore, it is important to focus attention on executing an efficient rejoin process.

To address churn issue, Peng et al. [12] applied multiple superpeer in each cluster instead of a single superpeer. As a consequence, this design has an additional table to maintain by each superpeer, that is list of superpeer in cluster. Normally, in hierarchical Chord-based P2P, a superpeer has two table to maintain, successor list table and normal peers table. Moreover, the flooding mechanism applied in each cluster lead to lower efficiency than Chord's as stated by the author. The works of Zoels et al. [13] concluded that simple star-like design is an optimum design for the lower layer of hierarchical P2P. Unfortunately, to our opinion, this design has not address the importance of efficient rejoin process yet. In other publication [18], the authors suggested to let each normal peer has a copy of successor list from its static peer if needed. When a superpeer fails, each normal peer send a rejoin request to the first superpeer in the list (perform rejoin individually). We argue that this rejoin mechanism, will produce a very high volume of rejoin traffic, specially when cluster/group size is big. Moreover, since the disconnected normal peers from the previous superpeer (which is the failed superpeer) will join to the same superpeer (which is the successor of the failed superpeer), the superpeer may overloaded. Interested in this optimum design [13], we proposed a low traffic rejoin approach, termed collective rejoin algorithm, in order to improve benefits of this design. In this paper we describe the efficiency of bandwidth consumption if the rejoin process is performed using our proposed collective rejoin algorithm rather than an individual rejoin algorithm.

II. RESEARCH METHOD

The proposed rejoin algorithm was implemented on a hierarchically structured P2P, and is part of a complete P2P protocol for heterogeneous networks [19]. A bandwidth consumption model was utilized to analyze the efficiency of the collective rejoin algorithm in consuming bandwidth and compared to that of the individual rejoin algorithm.

A. System Architecture

The hierarchical architecture implemented is a two-level hierarchical P2P as illustrated in Fig. 1. The top layer is a group of superpeers, which are nodes with high capacity, whereas the lower layer consists of groups of

normal peers. Nodes in the lower layer are organized in a simple star-like shape as suggested by Zoels et al. [13].

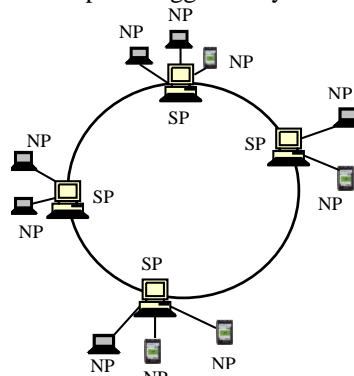


Figure 1. Hierarchical architecture

A new attribute, namely *neighbor_table*, was created in order to support the proposed collective rejoin algorithm. When a new normal peer joins the group, the superpeer should update the list and send it to the newly joining normal peer in response to its join request. Other normal peers within the group will receive this updated list as a response message during periodic maintenance. This list is used to record members of the group and the capacity of each node. In the generic framework, the category of the capacity is based on electric power capacities, which are categorized into three categories: the node with an unlimited electric power resource such as a desktop PC (cap1); a node with a moderate limited battery life, such as a computer notebook (cap2); and a node with a very limited battery life, such as a smartphone (cap3). A node has the chance to have role as a superpeer if it is in cap1 or cap2 category although nodes in the cap1 category have a higher probability of being assigned as a new superpeer than nodes in the cap2 category. We do not consider a cap3 category node to be a superpeer, as Zoels et al. [20] do, because its limited computing resources and electric power are not suitable for handling other peers. Another proposed attribute is time (*age*), which refers to how long the node has been joining the system. This information is required to guarantee that a peer that sends a request to join message with *age* > 0 (which is a peer that performs the rejoin process) will be assigned as a superpeer.

B. Collective Rejoin Algorithm

In the current hierarchical P2P architecture, when a superpeer experiences a failure, the disconnected normal peers should perform a rejoin process using the successor list which is copied from their superpeer. Using this approach, since each normal peer has to perform the rejoin mechanism individually, the join traffic increases linearly to the number of disconnected normal peers. In contrast, in the proposed approach during which the rejoin process is performed collectively, the amount of join traffic only depends on the number of failed superpeers. This benefit occurs because in a group that must perform a rejoin process, an elected normal peer performs the process on behalf of other members within

that group. The elected normal peer is a normal peer with highest capability in the group.

C. Bandwidth Consumption Model

Bandwidth consumption is the number of messages produced by the stabilization and rejoin process when a superpeer failure occurs. Basically, a network with N peers and α superpeer ratio will be grouped into G , as in

$$G = \alpha \times N, 0 < \alpha \leq 1 \quad (1)$$

Superpeer ratio is a proportion between the number of superpeers to all peers in an overlay network. Value of $\alpha=1$ means the system is a flat Chord. In the implemented architecture as illustrated in Fig.1, these groups form the upper layer of the hierarchical P2P, and G is the ring size. Suppose the normal peers are distributed uniformly among the superpeers, then each group has a size of $1/\alpha$. This means that each superpeer is responsible for handling β normal peers, which is

$$\beta = \frac{1}{\alpha} - 1 \quad (2)$$

Zoels et al. [14] found that the optimal operation for this type of architecture is for the superpeer ratio to be up to 25%. The basic Chord formula described by Stoica et al. [6] is used for traffic calculations that only consider the type of operation and not the number of messages needed to perform each operation. The idea behind this consideration is to form a generic bandwidth consumption analysis of the proposed algorithm.

In the implemented hierarchical architecture, when a node joins as a superpeer, it follows the Chord join algorithm. To find its right place in the ring, the node uses a mechanism similar to Chord's lookup query, which is costs $\log N$. In a ring with N peers, and α superpeer ratio, the cost for finding its position in the ring (*finding_pos*) becomes:

$$finding_pos = \log(G). \quad (3)$$

Once the node finds its position, it has to build its *finger_table* and *successor list*. In Chord [6], the cost for maintenance is $\log^2 N$, for P2P overlay contains N peers. Therefore, the traffic produced by the joining node for this process is

$$maint = \log(G) \times \log(G). \quad (4)$$

From (3) and (4) we can summarize that if a node joins as a superpeer, it will generate traffic by the amount of

$$join_s = finding_pos + maint = (\log(G) \times (1 + \log(G))). \quad (5)$$

On the other hand, according to Zoels et al. [18], if a node joins as a normal peer it only needs to send a request to the nearest static peer. For the system overall, this means sending one response message per rejoining normal peer. Suppose f is the ratio of superpeer failure, then the number of disconnected normal peers will be

$$D = f \times \alpha \times N \times \left(\frac{1}{\alpha} - 1\right) = f \times N(1 - \alpha) \quad (6)$$

The collective rejoin traffic (*rejoin_c*) will follow (5) with the multiplier factor of $f\alpha N$, that is

$$rejoin_c = f \times G \times join_s \quad (7)$$

The individual rejoin will follow both (5) and (6) with a proportion rejoining as superpeers and others rejoining as normal peers. Suppose p is the proportion of D that rejoin as normal peers, then the individual rejoin will be

$$rejoin_i = (1 - p) \times D \times join_s + p \times D \quad (8)$$

In the case of all the disconnected normal peers performing rejoin as normal peers ($p=100\%$), then the rejoin traffic depends only on (6). According to individual rejoin that described by Zoels et al. [18], the number of normal peers of the successor of the failed superpeer will double. Suppose β is maximum capacity of superpeers, then a load balancing mechanism must be executed in order not to overload the current superpeer. Otherwise, suppose all the normal peers are distributed randomly, then the new group size will be

$$\beta_1 = \frac{N \times (1 - \alpha \times j)}{(1 - j) \times \alpha \times N} = \frac{\frac{1}{\alpha} - f}{1 - f} \quad (9)$$

which is still bigger than β and a load balancing mechanism is needed.

III. RESULTS AND DISCUSSION

As shown in (7) and (8), the difference between the traffic of individual and collective rejoin depends on the following factors: number of peers, superpeer ratio, superpeer failure ratio, and the proportion of disconnected normal peers performing rejoin as normal peers. The analysis will be focused on the impact of the aforementioned factors on the amount of traffics generated by the rejoin process. The simulation was performed to validate the formal analytical results. Based on (7) and (8), varying α and N will have similar impacts. Therefore, without loss of generality, a static number of peers is used in the whole analysis as well as in the simulation. Otherwise stated, we set the network size to 1000 peers.

A. Impact of Ratio of Superpeer (α) on Bandwidth Consumption

In this section, the impact of the superpeer ratio to the bandwidth consumed by both rejoin algorithms is analyzed. Suppose a P2P overlay with N peers and α superpeer ratio experiences a failure on 10% of its superpeers, then the amount of failed superpeers is $0.1(\alpha N)$. The graph in Fig. 2 depicts traffic produced by individual rejoin (8) as suggested in [18] compare to our proposed collective rejoin (7), in various superpeer ratios (α). The higher the value of α means the smaller the size of the group (or in other words, the lower the workload of the superpeer). The value of α is varied from 1% to 10%, based on experimental results reported by Silverston et al. [21] that recorded no more than 10% of peers staying connected during the application session. Zoels et al. [14]

also suggested keeping the superpeer ratio below 25% of the network size in order to achieve optimum operation costs. It is assumed that 90% of all the disconnected normal peers will rejoin the system as normal peers (a more complete analysis for various normal peers join proportions are presented in the following section). The collective rejoin plots follow (7), while the individual ones follow (8) with p 90%. Although the increase of the bandwidth consumption of individual rejoin is slower than that of collective rejoin, the value is still higher. This is the normal impact, since in individual rejoin, each peer must perform a separate rejoin request. While in our proposed collective rejoin, the rejoin request is performed by the new superpeer on behalf of other peers in the group.

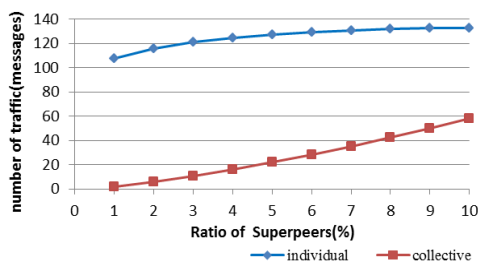


Figure 2. Number of rejoin traffic in various α

To have a representative comparison, the proportion of disconnected normal peers that perform individual rejoin as normal peers (p) is varied for various superpeer ratios (α). The same range of α as shown in Fig. 2 (1% - 10%) is used. Fig. 3 shows the plots of total individual rejoin when 50%, 75%, 90%, and 100% of the disconnected peers rejoin the system as normal peers (inversely 50%, 25%, 10%, and 0% join as superpeers). The graph shows that the higher the proportion of disconnected normal peers which perform rejoin as normal peers, the less bandwidth that is consumed. Moreover, it shows that at p 100% , the amount of traffic is likely to decline, as depicted with equation (8). In complete experimental results for $N = 1000$, the amount of individual rejoin traffic for this proportion will be lower than the collective one at α 15%. This eagers the result in [14], which concluded that the maximum value of superpeer ratio for optimum operational cost is 25%. Nevertheless, the amount of new joining peers may lead to an overload of the superpeer, as discussed in Section 2, and an additional load balancing mechanism should be executed to redistribute the normal peers.

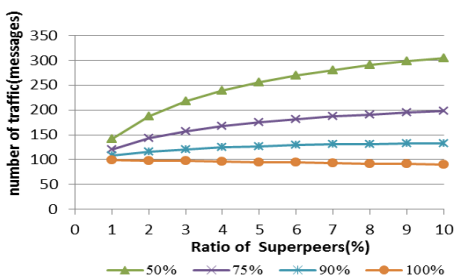


Figure 3. Number of individual rejoin traffic for various p and α

B. Impact Analysis of Ratio of Superpeer Failure (f) on Bandwidth Consumption

Fig. 4 and Fig. 5 represent the correlation between the superpeer failure ratio and the bandwidth consumption of both rejoin algorithms. As suggested by Silverston et al. [21], the value of α is set to 10%. As in part A of this section, p value of 90% is used. For all plots, the higher failure ratio of the superpeer produces higher rejoin traffic, which is a logical consequence of the increase in the number of failed superpeers. Since more failed superpeers cause more disconnected normal peers, the ratio of superpeer failures impacts individual rejoin more significantly than collective rejoin (in Fig. 4 this conclusion is indicated by the higher gradient value of the individual rejoin graph than that of the collective rejoin).

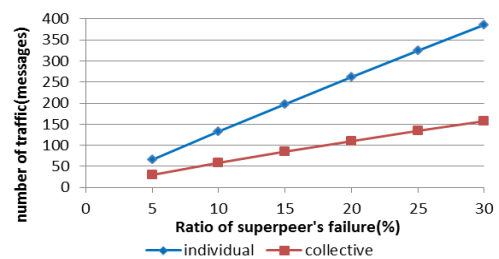


Figure 4. Number of rejoin traffic in various f

For all proportions of disconnected normal peers performing rejoin as normal peers, the higher superpeer failure ratio (f) generates more rejoin traffic. A pattern similar to that depicted in Fig. 3 occurs, where the lower p value produces the higher traffic. Even though at p 100% the lowest bandwidth consumption value is produced, this situation may lead the superpeer being overloaded.

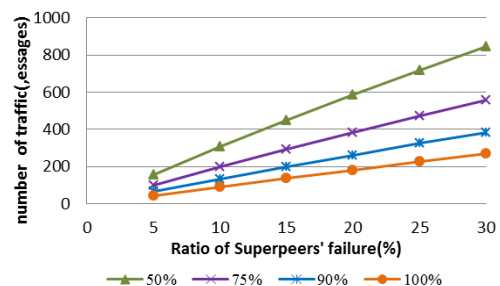


Figure 5. Number of individual rejoin traffic for various p and f .

C. The Bandwidth Consumption Efficiency

In Fig. 6 and Fig. 7 the bandwidth consumption efficiencies of collective rejoin are presented. The bandwidth consumption efficiency (eff) is obtained using the formula

$$eff = \frac{rejoin_i - rejoin_c}{rejoin_i} \tag{10}$$

The graph in Fig. 6 shows the bandwidth consumption efficiency of collective rejoin for various ratios of superpeer (α) when various proportion of disconnected normal peers (50%, 75%, 90%, and 100%) perform rejoin as normal peers. Using the same variation of f as in all

part of this paper [5%..30%], the bandwidth consumption efficiency for various ratios of superpeer failure are presented in Fig. 7. For all variations of p , the bandwidth consumption efficiency increases with the higher superpeer failure ratio. The graphs in Fig. 6 and Fig. 7 demonstrate that the collective rejoin approach produced less traffic than that of individual rejoin approach. The decreasing of efficiency resulting from the higher superpeer ratio as shown in Fig. 6 is a logical impact of the smaller number of normal peers when the value of α is increased. As stated in (2), the higher value of α means the smaller size of group.

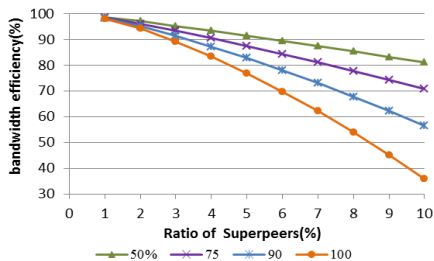


Figure 6. Bandwidth consumption efficiency for various α

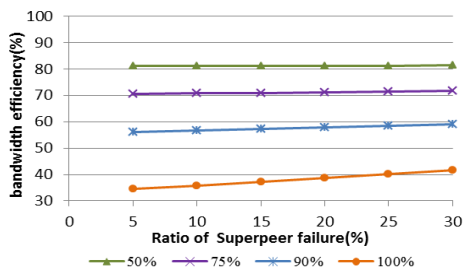


Figure 7. Bandwidth consumption efficiency for various f

D. Simulation Results

To validate the results of previous formal analyses, a Java package was developed to implement the proposed collective rejoin algorithm. The package is bundled as an additional package on Peersim [22]. Separate configuration file was used for each of the simulation scenarios. The generic scenario is as follows: the overlay was created first, and then the superpeer failure event was performed for one hour simulation time (Peersim uses a simulation unit time instead of computer system unit time; we equated one second to 100 simulation units time).

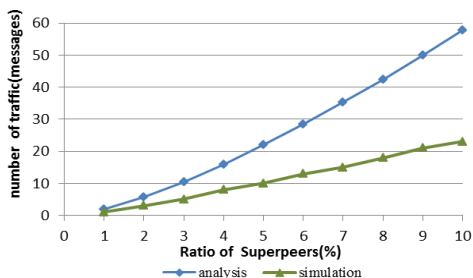


Figure 8. Collective rejoin traffic for various α

In Fig. 8, the formal analytical result of the collective rejoin algorithm for various ratios of superpeer (α) is

compared to the output of simulation. The same range of α as in previous formal analysis [1%...10%] is used. Although both plots have different values, but they have similar pattern. The difference between analysis result and simulation result occurs, since the value in analysis result is the maximum one. The same behaviour as in Fig. 8 occurs, as shown in Fig. 9, when the similar comparison perform for various ratios of superpeer failure (f).

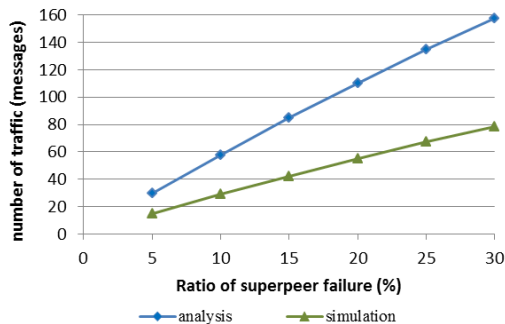


Figure 9. Collective rejoin traffic for various f

IV. CONCLUSION

Hierarchically structured P2P is an appropriate architecture for heterogeneous networks. In order to minimize the performance degradation produced by the normal peers rejoin process following a superpeer failure event, the proposed collective rejoin algorithm is an appropriate solution. The analytical results show that our proposed collective rejoin algorithm has lower bandwidth consumption than the individual rejoin algorithm in all scenarios. The efficiency is even more significant, in the case of the high dynamic overlay which is represented by the high superpeer failure ratios. As it is showed in Fig. 4, for the same incremental value of the superpeer’s failure ratio, the increase of traffic produced by individual rejoin is higher than that of collective rejoin. Although the graphs in Fig. 6 shows that for the higher superpeer ratio, the efficiency is decreased, but the value is still more than 30%. From these two experimental schemes we conclude that the collective rejoin algorithm is an efficient approach for a P2P overlay network that highly dynamics as well as for that with a big gopup size. These results are validated by the simulations ouput that produced similar patterns for our proposed collective rejoin algorithm. The similar pattern also occurs in the comparison of formal analytical result and simulation result for individual rejoin algorithm.

ACKNOWLEDGMENT

The main author of this paper thanks the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University for facilitating her study. Part of this study is also supported by The Directorate General of Higher Education, Ministry of Education and Culture, The Republic of Indonesia under Sandwich-Like 2012 Programme (No. 2954.6/E4.4/2012).

REFERENCES

- [1] J. Kangasharju, "Peer-to-Peer System," in *Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises*, Ney York: IGI Global, 2008, pp. 174-189.
- [2] <http://www.napster.com>
- [3] <http://www.gnutella.com>
- [4] <http://www.kazaa.com>
- [5] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network". *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '01)* pp. 161-172, 2001.
- [6] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup protocol for Internet applications," *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 17 - 32, Feb. 2003.
- [7] A. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems", *Lecture Notes in Computer Science*, 2218:329-350, 2001
- [8] B. Y. Zhao, L. Huang, J. Stribling, S. Rhea, A. D. Joseph, and J. D. Kubiatowicz, "Tapestry: A Resilient Global-scale Overlay for Service Deployment," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 1, pp. 41-53, 2004.
- [9] L. Garces-Erice, E. W. Biersack, P. A. Felber, K. W. Ross, and G. Urvoy-Keller, "Hierarchical Peer-to-Peer Systems", *Proceeding of the ACM/IFIP International Conference on Parallel and Distributed Computing (Euro-Par)*, Klagenfurt, Austria, 2003
- [10] A. Montresor, "A Robust Protocol for Building Superpeer Overlay Topologies", *Technical Report UBLCS-2004-8*, 2004.
- [11] S. Zoels, Z. Despotovic, and W. Kellere, "On hierarchical DHT systems - An analytical approach for optimal designs", *ScienceDirect, Computer Communications* 31 pp. 576-590, 2008.
- [12] Z. Peng, Z. Duan, J. Qi, Y. Cao and E. Lv, "HP2P: A Hybrid Hierarchical P2P Network", *Proceedings of the 1st International Conference on the Digital Society (ICDS'07)*, 2007.
- [13] M. Pandey, S. M. Ahmed, and B. D. Chaudhary, "2T-DHT: A Two Tier DHT for Implementing Publish/Subscribe", *Proceeding of The International Conference on Computational Science and Engineering (CSE '09)* pp. 158-165, Vancouver, 29-31 Aug. 2009
- [14] S. Zoels, Q. Hofstatter, Z. Despotovic, and W. Kellerer, "Achieving and Maintaining Cost-Optimal Operation of a Hierarchical DHT System". *Proceeding of The IEEE International Conference on Communications ICC '09*, 2009.
- [15] J. Li, C. Li, Z. Fang, and H. Wang, "Layer Optimization for DHT-based Peer-to-Peer Network", *Proceeding of The 10th Annual Conference on Networks (ICN) 2011* pp. 382-387.
- [16] S. Rhea, D. Geels, T. Roscoe, J. Kubiatowcz, "Handling Churn in a DHT", *The USENIX 2004 Annual Technical Conference, Boston, MA, USA*, 2004.
- [17] S. Wahjuni and K. Ramli, "Considering The Nodes Join/Leave Behavior in The Analysis of Chord Stabilization in The Heterogeneous Network", *International Journal of Computer Science and Network Security*, Vol. 11 No. 12 pp. 57-61, 2011.
- [18] S. Zoels, Z. Despotovic, and W. Kellerer, "Cost-Based Analysis of Hierarchical DHT Design," in *Sixth IEEE International Conference on Peer-to-Peer Computing, 2006. P2P 2006*, 2006, pp. 233-239
- [19] S. Wahjuni, A. A. P. Ratna, and K. Ramli, "Efficient normal peers group recovery in hierarchical peer-to-peer," in *2012 IEEE International Conference on Communication, Networks and Satellite (ComNetSat)*, 2012, pp. 6-10.
- [20] S. Zds, R. Schollmeier, W. Kellerer, and A. Tarlano, "The hybrid chord protocol: a peer-to-peer lookup service for context-aware mobile applications," in *Proceedings of the 4th international conference on Networking - Volume Part II*, Berlin, Heidelberg, 2005, pp. 781-792.
- [21] T. Silverston, O. Fourmaux, A. Botta, A. Dainotti, A. Pescapé G. Ventre, and K. Salamatian, "Traffic analysis of peer-to-peer IPTV communities," *Computer Networks*, vol. 53, no. 4, pp. 470-484, Mar. 2009.
- [22] <http://sourceforge.net/projects/peersim>.

Sri Wahjuni is currently working toward a Ph. D degree in the Department of Electrical Engineering, University of Indonesia. She is a member of the Laboratory of Net-Centric Computing and a faculty member of the Department of Computer Science, Bogor Agricultural University. Her research interests include embedded systems, mobile computing, and ubiquitous networks.

A.A. Putri Ratna is senior lecturer at the Faculty of Engineering, University of Indonesia. She obtained her Master's degree at Waseda University, Japan, in 1990 and her Doctoral degree at the University of Indonesia in 2006. Her research interests include computer networks and web-based information systems.

Kalamullah Ramli is Professor of Computer Engineering at the Faculty of Engineering, University of Indonesia. He finished his Master's degree in Telecommunication Engineering at the University of Wollongong, NSW, Australia, in 1997 and obtained a Doktor-Ingenieur in Computer Networks in 2003 from Universitaet Duisburg-Essen, NRW, Germany. His research interests include embedded systems, computer and communications, and mobile applications.

A Method of Case Retrieval for Web-based Remote Customization Platform

Yuhuai Wang^{1,2*}, Hong Jia², and Xiaojing Zhu³

1. Qianjiang College, Hangzhou Normal University, Hangzhou, China

2. Key Laboratory of E&M (Zhejiang University of Technology), Ministry of Education & Zhejiang Province, Hangzhou, China

3. EP Equipment CO., LTD. AnJi, China

*Corresponding author, Email: wyiya@hotmail.com, hjia@zjut.edu.cn, jackzhu@ep-ep.com

Abstract—A web-based remote customization platform is developed for hardware product in this paper. To realize the rapid product customization, a case-based reasoning approach based on fuzzy set is put forward. To retrieve the most similar case from the case base, a parabola membership function is constructed based on the fuzzy set, and synthesis weights are introduced by combining subjective weights with objective weights which are calculated based on the deviation information of similarity. Then the model for solving cases' global similarity is set up based on synthesis weights. To improve the accuracy of the similarity measurement, center distance revision method based on area is presented for the Bi-interval type which is one of fuzzy numeric attribute. Implementation example applying above methods is given in the area of electric drill customization. Results show that the presented approach helps to improve the accuracy of the similarity of the case product, and reduce the time and cost of product design process.

Index Terms—Individual Customization; Case-Based Reasoning; Membership Function; Similarity; Case Retrieval

I. INTRODUCTION

With the fast development of information technology, lots of manufacturing companies have set up remote online product customization system to meet user's need. Most of the systems could only provide product pictures, 3D product model information could not be demonstrated, all those limits the interactive communication between users and product designers, and this limitation will mostly increase the time and cost of product design process. To develop a web-based 3D product collaborative design system is an effective way to solve the problem.

Case-Based Reasoning is a good method in the fields of fast configuration design of mass customization. It has advantage in inducing and extracting the reasoning rules [1, 2]. Similar case indexing is the key of CBR [3]. Take impact drill as example, if product model with high similarity to the goal product could be found from the case base, then design period could be greatly shortened. As user's requirements are not clear enough in some circumstances, it is vital to calculate the similarity of

those unclear attributes such that the most similar case could be found.

Ref. [4] and Ref. [5] present a fuzzy logic approach which calculates the similarity and retrieves the best case based on the distance function and the fuzzy number converted from the exact number by the Gaussian function. However, cases' attributes could not be described by distance function which leads to the incomplete description of similarity and its inaccuracy. Ref. [6] and Ref. [7] propose the hybrid measure for comparing cases with a mixture of crisp and fuzzy features without considering the uncertainty for requirements. Similarity of fuzzy linguistic attribute and intervals could be calculated by applying the proposed measures. Accuracy of similarity is improved to some extent, and the measurement is used in the area of fault diagnosis, for example, abnormal tire wear. The limitations of the method are: (1) Hamming distance function is used for fuzzy attributes, (2) Overlapped area is repeated calculated of the similarity calculation of interval type. And this limitations set obstacle for obtaining the most accurate similarity and retrieving the best case. Ref. [8] and Ref. [9] solve the similarity to retrieve the case using the membership function constructed by using of the triangular, trapezoidal and Gaussian function. However, the membership functions are to some extent inaccuracy or complex [10]. Attributes of interval type are not considered in this method. And weights' average value is used as attribute weight. These could decrease the accuracy. On the other hand, triangular and trapezoidal function could not describe the characteristics of the cases' attribute correctly. And Gaussian function is too complicated for calculation.

To improve accuracy of similarity and to take cases' attributes into account, the paper puts forward a case-based reasoning approach based on fuzzy set (FSCBR). In this paper, first, the standard model set is defined. Then the parabola membership function of the model attribute is constructed, the similarity of the membership is calculated, the most matching case is searched and the rapid design platform based on the above approach is developed.

II. KEY TECHNOLOGIES OF CUSTOMIZATION PLATFORM

A. System Architecture

The proposed remote customization platform is composed of two parts which are server and client side. Customization design process could be described as following steps: first, user submit their demands to server from client side. Then, on server side, engineer from the company searching for the closest case from case base by applying the FSCBR method and the matched case with its 3D model are provided to user for 3D browsing and further processing. Fig. 1 shows the Model-View-Controller (MVC) architecture of the customization platform.

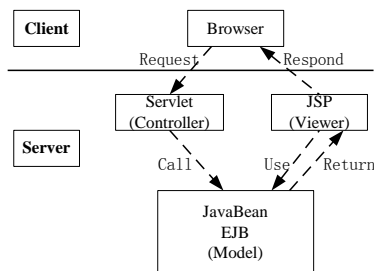


Figure 1. MVC architecture

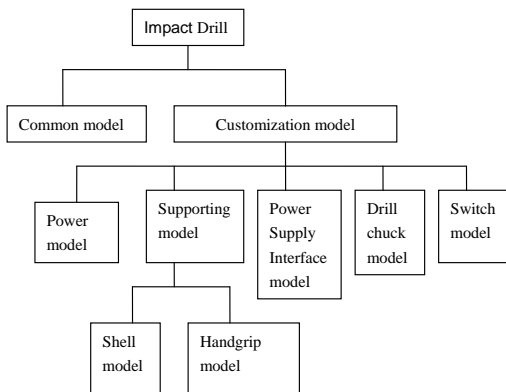


Figure 2. Components of impact drill model

B. 3D Interactive Browsing

AutoVue is used as 3D browser in the system. Functions like rotation, pan, zoom, sectioning and so on are provided. Geometrical and topological information could be extracted and enquired. 3D notation of design comments helps user better understands products. Part of the program codes is displayed here:

```

</tr>
<td width="15%">Product No.:</td>
<td width="15%" height="30"><%=project_code%></td>
<td rowspan="10">
<object id="AutoVueX" classid="clsid:B6FCC215-D303-11D1-
BC6C-0000C078797F" width="500" height="350">
<param name="src" value="<%=picture%>">
</object>
</td>
</tr>
    
```

C. Compact Drill Model

According to the structural analysis and the user requirements, components of compact drill are illustrated

as Fig. 2. The main attributes include: input power, body color, handle position, handle color, power supply voltage, maximum diameter of the drill, high idling speed, net weight, torque.

D. Proposed Approach

Case set is composed of product cases, it is written as M . Suppose there are n samples in M , and each sample has m attributes. Attribute set F is defined as

$$F = \{f_j \mid j = 1, 2, \dots, m\} \tag{1}$$

Attribute vector of i^{th} sample is defined as

$$\mathbf{c}_i = (z_{i1}, z_{i2}, \dots, z_{im}) \tag{2}$$

where: z_{ij} is the j^{th} attribute value of \mathbf{c}_i , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$. So M can be represented as

$$M = (\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n)^T = (z_{ij})_{n \times m} \tag{3}$$

Because of the inaccurate of the input and output variables under fuzzy condition, trapezoidal membership function, triangular function and Gaussian function are normally used to calculate the similarity [11].

According to the characteristic, case attributes can be divided into four types as crisp symbolic attribute (CS), crisp numeric attribute (CN), fuzzy linguistic attribute (FL) and fuzzy numeric attribute (FN). A membership function illustrates the degree of membership for each possible crisp value of the fuzzy variables.

(1) CS attribute membership function

For CS attribute, there's no realistic quantity relation among all its possible values. We define crisp symbolic attribute membership function as

$$\mu_s(z) = \begin{cases} 1, & v = z \\ 0, & v \neq z \end{cases} \tag{4}$$

where: v is the value of the input attribute, z is the attribute value of the case base.

(2) CN attribute membership function

For CN attribute, its value represents a point in the attribute space. Distance between points describes the difference among case attributes. To define membership function based on distance is a good way for calculating the similarity among case attributes. CN attribute membership function is defined as following,

$$\mu_s(z) = 1 - \frac{|z - v|}{\max(z_i, v)} \quad ; i = 1, 2, \dots, n \tag{5}$$

(3) FN attribute membership function

In order to improve the accuracy and flexibility of problem description, an estimative figure v is always provided in practice. Fuzzy set theory is a suitable approach for assessing the similarity among these fuzzy attributes.

Assume fuzzy set S in the discourse domain D as following

$$S = \{(z, \mu_s(z)) \mid z \in D\} \tag{6}$$

TABLE I. PARAMETER VALUES FOR THE MEMBERSHIP FUNCTION EQ. (8)

	equal(=)	Not less than (> or ≥)	Not greater than (< or ≤)
c	v	max(f)	min(f)
di	min(λv; v - min(f))	min(max(f) - (v - λv); max(f) - min(f))	0
ds	min(λv; max(f) - v)	0	min((v + λv) - min(f); max(f) - min(f))

where: $\mu_s(\cdot)$ is the membership function of S , $\mu_s(z) \in [0,1]$ describes the grade of membership of z in S . The nearer the value of $\mu_s(\cdot)$ to unity, the higher the grade of membership of z in S .

Unlike scalars and intervals, fuzzy numbers are uncertain numbers for which, in addition to knowing a range of possible values, one can say that some values are more plausible, or ‘more possible’ than others. Triangular fuzzy number and trapezoidal fuzzy number are the most widely used fuzzy number types for decision making under the condition of fuzzy environment [12].

TABLE II. PARAMETER VALUES FOR THE MEMBERSHIP FUNCTION EQ. (9)

	Inside the range
c1	v1
c2	v2
di	min(λv1; v1 - min(f))
ds	min(λv2; max(f) - v2)

Fuzzy set based on parabola is applied here to simulate fuzzy numeric interval attribute, its function is defined as following,

$$L(z) = R(z) = z^2 \tag{7}$$

The membership function is denoted as,

$$\mu_s(z) = \begin{cases} 0 & , z \in [0, c1 - di] \\ \frac{1}{di^2} (z - c1 + di)^2 & , z \in [c1 - di, c1] \\ 1 & , z \in [c1, c2] \\ \frac{1}{ds^2} (z - c2 - ds)^2 & , z \in (c2, c2 + ds] \\ 0 & , z \in (c2 + ds, +\infty) \end{cases} \tag{8}$$

If $c1 = c2$, the membership function will be the same as that of FN, which is illustrated as following,

$$\mu_s(z) = \begin{cases} 0 & , z \in [0, c - di] \\ \frac{1}{di^2} (z - c + di)^2 & , z \in [c - di, c] \\ 1 & , z = c \\ \frac{1}{ds^2} (z - c - ds)^2 & , z \in (c, c + ds] \\ 0 & , z \in (c + ds, +\infty) \end{cases} \tag{9}$$

Suppose f as the membership, $d(f)$ as the domain, $\min(f)$ and $\max(f)$ are the minimum value and maximum value of $d(f)$, then its membership function

could be represented as $f(di, ds, c1, c2)$. Tab. 1 and Tab. 2 show the parameter values.

There exist six relationships, which are equal, less than, greater than, not less than, not greater than and within the range. And these six relationships could be grouped into three types: type I (Fig. 3) as equal, type II (Fig. 4 and 5) as less than, greater than, not less than, not greater than, type III (Fig. 6) as within the range.

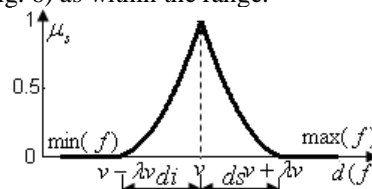


Figure 3. Type I

For fuzzy numerical interval attribute, if the input case attribute is also belong to type III, then there’s difficulty in calculating the attributes’ similarity by using above membership function. An approach based on overlapped area is put forward to revise the similarity for the Bi-interval type. Membership function is formulated based on Eq. (8), and the similarity of the fuzzy sets is calculated by computing the area overlapping rate of corresponding membership functions.

$$\begin{aligned} sim(x, y) &= A(x \cap y) / A(x \cup y) \\ &= A(x \cap y) / (A(x) + A(y) - A(x \cap y)) \end{aligned} \tag{10}$$

$$\begin{aligned} A(x) &= \int_{d(f)} \mu_s dz = \int_{c1-di}^{c2+ds} \mu_s(z) dz \\ &= \int_{c1-di}^{c1} \frac{(z - c1 + di)^2}{di^2} dz + \int_{c2}^{c2+ds} \frac{(z - c2 - ds)^2}{ds^2} dz + (c2 - c1) \times 1 \\ &= di / 3 + ds / 3 + c2 - c1 \end{aligned} \tag{11}$$

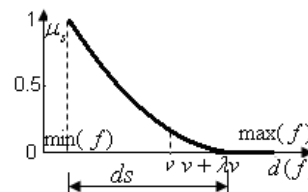


Figure 4. Type II (less than and not greater than)

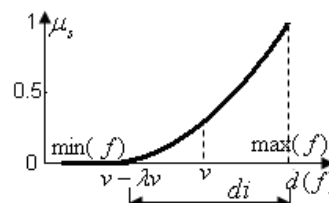


Figure 5. Type II (greater than and not less than)

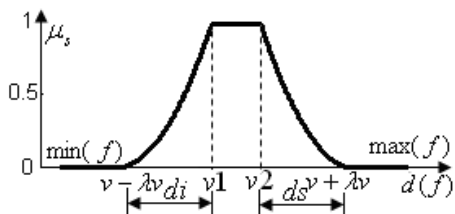


Figure 6. Type III

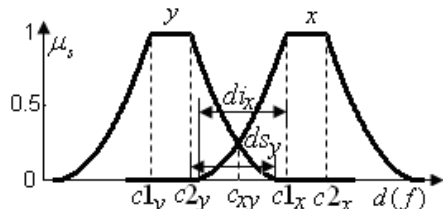


Figure 7. The representation of interval attribute and requirements

Take following Fig. 7 as example, the intersection area can be represented as

$$A(x \cap y) = \begin{cases} 0 & ; \quad x \cap y = \emptyset \\ \min(A(x); A(y)) & ; \quad x \subseteq y \quad \text{or} \quad y \subseteq x \\ \frac{(c_{xy} - c1_y + di_y)^3}{3di_y^2} + \frac{(c2_x + ds_x - c_{xy})^3}{3ds_x^2} & ; \quad x \cap y \neq \emptyset; c2_x \leq c1_y \\ \frac{di_y}{3} + \frac{ds_x}{3} + c2_x - c1_y & ; \quad x \cap y \neq \emptyset; c1_y < c2_x < c2_y \\ \frac{di_x}{3} + \frac{ds_y}{3} + c2_y - c1_x & ; \quad x \cap y \neq \emptyset; c1_x < c2_y < c2_x \\ \frac{(c_{xy} - c1_x + di_x)^3}{3di_x^2} + \frac{(c2_y + ds_y - c_{xy})^3}{3ds_y^2} & ; \quad x \cap y \neq \emptyset; c2_y \leq c1_x \end{cases} \quad (15)$$

For the Bi-interval type, if $x_i \subseteq y_i$, the coefficient k is introduced to revise the similarity based on the relative area.

Let $C_x = \frac{c1_x + c2_x}{2}$ and $C_y = \frac{c1_y + c2_y}{2}$, so

$$\varepsilon = |C_x - C_y| / (|c2_y - c1_y| / 2) \quad (16)$$

$$\overline{sim}(x, y) = ksim(x, y) \quad (17)$$

where:

$$k = \begin{cases} 1 & ; \varepsilon \geq 1 \\ 1 - \varepsilon & ; \varepsilon < 1 \end{cases} \quad (18)$$

(4) FL attribute membership function

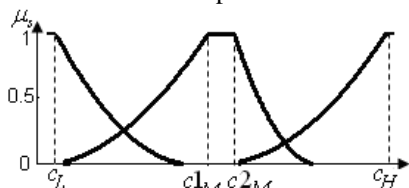


Figure 8. The representation of FL

FL attribute is a fuzzy concept which is associated with a certain fuzzy region. Because of the uncertainty of FL

$$A(x \cap y) = \int_{c1_x - di_x}^{c_{xy}} \frac{(z - c1_x + di_x)^2}{di_x^2} dz + \int_{c_{xy}}^{c2_y + ds_y} \frac{(z - c2_y - ds_y)^2}{ds_y^2} dz \quad (12)$$

$$= \frac{(c_{xy} - c1_x + di_x)^3}{3di_x^2} + \frac{(c2_y + ds_y - c_{xy})^3}{3ds_y^2}$$

where: c_{xy} is the value of the intersection point of fuzzy set border x, y .

Suppose $\mu_s(z_y) = \mu_s(z_x)$, and according to Eq. (8), we got

$$(z - c2_y - ds_y)^2 / ds_y^2 = (z - c1_x + di_x)^2 / di_x^2 \quad (13)$$

and

$$c_{xy} = (c2_y di_x + c1_x ds_y) / (di_x + ds_y) \quad (14)$$

The solution to intersection area could be written as:

attribute, the constructed membership function is shown in Fig. 8. The similarity can be measured by the FNI.

E. The Global Similarity

(1) Assess of hybrid weight

Hybrid weight is composed of two kinds of weights. For weight which illustrates the attributes from experts' subjective experience, it is defined as

$$\mathbf{w}^* = (w_1^*, w_2^*, \dots, w_m^*) \quad (19)$$

For objective weight which illustrates the attributes of the case product, it is defined as

$$\mathbf{w}^* = (w_1^*, w_2^*, \dots, w_n^*) \quad (20)$$

Suppose M_x as the case to be designed, s_{ij} is the similarity of the j^{th} attribute of M_x and M_i , similarity matrix of all the case attributes from M_x and M_i can be written as,

$$\mathbf{s} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1m} \\ s_{21} & s_{22} & \dots & s_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{nm} \end{pmatrix} \quad (21)$$

Objective weight can be calculated by following formula.

TABLE III. REQUIREMENTS FOR RAPID DESIGN OF DRILL

Attribute	Input power/W	Body color	Handle position	Handle color	Power supply voltage/V	Maximum diameter of the drill/mm	High idling speed/rpm	Net weight /Kg	Torque /Nm
M_1	850	Black	bottom	Black	230	19	2800	2.7	5~15
M_2	1200	Black	middle	Grey	110	13	550	1.8	45~60
M_3	950	Red	bottom	Red	220	16	2000	2.5	25~35
...
M_s	900~1000	Red	bottom	Red	220	16	normal	<=3.0	10~40
Uncertainty degree	30%	0	0	0	0	40%	20%	20%	20%
Can be ignored?	No	No	No	Yes	No	No	No	No	No
Type	FN	CS	CS	CS	CN	FN	FL	FN	FN
Relationship	within	-	-	-	-	equal	mean	<=	interval
Weight(w^*)	0.25	0.05	0.05	0	0.1	0.15	0.15	0.1	0.15

$$w_j^* = [\sum_{i=1}^{n-1} \sum_{k=i+1}^n (s_{ij} - s_{kj})^2] / \sqrt{\sum_{j=1}^m [\sum_{i=1}^n \sum_{k=i+1}^n (s_{ij} - s_{kj})^2]} \quad (22)$$

where $\sum_{i=1}^n \sum_{k=i+1}^n (s_{ij} - s_{kj})^2$ represents the sum of square of similarity deviations of the j^{th} attribute of each case.

The hybrid weight of case M_i can be given by following formula.

$$w = (\frac{w_1^* w_1^*}{\sum_{j=1}^m (w_j^* w_j^*)}, \frac{w_2^* w_2^*}{\sum_{j=1}^m (w_j^* w_j^*)}, \dots, \frac{w_m^* w_m^*}{\sum_{j=1}^m (w_j^* w_j^*)}) \quad (23)$$

(2) Global similarity

Solution model for global similarity can be denoted as following,

$$sim = s_{n \times m} w_{m \times 1}^T = (sim_1, sim_2, \dots, sim_n)^T \quad (24)$$

where:

$$sim_i = \sum_{j=1}^m w_j s(i, j) / \sum_{j=1}^m w_j, \quad i = 1, 2, \dots, n$$

$$\sum_{j=1}^m w_j = 1.$$

III. CASE STUDY AND SYSTEM IMPLEMENTATION

A. Case Study of Impact Drill

Tab. 3 shows the examples and requirements for the customized electric drill

(1) Similarity measurement for CS

Take attribute “handle position” for example, suppose the requirement of this attribute is “bottom”, if attribute value of the matching case is bottom, then the similarity equals 1, otherwise is 0.

(2) Similarity measurement for CN

Take “power supply voltage” for example, according to Eq. (5), following results could be attained

$$\mu_s(230) = 1 - |230 - 220| / 230 \approx 0.957,$$

$$\mu_s(110) \approx 0.522,$$

$$\mu_s(220) = 1.$$

(3) Similarity measurement for FN

(a) Type I

Take attribute ‘Maximum diameter of the drill’ as example, its specification is 4-49. According to Tab. 1, we got $c = v = 16$ and $di = ds = 0.4 \times 16 = 6.4$. Applying Eq. (9), we obtained

$$\mu_s(19) = (19 - 16 - 6.4) / 6.4^2 = 0.282,$$

$$\mu_s(13) = (13 - 16 + 6.4) / 6.4^2 = 0.282,$$

$$\mu_s(16) = 1.$$

(b) Type II

Take attribute ‘Net weight’ as example, the general weight range of drill is [1.5, 7.0]. Design requirement is less than or not greater than 3.0 kg. According to Tab. 1, we got $v = 3.0, \max(f) = 7.0, \min(f) = 1.5$, $\lambda v = 0.2 \times 3.0 = 0.6$ and $ds = 2.1$.

Applying Eq. (9), we obtained

$$\mu_s(z) = (z - 3.6)^2 / 2.1^2 \quad (25)$$

Then, $\mu_s(2.7) = 0.184, \mu_s(2.5) = 0.274, \mu_s(1.8) = 0.73$.

(c) Type III

Take attribute ‘Input power’ as example, design requirement is ‘900-1000’. According to Tab. 2, we got $c1 = v1 = 900, c2 = v2 = 1000, di = 0.3 \times 900 = 270$ and $ds = 0.3 \times 1000 = 300$.

Applying Eq. (9), we obtained

$$\mu_s(850) = (850 - 630)^2 / 270^2 = 0.664,$$

$$\mu_s(1200) = 0.111,$$

$$\mu_s(950) = 1.$$

(d) Bi-interval type

Take attribute ‘Torque’ as example, both requirement attribute and the case attribute are interval type. Suppose case1, case2, case3 and requirement attribute as $x1, x2, x3, y$. According to Eq. (11) and Tab. 2, we got

$$A(x1) = 11.333, A(x2) = 22,$$

$$A(x3) = 18.667, A(y) = 33.333$$

As there’s point of intersection between the torque curves of case 2 and the requirement, according to Eq. (14), we got

$$c_{x2y} = 42.353$$

Substituting it into Eq. (15) and Eq. (10), then we got

$$sim(x1, y) = 6.667 / (33.333 + 11.333 - 6.667) = 0.175,$$

$$sim(x2, y) = 1.993 / (33.333 + 22 - 1.993) = 0.037 ,$$

$$sim(x3, y) = 18.667 / 33.333 = 0.56 .$$

And, considering

$$c_{x3} = (20 + 35) / 2 = 27.5$$

$$c_y = (10 + 40) / 2 = 25$$

$$\epsilon = |27.5 - 25| / |(40 - 10) / 2| \approx 0.167$$

Then,

$$sim(x3, y) = (1 - 0.167) \times 0.56 = 0.466 .$$

(4) Similarity measurement for FL

Take attribute ‘High idling speed’ as example, design requirement is ‘normal’. The survey shows that the high, normal and low attributes of ‘High idling speed’ are [2500, 3200], [1000, 2500] and [500, 1000] respectively. So, the similarity can be measured by type II.

$$\mu_s(2800) = 0.160, \mu_s(550) = 0, \mu_s(2000) = 1$$

If the case attribute is also the fuzzy concept, the similarity can be measured according to Bi-interval type.

So, the similarity matrix can be given as following

$$s = \begin{pmatrix} 0.664 & 0 & 1 & 0.957 & 0.282 & 0.16 & 0.184 & 0.175 \\ 0.111 & 0 & 0 & 0.522 & 0.282 & 0 & 0.73 & 0.037 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0.274 & 0.466 \end{pmatrix}$$

According to Eq. (22), objective weight vector of similarity w^* can be given as

$$w^* = (0.32, 0.53, 0.53, 0.11, 0.28, 0.46, 0.14, 0.08)$$

Considering expert’s subjective weight

$$w^* = (0.25, 0.05, 0.05, 0.1, 0.15, 0.15, 0.1, 0.15) ,$$

hybrid weight vector w can be attained by applying formula (23)

$$w = (0.29, 0.09, 0.09, 0.04, 0.15, 0.25, 0.05, 0.04) .$$

Global similarity can be calculated as following by applying Eq. (24)

$$sim = s_{3 \times 8} w^T_{8 \times 1} = (0.421, 0.131, 0.943)^T .$$

Figure 9. Interface for inputting user requirements

B. System Implementation

Applying SQL Server and tomcat 6.0 together with the presented approaches, a Web-based remote customized

impact drill system is developed. Based on the parameters submitted by the user from client side shown in Fig. 9, the system finds the most suitable case from the existing case base by employing FSCBR algorithm. The corresponding 3D STEP model illustrated in Fig. 10 is given at the same time, which provides user the possibility to browse the product model and further interactive operation with product designer.

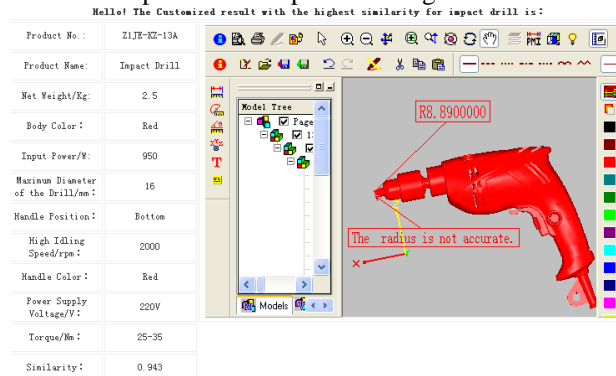


Figure 10. Illustration of the best case and its 3D model

C. FSCBR Retrieving Results

The similarity of all attributes shown in Tab. 3 is calculated by the nearest neighbor method, Hamming distance method and fuzzy similarity method (FSM) presented by Ref. [6], respectively.

The matrix of the attributes similarity with the nearest neighbor method is as following,

$$\begin{pmatrix} 0 & 0 & 1 & 0.917 & 0.5 & - & 0.75 & 0.091 \\ 0.286 & 0 & 0 & 0 & 0 & - & 0 & 0.455 \\ 1 & 1 & 1 & 1 & 1 & - & 0.583 & 1 \end{pmatrix} .$$

The matrix of the attributes similarity with Hamming distance method is as following,

$$\begin{pmatrix} 0.714 & 0 & 1 & 0.917 & 0.5 & - & 0.75 & 0.727 \\ 0.286 & 0 & 0 & 0.083 & 0.5 & - & 0 & 0.5 \\ 1 & 1 & 1 & 1 & 1 & - & 0.583 & 0.909 \end{pmatrix} .$$

The matrix of the attributes similarity with FSM is as following,

$$\begin{pmatrix} 0.664 & 0 & 1 & 0.957 & 0.282 & 0.16 & 0.184 & 0.175 \\ 0.111 & 0 & 0 & 0.522 & 0.282 & 0 & 0.73 & 0.037 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0.274 & 0.56 \end{pmatrix} .$$

The global similarity between M_x and M_i ($i=1, 2, 3$) listed in Tab. 3 is shown in Tab. 4.

Illustrated examples show that the final case found by FSCBR is the same as by Hamming Distance, Nearest Neighbor Method and FSM. M_3 is the closest case and M_2 is the least similar case. However, imilarity differences exist among those methods. The $sim(M_x, M_3)$ is almost the same while using FSCBR and Hamming Distance. And it’s larger than the similarity calculated by Nearest Neighbor Method and FSM. The $sim(M_x, M_1)$ is very close while using FSCBR and FSM. And it’s larger than the similarity calculated by Nearest Neighbor Method but smaller than that of Hamming Distance. The $sim(M_x, M_2)$ is almost the same while using FSCBR and Nearest

Neighbor Method. And it's smaller than the similarity calculated by Hamming Distance method and FSM.

TABLE IV. THE GLOBAL SIMILARITY CALCULATED BY FOUR METHODS

	Sim(M_1, M_2)	Sim(M_1, M_3)	Sim(M_2, M_3)	Time/s
the nearest neighbor method	0.305	0.140	0.808	0.62
Hamming distance method	0.729	0.380	0.945	0.54S
FSM presented by Ref. [6]	0.423	0.201	0.861	1.01S
FSCBR presented in this paper	0.421	0.131	0.943	1.09

The reasons caused the limitations of the traditional methods are analyzed as following, (1) The similarity of FL attribute 'High idling speed' is omitted, because requirement 'normal' can't be recognized. (2) For attribute of FN type, the medium value of the closed interval or the border value of the open interval is taken for calculating similarity which leads to inaccuracy of the final result. (3) For FSM, its similarity lacks high accuracy due to following reason: (a) Overlapped area is repeated calculated; (b) attribute weight is simply by using the weights' average. (4) Uncertainty factors are ignored in calculating similarity in traditional methods while FSCBR take it into accounts which result in better similarity accuracy.

IV. CONCLUSION

A case-based reasoning approach based on fuzzy set is presented for remote customization platform of hardware product and impact drill. This approach includes algorithm for generalized membership function, method for center distance correction based on relative area and calculation model for global similarity based on mixed weights. By applying the method, the limitation of the traditional similarity calculation methods which are mostly based on distance functions is to some extent solved, and the accuracy of the similarity is also improved. It provides a new solution to similarity measurement in the application of product customization.

ACKNOWLEDGMENT

The authors wish to acknowledge the Science and Technology Department of Zhejiang Province under Grant No. 2009C31039, No. 2012C01012-4 and No. 2010R50002-001 for their support. And our work is supported in part by Research Foundation and Key Disciplines Foundation of Qianjiang College and Hangzhou city.

REFERENCES

[1] T. Y. Slonim and M. Schneider, "Design issues in fuzzy case based reasoning", *Fuzzy Sets and Systems*, vol. 117, no. 2, pp.251-267, 2011.

[2] J. R. Tan, T. Li and R. Y. Dai, "Research on configuration design system supporting mass customization", *Journal of Computer Aided Design & Computer Graphics*, vol. 15, no. 8, pp. 931-936, 2003.

[3] E. Armengol and E. Plaza, "Symbolic explanation of similarities in case-based reasoning", *Computing and Informatics*, vol. 25, no. 2-3, pp.153-171, 2006.

[4] K. Tahera, R. N. Ibrahim and P. B. Lochert, "A fuzzy logic approach for dealing with qualitative quality characteristics of a process", *Expert Systems with Applications*, vol. 34, no. 4, pp. 2630-2638, 2008.

[5] B. S. Yang, S. K. Jeong and Y. M. Oh, "Case-based reasoning system with Petri nets for induction motor fault diagnosis", *Expert Systems with Applications*, vol. 27, no. 2, pp. 301-311, 2004.

[6] B. S. Zhang and Y. L. Yu, "Hybrid similarity measure for retrieval in case-based reasoning system", *Systems Engineering-theory & Practice*, no.3, pp. 131-136, 2002.

[7] C. Y. Chen, D. S. Zhang and P. H. Ren, "Synthetic similarity measure for case retrieval in case-based reasoning diagnosis system", *Chinese Journal of Mechanical Engineering*, no. 5, pp. 48-52, 2004.

[8] N. Stephane and L. L. J. Marc, "Case-based reasoning for chemical engineering design", *Chemical Engineering Research and Design*, no. 6, pp. 648-658, 2008.

[9] J. J. Li, J. Qi, J. Hu and Y. H. Peng, "Similarity measure method based on membership function and its application", *Application Research of Computers*, no. 3, pp. 891-893, 2010.

[10] T. W. Liao and Z. Zhang. "A review of similarity measures for fuzzy systems", *New Orleans, Louisiana*, pp. 100-106, 1997.

[11] Z. Kowalski, M. Meler-Kapcia and S. Zielinski, "CBR methodology application in an expert system for aided design ship's engine room automation", *Expert Systems with Applications*, vol. 29, no. 2, pp. 256-263, 2005.

[12] K. Y. Ju, D. Q. Zhou and J. M. Wu. "Multi-galois lattice for similarity measurement in case retrieving", *Control and Decision*, no. 7, pp. 987-992, 2010.

Yuhuai Wang received the M.S. degree in Mechatronic Engineering from Zhejiang University of Technology of China, in 2006. He is a PH.D. Candidate in College of Mechanical Engineering, Zhejiang University of Technology. He is currently a Lecturer in Qianjiang College of Hangzhou Normal University. His main research interests are CAD/CAE/CAM/PDM, incremental forming, machine vision and image processing.

Hong Jia received her Ph.D. degree in Chemical Process Equipment from Zhejiang University of Technology of China in 2013. She is currently an Associate Professor in Zhejiang University of Technology. Her major research interests are CAD/CAE/PDM and its integration.

Xiaojing Zhu received the M.S. degree in Mechatronic Engineering from Zhejiang University of Technology of China, in 2009. He is currently a R&D Manager in EP Equipment CO., LTD. His main research interests are CAD/CAE /PDM and Product research and development (PR&D).

Symbol Timing Estimation with Multi-h CPM Signals

Sheng Zhong, Chun Yang, and Jian Zhang

Institute of Electronic Engineering, China Academy of Engineering Physics, Mianyang 621900, China

Email: {zhongsheng0621, ychun507, zjian3000}@163.com

Abstract—A new symbol timing estimation algorithm for Multi-h CPM signals is proposed. It is an extension to an existing non-data aided symbol timing estimator that is only for Single-h CPM signals. A reduced-complexity scheme of proposed algorithm is introduced with negligible performance losses. Performance in AWGN channel is assessed by computer simulation, and the simulation results show that the proposed algorithm is suitable for full response and partial response formats. Meanwhile, it has a good performance and is not sensitive to the carrier frequency.

Index Terms—Symbol Timing Estimation; Multi-h CPM; Non-Data Aided; Reduced-Complexity

I. INTRODUCTION

Continuous phase modulation (CPM) is advantageous for its efficient use of power and bandwidth [1]. Another advantage of CPM is constant envelope, which is not sensitive to nonlinear amplifier and nonlinear channel, and makes it extensive application in the wireless communication system [2] [3]. An interesting type of CPM is Multi-h CPM, which is different from the general single-h CPM, that is multiple modulation index cyclic change. This leads to improving error performance, and making its spectrum more compact, out-of-band rolled down faster. Under the condition of limited bandwidth and power, Multi-h CPM has more excellent transmission performance than single-h CPM. However, it suffers from high implementation complexity and synchronization problems, such as symbol timing recovery. Therefore, how to effectively achieve the Multi-h CPM symbol timing recovery has become the key to Multi-h CPM research. In [4] and [5], a data-aided symbol timing estimate algorithm for single-h CPM were developed using the Fourier series of known training sequence to complete its timing error estimation, but its transmission efficiency was reduced owing to the use of additional training sequence. In [6]-[10], a decision directed symbol timing recovery scheme was proposed based on single-h CPM decomposition representation, but all of them had false locks problems with multilevel and partial response CPM signals. In [11], a symbol timing tracking scheme which extracting timing error form phase was proposed for MCPFSK, but it was only suitable for non-coherent differential receiver system. In [12]-[13] a non-data aided symbol timing scheme was proposed for single-h CPM, but not for Multi-h CPM. In [14], a phase and clock

synchronization scheme was proposed for Multi-h CPM, whose timing error was extracted by calculating the branch metric and its implement complexity was very high.

In this paper, a non-data aided symbol timing estimation algorithm for Multi-h CPM signal is proposed. The symbol timing estimation is acquired, using maximum likelihood principles. Meanwhile, a reduced-complexity scheme for the proposed algorithm is developed that yields an extremely low-complexity version with only negligible performance losses. Both non-reduced complexity scheme and reduced complexity scheme of numerical performance results against the modified Cramer-Rao bound (MCRB) are introduced.

The paper has the following outline. Signal model and basic notations are introduced in the next Section. Section III describes the symbol timing estimation algorithm. Reduced-complexity method for the proposed algorithm is given in Section IV. Analytical methods to assess its performance and simulation results are presented in Section V and finally the conclusions are drawn in Section VI.

II. SIGNAL MODEL

The complex envelope of Multi-h CPM signal

$$v(t) = \sqrt{\frac{E_s}{T}} \exp\{j\varphi(t; \alpha)\} \tag{1}$$

$$\varphi(t; \alpha) = 2\pi \sum_{n=-\infty}^{\infty} \alpha_n h_n q(t-nT) \tag{2}$$

where E_s is the symbol energy, T is the symbol duration, $\alpha = (\alpha_0, \alpha_1, \dots)$ is a sequence of M -ary data symbols.

$\{h_n\}_{n=0}^{N_h-1}$ is a set of N_h modulation index and h_n is unchanged in symbol duration T . The underlined subscript notation in(2) is defined as modulo- N_h , i.e. $\underline{n} \triangleq n \bmod N_h$. Phase response $q(t)$ is defined as the time integral of frequency pulse $f(t)$, which is limited to time interval $(0, LT)$. When $L=1$ the signal is called full-response formats and when $L>1$ the signal is called partial-response formats. Some general pulse shapes are length- LT rectangular (LREC), length- LT raised-cosine (LRC), and Gaussian.

In what follows, we refer to estimated and hypothesized values of a generic quantity x as \hat{x} and \tilde{x} , respectively. Also, \hat{x} and \tilde{x} can assume the same values as x itself. Superscript $*$ denotes the complex conjugate. Notations \otimes , $E(\cdot)$ denotes the convolution, mathematical expectation, respectively.

III. SYMBOL TIMING ESTIMATION

A. The Proposed Timing Estimator

We derive the symbol timing estimator using maximum likelihood principles. The signal observed at the receiver is

$$r(t) = \sqrt{\frac{E_s}{T}} \exp \left\{ j \left[\theta + 2\pi \sum_{n=-\infty}^{\infty} \alpha_n h_n q(t - nT - \tau - \zeta T) \right] \right\} + n(t) \quad (3)$$

where $n(t)$ is complex-valued additive white Gaussian noise(AWGN) with zero mean and power spectral density N_0 . The variables τ and θ represent the symbol timing offset and carrier phase, respectively. ζ is modulation index timing offset parameter which belongs to a discrete finite set $\Gamma = \{0, 1, \dots, N_h - 1\}$. In practice, all of these variables are unknown to the receiver and must be recovered.

Denoting $0 \leq t \leq L_0 T$ as the observation interval and assuming that L_0 is an integer multiple of N_h , the joint likelihood function for $\tilde{\alpha}$, $\tilde{\theta}$, $\tilde{\tau}$ and $\tilde{\zeta}$ is described by

$$\Lambda(r | \tilde{\alpha}, \tilde{\theta}, \tilde{\tau}, \tilde{\zeta}) = \exp \left\{ \frac{1}{N_0} \sqrt{\frac{E_s}{T}} \operatorname{Re} \left[e^{j\tilde{\theta}} \times \int_0^{L_0 T} r(t) e^{-j2\pi \sum_n \tilde{\alpha}_n h_n q(t - nT - \tilde{\tau} - \tilde{\zeta} T)} dt \right] \right\} \quad (4)$$

In order to achieve the final likelihood function $\Lambda(r | \tilde{\tau})$, let us define

$$R = \int_0^{L_0 T} r(t) e^{j\tilde{\theta}} \times e^{-j2\pi \sum_n \tilde{\alpha}_n h_n q(t - nT - \tilde{\tau} - \tilde{\zeta} T)} dt \quad (5)$$

where $R = |R| e^{j\phi_R}$ is complex variables. Substituting (5) into (4) and noting R is independent of $\tilde{\theta}$. Then, the joint likelihood function (4) may be expressed as

$$\Lambda(r | \tilde{\alpha}, \tilde{\theta}, \tilde{\tau}, \tilde{\zeta}) = \exp \left\{ \frac{1}{N_0} \sqrt{\frac{E_s}{T}} |R| \cos(\phi_R - \tilde{\theta}) \right\} \quad (6)$$

Hence, assuming carrier phase $\tilde{\theta}$ uniformly distributed over $[-\pi, \pi]$ [12] and averaging (6) with respect to $\tilde{\theta}$ yields

$$\Lambda(r | \tilde{\alpha}, \tilde{\tau}, \tilde{\zeta}) = I_0 \left(\frac{1}{N_0} \sqrt{\frac{E_s}{T}} |R| \right) \quad (7)$$

where $I_0(x)$ is the modified Bessel function of zero order and in a low SNR it is satisfied with the

approximation $I_0(x) \approx 1 + x^2/4$. Symbol data $\tilde{\alpha}$ is independent of $\tilde{\tau}$ and $\tilde{\zeta}$, averaging (7) with respect to $\tilde{\alpha}$. Then we have

$$\Lambda(r | \tilde{\tau}, \tilde{\zeta}) = E_{\tilde{\alpha}} \left(E_{\tilde{\tau}} \left(|R|^2 \right) \right) \quad (8)$$

Performing some algebraic manipulation and removing all nuisance parameters, the joint likelihood function for $\tilde{\tau}$ and $\tilde{\zeta}$ is described by

$$\Lambda(r | \tilde{\tau}, \tilde{\zeta}) \approx \sum_{k_1=0}^{N_{L_0}-1} \sum_{k_2=0}^{N_{L_0}-1} r(k_1) r^*(k_2) F_{\tilde{\zeta}}[(k_2 - k_1)T_s, k_2 T_s - \tilde{\tau} - \tilde{\zeta} T] \quad (9)$$

where $r(k)$ is samples of $r(t)$ and $T_s = T/N$ is sampling time. The function $F_{\tilde{\zeta}}[\Delta t, t]$ contains the modulation index timing offset parameter and is defined as

$$F_{\tilde{\zeta}}[\Delta t, t] = \prod_{i=-\infty}^{+\infty} \frac{1}{M} \frac{\sin[2\pi h_{i+\tilde{\zeta}} M q(\Delta t, t - iT - \tilde{\zeta} T)]}{\sin[2\pi h_{i+\tilde{\zeta}} q(\Delta t, t - iT - \tilde{\zeta} T)]} \quad (10)$$

$$q(\Delta t, t) = q(t) - q(t - \Delta t) \quad (11)$$

where $F_{\tilde{\zeta}}[\Delta t, t]$ for Multi- h CPM is periodic function of t with period T and, as such, it needs to be computed only on one symbol interval, say $t \in (0, T)$. For $t \in (0, T)$ and $\Delta t > 0$, the only non-unity factors are those with the index in the range $-[L + \text{floor}(\Delta t / t)] \leq i \leq 0$, where $\text{floor}(x)$ means "largest integer not exceeding x ". It is worth noting that $F_{\tilde{\zeta}}[\Delta t, t]$ is calculated by seeking the function limit, when $q(\Delta t, t)$ converges to zero.

Due to various modulation index in equal probability for Multi- h CPM, averaging over $\tilde{\zeta}$ results in

$$\Lambda(r | \tilde{\tau}) \approx \sum_{k_1=0}^{N_{L_0}-1} \sum_{k_2=0}^{N_{L_0}-1} r(k_1) r^*(k_2) \times \left\{ \frac{1}{N_h} \sum_{\tilde{\zeta}=0}^{N_h-1} F_{\tilde{\zeta}}[(k_2 - k_1)T_s, k_2 T_s - \tilde{\tau} - \tilde{\zeta} T] \right\} \quad (12)$$

For convenient analysis, let us define

$$F[\Delta t, t] = \frac{1}{N_h} \sum_{\tilde{\zeta}=0}^{N_h-1} F_{\tilde{\zeta}}[\Delta t, t] \quad (13)$$

Substituting (14) into (12), likelihood functions for symbol timing offset $\tilde{\tau}$ is defined as

$$\Lambda(r | \tilde{\tau}) \approx \sum_{k_1=0}^{N_{L_0}-1} \sum_{k_2=0}^{N_{L_0}-1} r(k_1) r^*(k_2) F[(k_2 - k_1)T_s, k_2 T_s - \tilde{\tau}] \quad (14)$$

Because $F[\Delta t, t]$ is also periodic function of t with period T , its Fourier series expansion can be used in evaluating (14). Likelihood functions for symbol timing offset $\tilde{\tau}$ is defined as

$$\Lambda(r|\tilde{\tau}) = \text{Re} \left\{ \sum_{m=1}^{\infty} A(m) e^{j2\pi m \tilde{\tau}/T} \right\} \quad (15)$$

with

$$A(m) = \sum_{k=0}^{NL_0-1} [r(k) e^{-j\pi m k/N}] y_m^*(k) \quad (16)$$

where

$$y_m(k) = \sum_{i=0}^{NL_0-1} [r(i) e^{j\pi m i/N}] h_m(k-i) \quad (17)$$

$$h_m(k) = \frac{1}{T} e^{j\pi m k/N} \int_0^T F[-kT_s, u] e^{j2\pi m u/T} du \quad (18)$$

The impulse responses $h_1(k)$, which is computed in(18) is shown in Fig.1 for Multi- h CPM signal with $M=4, 1RC$, and $h=(4/16, 5/16)$. For convenience, it can be called as 4M1RC with $h=(4/16, 5/16)$.

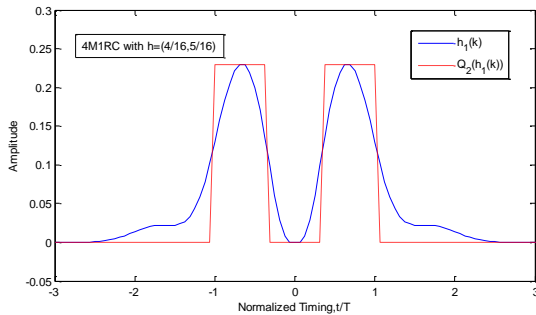


Figure 1. The quantized and unquantized impulse responses of $h_1(k)$ for 4M1RC with $h=(4/16, 5/16)$

The pulses $h_m(k)$ observed here, which is the same as that observed in [12], is that the energy in the pulses $h_m(k)$ decreases rapidly as the Fourier series harmonic index m increases. Therefore, the likelihood function in (15) is well approximated by the single term where $m=1$. Because $h_1(k)$ is also a physical impossibility to achieve non-causal filter. So, following the approach in [12], we shift $h_1(k)$ by ND steps and choose D as the smallest integer such that $h_1(k)$ is confined with the interval $0 \leq k \leq 2ND$. For example, $D=4$ is an appropriate choice for 4M1RC with $h=(4/16, 5/16)$ format. Shifting $h_1(k)$ rightward implies delaying the filter output by the same amount. Hence, (16) and (17) become to

$$y_1(k-ND) = [r(k) e^{j\pi k/N}] \otimes h_1(k-ND) \quad (19)$$

$$A(1) = \sum_{k=ND}^{N(L_0+D)-1} [r(k) e^{j\pi(k-ND)/N}] y_1^*(k-ND) \quad (20)$$

Substituting (19), (20) into (15), and setting likelihood functions equal to zero after taking the partial derivative with respect to $\tilde{\tau}$, the symbol timing offset estimation is given by

$$\hat{\tau} = -\frac{T}{2\pi} \arg \{A(1)\} \quad (21)$$

B. Timing Adjustment

Because the received signal is cyclostationary processes with period T , the feedforward timing recovery schemes, such as proposed algorithm, exhibit jumps of T seconds in passing from one block to the other. If the phenomenon is not properly recognized, it makes BER performance severely deteriorated. To cope with this problem, the timing estimate $\hat{\tau}$ must be unwrapped. Following [12], the unwrapped timing estimate $\hat{\tau}_u$ is obtained by the following equation

$$\hat{\tau}_u^{(l)} = \hat{\tau}_u^{(l-1)} + \text{SAW}(\hat{\tau}^{(l-1)} - \hat{\tau}_u^{(l-1)}) \quad (22)$$

where l denote the l -th symbol timing estimation observation interval. $\text{SAW}(x)$ is a sawtooth function that reduces x to the interval $[-0.5T, 0.5T]$. Formally

$$\text{SAW}(x) = (x + 0.5T)_{\text{mod } T} - 0.5T \quad (23)$$

Once the unwrapped timing estimation are available, in reality, the receiver need to change the A/D sampling clock phase (synchronous sampling recovery) or interpolate the received signal (asynchronous sample recovery) to recover the optimal sampling point. In all-digital receivers, interpolation scheme is usually adopted, which synchronizes the received signal based on an interpolation filter. Due to not changing the clock phase, this scheme makes receivers more stability. Generally, the filter with Farrow structure is usually used for interpolating the received signal, and the output of interpolation filter can be expressed as

$$Z(iT_s) = \sum_k h_l [kT_s + u_k^{(l)} T_s] r[(m_k^{(l)} - k)T_s] \quad (24)$$

where, $l = k \text{ mod } NL_0$, $h_l(t)$ is so-called the filter impulse response, m_k and u_k denotes the basepoint index and fractional interval, respectively, which can be expressed as

$$u_k^{(l)} = \text{frc} \left(u_k^{(l-1)} + \frac{\hat{\tau}_u^{(l)} - \hat{\tau}_u^{(l-1)}}{T_s} \right) \quad (25)$$

$$m_k^{(l)} = m_k^{(l-1)} + \text{int} \left(u_k^{(l-1)} + \frac{\hat{\tau}_u^{(l)} - \hat{\tau}_u^{(l-1)}}{T_s} \right) \quad (26)$$

where $\text{frc}(x) = x - \text{int}(x)$

IV. REDUCED COMPLEXITY

From (19), filter operation determines the complexity of the proposed algorithm. For convenience, the filter is set as a direct FIR structure, whose outputs requires $NL_0(2DN+3)$ complex multiplication and $2DN(NL_0-1)+1$ complex addition per block time. Hence, complexity can be reduced by quantizing impulse response $h_1(k)$ while maintaining its performance. The quantization formula is expressed as

$$Q_l(x(k)) = \text{round}\left(\frac{x(k)2^l}{M_x}\right) \frac{M_x}{2^l}, l > 1 \quad (27)$$

$$M_x = \max(|x(k)|) \quad (28)$$

where $\text{round}(x)$ denotes ‘‘round towards the nearest integer’’, l denotes that l bits used to quantize the input, and the leftmost bit is used as the sign bit. The impulse response and amplitude frequency response of $Q_2(h_1(k))$ for 4M1RC with $h=(4/16,5/16)$ are shown in Fig.1 and Fig.2, respectively. In the time domain, the shape of $Q_2(h_1(k))$ is similar to that of $h_1(k)$. In the frequency domain, the amplitude frequency response of $Q_2(h_1(k))$ remains to be same with that of $h_1(k)$ in spectrum main lobe, although the larger side lobe level leads to poorer characteristic in the stopband, which makes its timing estimation performance little deterioration and is proved in the subsequent simulation results.

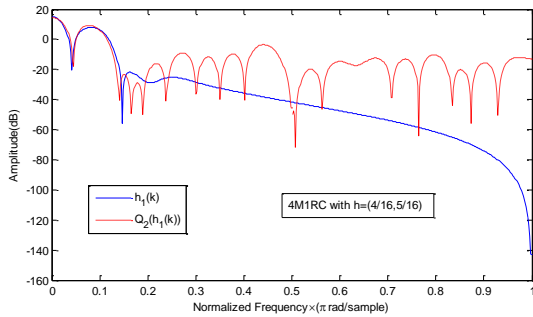


Figure 2. The quantized and unquantized amplitude frequency response of $h_1(k)$ for 4M1RC with $h=(4/16,5/16)$

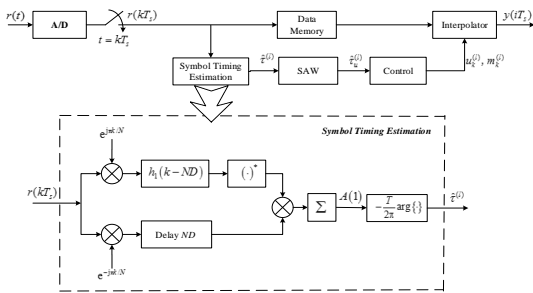


Figure 3. Discrete-time implementation of non-data aided symbol timing recovery system for Multi-h CPM

Multiplication with cumulative addition is simplified to accumulation by quantization $h_1(k)$ to $Q_2(h_1(k))$, making the filter become no multiplication filter. As an example of 4M1RC with $h=(4/16,5/16)$, when $N=4$ and $D=4$, 35840 complex multiplication per block time decreases to zero, but complex addition remains 32737 per block time. Discrete-time implementation of non-data aided symbol timing recovery system for Multi-h CPM is shown in Fig.3. When impulse responses $h_1(k)$ becomes

to $Q_2(h_1(k))$, the system reduces to a reduced-complexity timing recovery version.

V. PERFORMANCE ANALYSIS AND NUMERICAL RESULTS

A. Modified Cramer-Rao Bound

We use the modified Cramer-Rao bound(MCRB) to establish to lower bound on the degree of accuracy to which τ can be estimated for Multi- h CPM. The MCRB for Multi- h CPM was introduced by Perrins *et al.*[6]. Without going into the details of derivations of the bounds, we will apply the MCRB (normalized to the symbol rate) for timing offset estimation using the following formula [12]:

$$\frac{1}{T^2} \times \text{MCRB}(\tau) = \frac{1}{8\pi^2 \bar{h}^2 C_\alpha C_f L_0} \times \frac{1}{E_s / N_0} \quad (29)$$

where $C_\alpha = E\{\alpha_n^2\} = (M^2 - 1)/3$ for uncorrelated M -ary data symbols. For the special case of LREC we have $C_f = C_{LREC} = 1/(4L)$, and for the special case of LRC we have $C_f = C_{LRC} = 3/(8L)$. E_s / N_0 is the SNR, and \bar{h}^2 is the mean-squared modulation index, which is the modification needed to accommodate Multi- h CPM.

B. Numerical Results

The performance of the proposed algorithm for Multi- h CPM signals is evaluated by simulation. In all simulations, format of Multi- h CPM signals is choose by simulation program, the oversampling factor N is chosen to be 4, the block length $L_0=256$ and the channel is additive AWGN channel. The performances of symbol timing offset estimation is determined via simulation by measuring the *normalized timing error variance*

$$\frac{1}{T^2} \times \delta_\tau^2 = \frac{1}{T^2} \times \text{Var}\{\hat{\tau}[n] - \tau\} \quad (30)$$

where the minimum achievable normalized timing error variance is lower-bounded by the MCRB, which is given for Multi- h CPM.

Fig. 4 shows the normalized timing error variance performance for Multi- h CPM signals with same modulation index $h=(4/16,5/16)$. In the figure, it shows that both of the quantized and unquantized scheme have good simulation results for those signals with several different frequency pulse shapes: 1REC, 1RC and 2RC. The performance of 2RC signal is the worst, because $L=2$ brings more inherent symbol interference (ISI). But with the increase of SNR, estimate performance has reached an error platform, mainly by the likelihood function form of this algorithm and the characteristics of Multi- h CPM signals. On the one hand, the Multi- h CPM signals improve the spectrum efficiency by introducing ISI, on the other hand the likelihood function item $\sum_{k_1} \sum_{k_2} r(k_1)r^*(k_2)$ is equivalent to an ISI, which results in a fixed interference. In low SNR, the interference of

noise is dominant, and at high SNR, ISI became the main interference. Because the ISI is fixed and does not reduce with the increase of SNR, estimate performance is almost independent of SNR. Generally speaking, the proposed algorithm can not only provide good performances in full response receiving systems of 1RC and 1REC, but also provide satisfying performances in partial response receiving systems such as 2RC. The figure also shows that the performance of reduced complexity scheme is a little poorer than that of non-reduced complexity scheme owing to the larger side lobe level of the amplitude frequency response $Q_2(h_1(k))$, but it reduces the complexity of the system considerably.

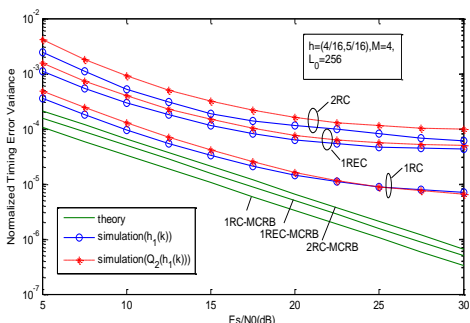


Figure 4. MCRB vs. normalized timing error variance for several formats of Multi-h CPM

Fig. 5 MCRB vs. normalized timing error variance for Multi-h CPM signals, when existing carrier frequency. When normalized carrier frequency $fT=0.01$, both of estimation performance are not worse. Hence, the proposed algorithm is immune to carrier frequency.

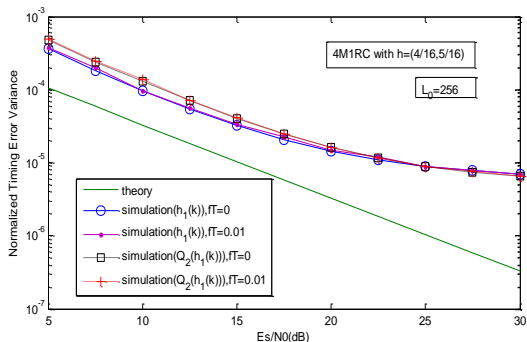


Figure 5. MCRB vs. normalized timing error variance for 4M1RC with $h=(4/16,5/16)$, when existing carrier frequency

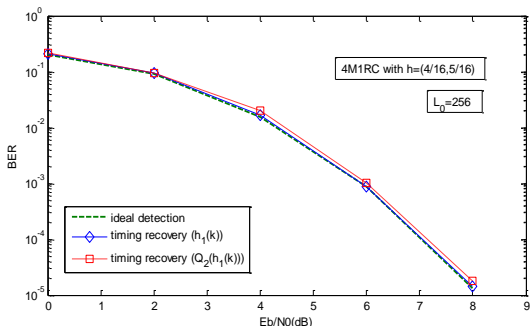


Figure 6. BER for 4M1RC with $h=(4/16,5/16)$ format

Fig. 6 shows the bit error rate (BER) performance for 4M1RC with $h=(4/16,5/16)$ whose timing estimations come from the quantized and unquantized schemes discussed above. Both of them can achieve near ideal-detection BER performance. The quantized scheme has only 0.02dB loss at $BER=10^{-5}$. This demonstrates the usefulness of the quantitation impulse response $h_1(k)$ to $Q_2(h_1(k))$, which effectively reduces the computational complexity of the proposed algorithm.

VI. CONCLUSIONS

A non-data-aided symbol timing estimation for Multi-h CPM has been presented. The proposed algorithm has been derived using maximum-likelihood principles and reduced-complexity method has been introduced. The performance of the proposed algorithm has been assessed in an AWGN environment by computer simulation. Simulation results have shown that the proposed algorithm is suitable for full response and partial response Multi-h CPM and provides good performance while it is not sensitive to carrier frequency. However, the quantization scheme of proposed algorithm has a lower implementation complexity and its performance loss is significantly small, implying lower power consumption, which is a critical requirement for software radio system.

ACKNOWLEDGMENT

This work was supported by National Science Fund of China (Grant No. 10876103).

REFERENCES

- [1] A. Perott and S. Benedett, "Capacity-achieving cpm schemes," *Information Theory, IEEE*, vol. 56, no. 4, pp.1521-1541, 2010.
- [2] H. Zhou and J. Bruck, "Efficient generation of random bits from finite state markov chains," *Information Theory, IEEE*, vol. 58, no. 4, pp. 2490-2506, 2012.
- [3] J. Li, "An estimate algorithm for the snr of continuous phase modulation signals," *Information and Electronic Engineering*, vol. 8, no. 1, pp. 71-76, 2010.
- [4] W. L. Shen, M. j. Zhao, and P. L. Qiu, "Data aided symbol timing estimate in space-time coding systems with continuous phasemodulation," *European Transactions on Telecommunications*, vol. 20, no. 2, pp. 227-232, 2009.
- [5] E. Hosseini and E. Perrins, "Training sequence design for data-aided synchronization of burst-mode cpm," in *Global Communications Conference, 2012 IEEE*, 2012, pp. 2234-2239.
- [6] E. Perrins, E. Bose, and M. P. Wylie-green, "Timing recovery based on the pam representation of cpm," in *Military Communications Conference, 2008 IEEE*, 2008, pp. 1-8.
- [7] Q. Zhao and G. L. Stuber, "Robust time and phase synchronization for continuous phase modulation," *Communication, IEEE*, vol. 54, no. 10, pp. 1857-1869, 2006.
- [8] X. M. Liu, C. Liao, and W. M. R., "Joint timing and phase estimation algorithm for cpm signals," *Computer engineering*, vol. 38, no. 21, pp. 103-106, 2012.
- [9] E. Andrea, Z. Francesca, and G. B., "Spread-spectrum continuous-phase-modulated signals for satellite

- navigation,” *Aerospace and electronic systems, IEEE*, vol. 48, no. 4, pp. 3234–3249, 2012.
- [10] W. Y. Tang and E. Shwedye, “ML estimate of symbol timing and carrier phase for cpm in walsh signal space,” *Communication, IEEE*, vol. 49, no. 6, pp. 969-974, 2002.
- [11] L. Zhong, M. J. Zhao, and J. Zhong, “A symbol timing tracking algorithm for mcpsk,” *Journal of circuits and systems*, vol. 17, no. 3, pp. 1-5, 2012.
- [12] A. N. D’Andrea, U. Menguli, and M. Moreli, “Symbol timing estimate with cpm modulation,” *Communication, IEEE*, vol. 44, no. 10, pp. 1362-1372, 1996.
- [13] W. Y. Tang and E. Shwedyk, “Reduced-complexity nondata-aided timing recovery for pam-based m-ary cpm receivers,” *Communication, IEEE*, vol. 4, no. 6, pp. 87-96, 2012.
- [14] G. V. Kulikov, A. U. Unger, and S. P. G., “Phase and clock synchronization of the viterbi demodulator of continuous phase modulation signals,” *Journal of Communications Technology and Electronics*, vol. 52, no. 6, pp. 656-662, 2011.

Sheng Zhong received the B.S. and M.S. degree in electronic information technology and Circuit and System from Southwest University of Science and Technology, China, in 2004 and 2007, respectively. He is currently working towards his Ph. D. degree in wireless physics at Graduate School, China Academy of Engineering Physics, China. His current research interest includes multi-h CPM technology.

Chun Yang received his B. S. degree in electronic information technology from Shandong University, China in June 1996 and his M.S. degree in wireless physics from Graduate School, China Academy of Engineering Physics, China in June 1997. He is currently working towards his Ph.D. degree in wireless physics at Graduate School, China Academy of Engineering Physics, China. His current research interest includes telemetry technology.

Jian Zhang received his Ph. D degree signal and information processing from University of Electronic Science and Technology of China, China, in 2004. His research interests include telemetry technology and terahertz communications.

Vector-Based Sensitive Information Protecting Scheme in Automatic Trust Negotiation

Jianyun Lei and Yanhong Li*

School of Computer Science, South-Central University for Nationalities, Wuhan, China

*Corresponding Author, Email: lejiaanyun@mail.scuec.edu.cn, anddylee@163.com

Abstract—The existing sensitive information protecting schemes can not satisfy the actual security requirement of some applications. A vector based sensitive information protect scheme is presented based on the existing schemes. One side in trust negotiation can selectively exposes the sensitive attributes to the other side in trust negotiation process based on personal security policy and the trust evaluation result of the other. The implementation process is given in concrete application instances and the scheme is analyzed.

Index Terms—Access Control; Automatic Trust Negotiation; Sensitive Information Protecting; Trust Evaluation

I. INTRODUCTION

Resource sharing is very important in large-scale application systems in distributed multi-domain environments [1-3], and access control technique used in these systems is the key to information security issues. In a distributed multi-domain environment, trust management [4] proposed by Blaze is a relatively mature access control technique. The visitor of the resources must provide his certificates to prove the appropriate access rights, and the owner of resources make the appropriate decision whether to allow access or not based on the certificates provided by the visitor. The automatic trust negotiation [5] proposed by Winsborough is also an access control technique in distributed multi-domain environments. Resources visitor and owner establish trust and make access control decisions through repeated exchanging their certificates without the third-party [6-8].

Certificates in trust management system and trust negotiation always contain some sensitive attributes which are necessary to be protected. However, the trust management system does not take this into consideration. It is a deficiency of trust management system. Many scholars have done a lot of research works in how to protect sensitive attributes of certificates in the process of automatic trust negotiation [9-11]. The existing solutions are not able to fully meet the requirements of the users in some applications. The users can't selectively expose some sensitive attributes of certificates according to their own access control policy and the trust assessment to the others [12-14].

Trust negotiation is a method that establishes trust relationship between entities in distributed domain environment. The entities do not know each other before,

but they establish trust relationship step by step through exchanging digital certificates again and again [15]. A trust negotiation system is consisted of the entities of negotiation, digital certificates and the police of exposing certificates, etc.

Digital certificate is digitalized tool that contains user identification and attribution, according to the different application background; there are identification certificate and attribution certificate. Digital certificate is signed by the issuer, so it has the unforgeability and verifiability [16, 17].

Access control policy is used to ensure the information not be accessed by the illegal users, so its function is to provide all kinds of the access operation to the data source of the legal user [18]. The access control policy in the trust negotiation resolves how to exchange certificates during the negotiation process, which is just the sequence of the exposing of all kinds of certificates [19, 20].

Compared with the access control system based on identification [21], trust negotiation contains the obvious advantages: (1) Two sides of the negotiation do not have to know the identification and attribution each other before, they establish the trust relationship during the process of exchanging the digital certificates, and this is appropriated in the distributed multi-domain system where the entities do know each other. (2) Two sides of the negotiation can define their own access control policy to provide the access to own sensitive resource. (3) There needn't the trusted third party during the trust negotiation.

In the trust management and trust negotiation systems based on certificates, the relative research on the protection of the sensitive attribution include [22], oblivious signature-based envelope [23], hidden credentials [24] and secret handshakes from pairing-based key agreements [25], etc.

In the secret handshakes from CA-oblivious encryption scheme, based on the scheme of zero-knowledge proof protocol, Bob promise to Alice that he contain a certain attribution, Alice and Bob work according to the protocol, Alice sends Bob an envelop, only when the attribution Bob promised satisfies the assert of Alice, Bob can open the envelop, and Alice knows nothing about any attribution of Bob. The theory of secret handshakes from CA-oblivious encryption and oblivious signature-based envelope are almost the same, the difference between them lies in: oblivious signature-based envelope use the

signature of the attribution but not the promise of the attribution.

In the hidden credentials scheme, Alice and Bob exchange some random information, and then Alice uses the random information that provided by Bob and his access control policy to encrypt the information, if Bob has corresponding credential, he can use the random information provided by Alice to decrypt the information. In secret handshakes from pairing-based key agreements scheme, all the users in a group share secret by exchanging information.

All the scheme above can not satisfy the requirement that Alice want to expose part of her sensitive attribution to Bob, and to the certain trust negotiation entity, such as Bob and Charlie, can access the different sensitive attribution in Alice's certificate under the policy that Alice set before. Further more, the schemes above need trusted third party, the safety of the system lies on the trusted third party and the storage also communication cost of the system will increase. A vector based sensitive information protect scheme is presented based on the existing schemes. One side in trust negotiation can selectively exposes the sensitive attributes to the other side in trust negotiation process based on personal security policy and the trust evaluation result of the other, and there needn't any trusted third party.

II. A SIMPLE PROTECTING SCHEME OF SENSITIVE INFORMATION

In a digital certificate, properties can be represented by the ordered pair $\langle attr_name, attr_value \rangle$, where $attr_name$ stands for the property name, and $attr_value$ stands for the property value. If a property is a piece of sensitive information, the property value should be stored in cipher text. For an example, Alice's certificate contains n attributes, attributes names are N_1, N_2, \dots, N_n , the corresponding attribute values are V_1, V_2, \dots, V_n . If there are i ($i \leq n$) attributes (subscript are denoted as j_1, j_2, \dots, j_i) are sensitive information, then the corresponding i properties of the certificate C are stored in cipher text, and the other $n-i$ attributes are stored in plain text. The publisher of certificate C generates the digital signature, and then sends C and decryption key which used to encrypt i sensitive attributes to Alice. The key must be sent through reliable channels, public key system can also be used to ensure the security. When Alice request services or resources from Bob, Alice must submit C to Bob, and selectively expose part or all of the sensitive attributes of C to Bob according to Bob's attributes (or privileges). Alice sends Bob the corresponding decryption keys of the sensitive information which Bob has authority in C through trusted channel. Bob can then get the encrypted property value by using the received decryption key. This scheme of protecting sensitive attributes in certificate has the advantage of being simple and easy to understand, but there are following inadequacies. The simple protecting scheme of sensitive information is shown in Figure 1.

(1) For each sensitive attribute there is no specific trust assessment, Alice identifies which attributes are sensitive

in certificate unilaterally. There is no measure of sensitivity, and no algorithm of exposing which sensitive attributes to Bob according to the specific circumstances of Bob.

(2) There is no specific data structure to indicate which sub-keys of the key K should be sent to Bob when Alice sends certificate to Bob. When Bob receives a certificate from Alice, he won't know which sub-key is for the corresponding properties immediately. If Bob use every sub-key of K to tentatively decrypt every sensitive attribute in certificate C one by one, it will greatly increase time and storage overhead of Bob.

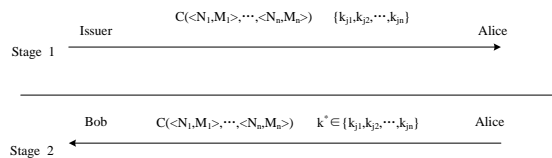


Figure 1. Simple protecting scheme of sensitive information

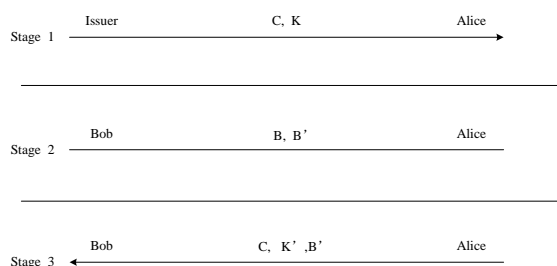


Figure 2. Vector-based sensitive information protect scheme

III. VECTOR-BASED SENSITIVE INFORMATION PROTECTING SCHEME

In order to facilitate formal describing, we use the issuer to represent the certificates publisher; AN and AV to represent attribute name and attribute value, each property has a trust threshold T ($0 \leq T \leq 100$), issuer defines the trust threshold according to the sensitivity of the attributes. Only when the corresponding value of the property on the other side is greater than the trust threshold, the property will be open to him. Obviously, the threshold value of non-sensitive property is 0. A triad $\langle AN, AV, T \rangle$ stands for property name, property value and the trust threshold. $E_k(m)$ represents decrypt information m using key k . C represents the certificate.

Based on the previous simple scheme, with the following adjustments, sensitive information protecting scheme based on vector is presented. The whole process of scheme is divided into three stages: The first is the certificate generation phase, the certificate issuer generates certificate and sends it to Alice; The second stage is trust assessment phase, Alice has a trust evaluation process to Bob according to her own attributes, generates a vector B , and then calculates the open vector B' to Bob according to B and trust threshold of each attribute corresponds in certificate C ; The third stage is the certificate exchange phase, Alice submit the certificate C to Bob, and expose part or all sensitive properties of certificate C to Bob according to B' . The

protocol of vector-based sensitive information protecting scheme is shown in figure 2.

A. Certificate Generation Stage

Certificate issuer generates a certificate C and sends it to the holder of the certificate Alice. There are two major steps in the phase.

(1) The issuer generates a certificate C based on the T value of the properties. In the certificate C, the values whose T = 0 (Non-sensitive properties) are saved in plain text like <AN,AV>. Suppose there are j sensitive properties, For each sensitive property the property value is stored in cipher text like <AN, E_{K_i}(AV)>, i=1..j, The encryption sub-key K_i Here is randomly chosen.

(2) The certificate C and encryption key K(K contains j sub-keys as K_i, i = 1..j) are sent to Alice, and the key is sent through reliable channels.

B. Trust Evaluation Stage

According to the history of certificates exchanging with Bob and other various factors, Alice can give an assessment of trust to Bob for each attribute in the certificate. Suppose there are N properties in Alice’s certificate, and the results of trust assessments to Bob is an N-valued vector containing the trust value, denoted using B, B=<T1,T2,...,TN>, where T_i is the trust value of Alice to Bob on the ith property. Then Alice will compare the trust value of each property in the certificate with B. There will generate a corresponding bit 1 if T_i in B is greater than the trust value of ith property, and that means the property can be disposed to Bob, otherwise there will generate a corresponding bit 0 and that means the property can not be disposed to Bob. A property open vector B’ can be obtained according to the chronological order of these bits. B’ is formed with these bits, and itself can be a binary number that contains these bits.

The xth bit f(x) in open vector B’ can be

$$f(x) = \begin{cases} 1, & \text{The } x^{\text{th}} \text{ property can be disposed} \\ 0, & \text{The } x^{\text{th}} \text{ property can not be disposed} \end{cases}$$

and B’ is shown in figure 3.

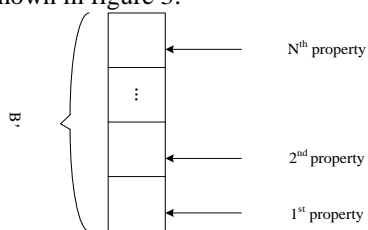


Figure 3. Corresponding relationship between property sensitivity and bits in open vector B’

C. Certificate Exchange Stage

In the certificate exchange stage, Alice sends the certificate C to Bob, and shows Bob about the non-sensitive properties in the certificate and the sensitive properties which are opened to Bob in the following four steps:

(1) Correspondence between the sensitive properties of the certificate C that can be exposed and binary bits B’ is established according to the chronological order.

(2) Alice organizes the properties whose trust threshold values are greater than 0 and the corresponding bits in the B’ is 1 and those corresponding encryption sub-keys K_i, to form a new group key K’.

(3) Alice sends C, K’ and B’ to Bob, where K’ must be sent in trusted Channel.

(4) After receiving C, K’ and B’, Bob decrypts the sensitive properties in C whose corresponding value in B’ is 1 using the K’ as the decryption key.

Subheadings: should be 10 point, italic, left justified, and numbered with letters (A, B, ...), followed by a period, two spaces, and the title using an initial capital letter for each word. The paragraph description of the subheading line should be set for 6 points before and 3 points after.

IV. APPLICATION EXAMPLE AND ANALYSIS

The implementation of the above sensitive information protecting scheme based on vector is presented using a specific application example.

A. Application Example

The government departments issue each person a certificate about the basic personal information, it contains in this certificate the information such as name, age, gender, education, income, department and address. Supposed those fields like "age", "income", "department" and "address" are sensitive information. The corresponding trust value of the sensitive properties of the certificate are (30, 80, 70, 90). When Alice and Bob are in the trust negotiation, through the trust evaluation to Bob, Bob's trust value according to Alice's certificate for the sensitive attribute are (50, 75, 70, 75). It's easy to get B’ = 1010 (binary). Alice Sends B’ and the decryption key of “age” and “department” to Bob in the certificate exchanging stage. After Bob receives C, he can only get the values of non-sensitive properties (e.g. name, gender) and can decrypt the value of opened sensitive properties (age, department), and non-opened sensitive attributes (income, address) can not be to obtained due to the lack of access to the corresponding decryption keys.

B. Analysis of the Security Performance of the Scheme

The security of sensitive information protecting scheme based on vector is mainly reflected in two aspects: The one is whether the eavesdropper is able to get Bob’s key which is used to decrypt the sensitive properties, the other is whether Bob can get the value of non-opened sensitive properties.

Because the key distribution is sent through the reliable channel which is based on other cryptography and security measures, so the eavesdropper can not obtain the corresponding decryption key. The key distribution is based on the selective vector, thus Bob can not get the keys for non-opened sensitive properties.

C. Other Features of the Scheme

View from the time overhead, there exists the trust assessment and computing process of the vector on the basis of the original trust negotiation. The process takes time overhead is fixed. That is, its time complexity is

$O(1)$. And there exist the process of encrypting the sensitive properties in the certificate generation and certificate exchange stages, since the encryption is symmetric, the time cost compared with the signature and asymmetric encryption algorithm in the certificate exchange process is negligible.

View from the storage overhead, an additional storage overhead of vector B, properties open vector B' and decryption key is required in vector-based sensitive information protecting scheme, besides the holder of the certificate C requires the storage overhead of the certificate itself, but this is the basic requirement for the system implementation and it is acceptable.

View from the communication overhead, in vector based sensitive information protecting scheme, the communication overhead in certificate generation stage is the certificate C and the sensitive properties decryption key, and the communication overhead is the certificate and the decryption key of opened sensitive properties also the properties open vector B' in the stage of exchange certificates.

V. CONCLUSION

Compared with the disclosure tree model proposed by Yu [26], inadvertently attribute certificate scheme and inadvertently signed envelope scheme, those models operate the certificate as a whole, expose all the properties' information in the certificate or do not expose any information at all. But the protecting scheme in sensitive properties which proposed in this article classifies the properties. The properties are divided into sensitive properties and non-sensitive properties, and the sensitive properties are divided into opened sensitive properties and non-opened sensitive properties. The scheme can selectively expose all non-sensitive properties and opened sensitive properties, but non-opened sensitive properties are protected. This can meet the needs of practical applications better, and is also more flexible and convenient.

Liao, who proposed SDSA scheme in 2008^[27], can also selectively expose sensitive properties of some or all, but there exists no assessment of trust with each other, and no corresponding data structure to express and store the value of trust either, so he can only simply define which properties are opened and which are not, and lack of maneuverability. This scheme recovers the bug. The exposure of each sensitive attribute is determined by the trust evaluation and sensitive property vector.

ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of Hubei Province, China (No. 2013CFB445).

REFERENCES

- [1] D. Brickley and L. Miller, "FOAF vocabulary specification 0. 91. Namespace Document," Online: <http://xmlns.com/foaf/0.1.>, Nov 2007.
- [2] L. Ding, L. Zhou, T. W. Finin, and A. Joshi, "How the semantic web is being used: An analysis of foaf documents," in *HICSS. IEEE Computer Society*, 2005.
- [3] B. Carminati and E. Ferrari, "Privacy-aware Access Control in Social Networks: Issues and Solutions," in *Privacy and Anonymity in Information Management Systems*, J. Nin and J. Herranz, Eds. Springer, to appear.
- [4] M. Blaze, J. Feigenbaum, J. Lacy. Decentralized Trust Management. In: Proceedings of the 17th Symposium on Security and Privacy. Oakland, California, USA. Los Alamitos: *IEEE CS Press*, 1996, 164-173.
- [5] W. H. Winsborough, K. E. Seamons, V. E. Jones. Automated Trust Negotiation. In: Proceedings of DARPA Information Survivability Conference and Exposition. Hilton Head, South Carolina, Los Alamitos: *IEEE press*, Volume 1, January 2000, 88-102.
- [6] E. Ferrari, A. C. Squicciarini, and E. Bertino, "X-TNL: An XML Language for Trust Negotiations," *4th IEEE Workshop on Policies for Distributed Systems and Networks*, Como, Italy, June 2003.
- [7] W. Nejdl, D. Olmedilla, and M. Winslett, "PeerTrust: Automated Trust Negotiation for Peers on the semantic web," in *Workshop on Secure Data Management in a Connected World (SDM'04)*, Toronto, Canada, Aug. 2004.
- [8] B. Carminati, E. Ferrari, and A. Perego, "Enforcing access control in web-based social networks," *ACM Trans. Inf. Syst. Secur.*, vol. 13, no. 1, 2009.
- [9] F. Bonchi and E. Ferrari, Eds., Privacy-aware Knowledge Discovery: Novel Applications and New Techniques. Chapman and Hall/CRC Press, 2010.
- [10] J. Nin, B. Carminati, E. Ferrari, and V. Torra, "Computing Reputation for Collaborative Private Networks," in *COMPSAC '09: Proceedings of the 2009 33rd Annual IEEE International Computer Software and Applications Conference*, 2009, pp. 246-253.
- [11] T. Y. K. E. Seamons, M. Winslett, "Protecting privacy during on line trust negotiation," in *2nd Workshop on Privacy Enhancing Technologies*, San Francisco, CA, April 2002.
- [12] N. Li and J. C. Mitchell, "Datalog with constraints: A foundation for trust management languages," in *Proceedings of the Fifth International Symposium on Practical Aspects of Declarative Languages*, Jan. 2003.
- [13] K. E. Seamons, M. Winslett, and T. Yu, "Limiting the disclosure of access control policies during automated trust negotiation." in *NDSS*, 2001.
- [14] W. H. Winsborough and N. Li, "Safety in automated trust negotiation." in *IEEE Symposium on Security and Privacy*, 2004, pp. 147-160.
- [15] E. Bertino, E. Ferrari, and A. C. Squicciarini, "Privacy-Preserving Trust Negotiation." *Proceedings of 4th Privacy Enhancing Technologies Workshop*, Toronto, CA, May 2004.
- [16] A. C. Squicciarini, A. Trombetta, and E. Bertino, "Supporting Robust and Secure Interactions in Open Domains through Recovery of Trust Negotiations," in *ICDCS. IEEE Computer Society*, 2007, p. 57.
- [17] A. C. Squicciarini, A. Trombetta, E. Bertino, and S. Braghin, "Identitybased long running negotiations," in *Digital Identity Management*, E. Bertino and K. Takahashi, Eds. ACM, 2008, pp. 97-106.
- [18] A. C. Squicciarini, F. Paci, E. Bertino, A. Trombetta, and S. Braghin, "Group-based negotiations in p2p systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 99, no. PrePrints, 2010.
- [19] S. Braghin, I. Nai Fovino, and A. Trombetta, "Advanced trust negotiations in critical infrastructures," *International Journal of Critical Infrastructures*, vol. 6, no. 3, pp. 225-245, 2010.

- [20] Parikshit N. Mahalle, Bayu Anggorojati, Neeli R. Prasad and Ramjee Prasad, "Identity Establishment and Capability Based Access Control (IECAC) Scheme for Internet of Things," In *IEEE 15th International Symposium on Wireless Personal Multimedia Communications (WPMC - 2012)*, pp. 184-188. Taipei - Taiwan, September 24-27 2012.
- [21] Adjei J. K. and Olesen H., "Keeping Identity Private," In *IEEE Vehicular Technology Magazine*, Volume: 6, Issue: 3, pp: 70-79, September 2011.
- [22] C. Castelluccia, S. Jarecki, G. Tsudik. Secret Handshakes from Ca-oblivious Encryption. In *Advances in Cryptology – ASIACRYPT 2004: 10th International Conference on the Theory and Application of Cryptology and Information Security*. Volume 3329 of *Lecture Notes in Computer Science*, Springer, 2004, 293-307.
- [23] N. Li, W. Du, D. Boneh. Oblivious Signature-Based Envelope. In: *Proceedings of the 22nd ACM Symposium on Principles of Distributed Computing (PODC 2003)*. Boston, Massachusetts, USA, New York: ACM Press, July 2003, 182-189.
- [24] Holt J, Bradshaw R, Seamons K, et. al. Hidden Credentials. 2nd ACM workshop on Privacy in the Electronic Society. Washington DC: ACM Press, 2003, 1-8.
- [25] Balfanz D, Durfee G, Shankar N, et. al. Secret Handshakes from Pairing-Based Key Agreements. *Proceedings of the 2003 IEEE Symposium on Secret and Privacy*. Oakland CA, 2003, 80-196.
- [26] T. Yu, M. Winslett, K. E. Seamons. Supporting Structured Credentials and Sensitive Policies through Interoperable Strategies for Automated Trust Negotiation. *ACM Transactions on Information and System Security (TISSEC)*, February 2003, 6(1) pp. 1-42.
- [27] Junguo Liao, Fan Hong, Jun Li et. al. Keeping confidentiality of sensitive attributes in credential during trust negotiation. *Chinese Journal of Communications*. 2008, 29(6) pp. 20-25.



Jianyun Lei, born in September, 1972, in Zhejiang province, China, received the B.E degree in computer and application from South-Central University for Nationalities (SCUFN) in 1994, the M.S degree in computer software and theory from Huazhong University of Science and Technology (HUST) in 2004, and the Ph.D. degree in information security from Huazhong University of Science and Technology (HUST) in 2010.

Since 1994, he has been a faculty in South-Central University for Nationalities (SCUFN) which is in Wuhan, China, and he is currently an Associate Professor in the school of computer science. His recent research interests include information security, internet of things (IOT) etc.

Dr Lei is the member of China Computer Federation (CCF).



Yanhong Li was born on 26 April 1973 in Hunan, China. received the B.E. degree in mechanical engineering from Central South University of China(CSU) in 1993, The M.S. degree in computer application from Chongqing University of China (CQU)in 2004, and the Ph. D degree in computer software and theory from Huazhong University of Science and Technology of China (HUST) in 2011.

Since 2012, she has been a faculty in South Central University for Nationalities (SCUFN) which is in Wuhan, China, and she is currently a lecturer in the school of computer science. Her research interests include information security and multimedia network communication technology.

An Improved Byzantine Fault-tolerant Program for WSNs

Yi Tian

Shangluo University, Shangluo, China

Abstract—In order to increase the level of fault-tolerance of wireless sensor networks, thus to enhance the reliability and accuracy, we studies the traditional byzantine fault-tolerant program, and makes improvements in the environment of WSNs, reducing the number of rounds of message exchange between network nodes, thus improving the efficiency and reducing communication overhead and energy consumption. The simulation result shows that our program makes all normal network nodes reach an agreement, while the number of rounds of message exchange greatly decreases compared to the traditional byzantine program.

Index Terms—WSNs; Fault Tolerance; Byzantine Fault-Tolerant Program; M-tree

I. INTRODUCTION

Wireless sensor networks (WSNs) are a distributed self-governance measurement and control network system composed of a lot of tiny sensors with communication and computation abilities. WSNs can make people get a large quantity of detailed and reliable information at any time, at any place, or under any condition; therefore, it is widely applied in national defense and military, environmental monitoring, traffic management, medical treatment and public health, and other fields.

Different application fields have different security requirements for WSNs. In most non-commercial applications such as environmental monitoring and air humidity, the security problem is not very important. However, in other fields such as monitoring the enemy's military deployment in the area occupied by the enemy in terms of military, the things varying from data acquisition, data transmission to physical distribution of points have to be kept confidential to the enemy. In those security-sensitive applications, if the networks node in WSNs is captured or energy of the network node is exhausted, the consequence may be disastrous [1] [2] [3]. How to ensure the reliability of whole networks and the security of data transmission becomes the important content of security research on WSNs.

The traditional passive defense method such as encryption and identity authentication, can produce effective defense function on the attack from external part of WSNs. For example, it is unable to discover the originally normal network nodes controlled by the enemy by use of encryption and identity authentication, for there is private key in those network nodes. It is shown through previous research that there still exist loopholes in some

network nodes which are vulnerable to the attack no matter how many security defense measures are taken. At the same time, those abovementioned defense measures basically can't confront the security threat from the internal part, such as network node failure [4] [5]. Therefore, the fault-tolerant program becomes the second barrier to guarantee the security.

Due to various limitations (such as computing power and storage space) on WSNs, the traditional fault-tolerant program can't be directly applied in WSNs; therefore, the fault-tolerant program must meet the specific application demand of WSNs. We hereby suggest designing a fault-tolerant program for WSNs by use of the advantages of the byzantine fault-tolerant program.

The byzantine fault-tolerant program originated from the byzantine agreement (BA) problem which was put forward by Lamport et al. in 1982 [6]. This problem is defined by Lamport [6] as follows.

- (1) The messages sent out from network node can be correctly delivered.
- (2) The receiver node knows the sender node who sends this message.
- (3) When the number of network nodes is n , the number of network nodes in which the error happens will be $(n-1)/3$ at most.

Based on the above-mentioned hypothesis, BA problem can be solved [7], and the steps of solution are generally shown as below.

- (1) Any network node is selected as the source network node, and then the source network node sends its initial value $v(s)$ to all other network nodes.
- (2) All normal network nodes make use of the message received from the source network node through exchange with other network nodes to check whether other normal network nodes have received the same value.

This solution can meet following requirements.

- (1) A consensus can be reached among all normal network nodes.
- (2) If the source network node is normal, the common value for which a consensus is reached among all normal network nodes shall be same with the initial value of the source network node.

Currently, the research on applicability of byzantine fault-tolerant program is mainly in the field of distributed computing system and wireless mesh networks [8]-[14]. Wang [14] et al. establishes a fault-tolerant network structure for wireless mesh networks according to traditional byzantine fault-tolerant program and improves

routing algorithm, thus improving the reliability of networks. However, for large number of rounds of message exchange between network nodes, it only establishes minimum message exchange unit from fixed topology for control. This is not applicable to WSNs with topological dynamic changes. Klempous [15] et al. applied traditional byzantine fault-tolerant program into WSNs and considered that the performance of fault-tolerance depended on the scope of message exchange between network nodes. If the scope of message exchange is too large, the large number of rounds of message exchange between network nodes will cause burden to communication; if the scope of message exchange is too small, normal network nodes cannot reach a consensus comprehensively, thus causing misinformation.

In view of this, we find that the major problem of the application of the byzantine fault-tolerant program into WSNs is that the number of rounds of message exchange between network nodes is too great. Generally, when the total number of network nodes is n and the maximum number of error network nodes is $(n-1)/3$, the number of rounds of message exchange required by traditional program is $(n-1)/3+1$ and the amount of message exchange in the i th round ($i=1\dots(n-1)/3+1$) is $(n-1)(n-2)\dots(n-i)$. Which causes a huge communication overhead for WSNs with a large number of network nodes and meanwhile causes huge consumption for sensor with limited energy.

In the research on the traditional byzantine program, two features as follows are noticed: 1) the number of normal network nodes is at least $n-(n-1)/3$. 2) The total number of messages received from normal network nodes is always higher than that of messages received from error network nodes. If the abovementioned 2 points can be fully utilized, the efficiency of the traditional fault-tolerant program can be effectively improved, that is, the energy consumption and communication traffic can be reduced through reducing number of rounds of message exchange between network nodes so that such program can be applicable to WSNs. In this paper, we propose an improved byzantine fault-tolerant program.

The remainder of this paper is organized as follows. Section II describes the proposed fault-tolerant program. The simulations and their results are analyzed in Section III. Finally, our conclusions and future work are presented in Section IV.

II. IMPROVED BYZANTINE FAULT-TOLERANT PROGRAM

Firstly, the hypothesis of the proposed program is given.

- (1) The messages sent out from network node can be correctly delivered.
- (2) The receiver node knows the sender node who sends this message.
- (3) When the number of network nodes is n , the number of network nodes in which the error happens will be $(n-1)/3$ at most.

- (4) In the first round of message exchange, only one network node can send messages.

Furthermore, all error network nodes have abnormal behaviors, that is, the values sent by error network node to other network nodes are different or there is difference between normal value and value sent by error network node.

A. The M-tree Used by Fault-Tolerant Program

In the process of message exchange, each network node uses a specific tree to store messages received from other network nodes after message exchange. This tree is named as M-tree (message tree). Fig. 1 shows its structure. First, the source network node sends its original value $v(s)$ to other network nodes and itself. As the sender of message is recognizable, all normal network nodes can recognize that this message comes from the source network node and store this value in the root node of M-tree. However, this cannot judge whether the source network node is normal network node. Therefore, multi-round message exchange between network nodes is required so as to eliminate the influence of error source node. Starting from the second level of M-tree, each node is named with the sender of storing value and father node (the second level is not named with root node). For example, if the storing value of a node in the second level comes from network node x , this node is named as $v(x)$; if the father node of a node in the third level is $v(x)$ and its storing value comes from network node y , this node is named as $v(xy)$; the rest can be done in the same manner. Nodes with continuously repetitive network nodes in the name (such as $v(xx)$, $v(yy)$ and $v(xyy)$) are not stored in M-tree so as to prevent error network nodes from sending error messages periodically which makes error messages stored in M-tree repeatedly and causes misjudgment.

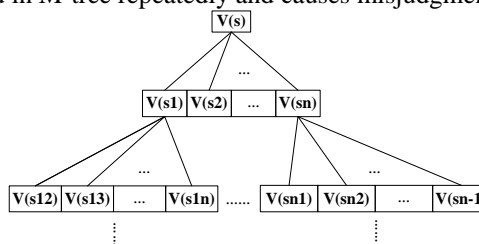


Figure 1. Structure of M-tree

B. Decision Process of Trusted Network Nodes

For all network nodes, the decision process of trusted network nodes can be implemented after three rounds of message exchange. First, it is necessary to judge whether the storing value of each node (marked as $v(cx)$) in the second bottom level of M-tree is equal to the value in the majority (expressed as $maj(cx)$) among all its corresponding child node values or not. If it is, it is required to judge whether the total number of child nodes (expressed as $cou_maj(cx)$) with the same value of $maj(cx)$ of each $v(cx)$ is not less than $(n-(n-1)/3-1)$, i.e. judge whether the total number of network nodes with the majority is not less than the minimum total number of normal network nodes. If both of abovementioned conditions are satisfied, the network nodes that sends the

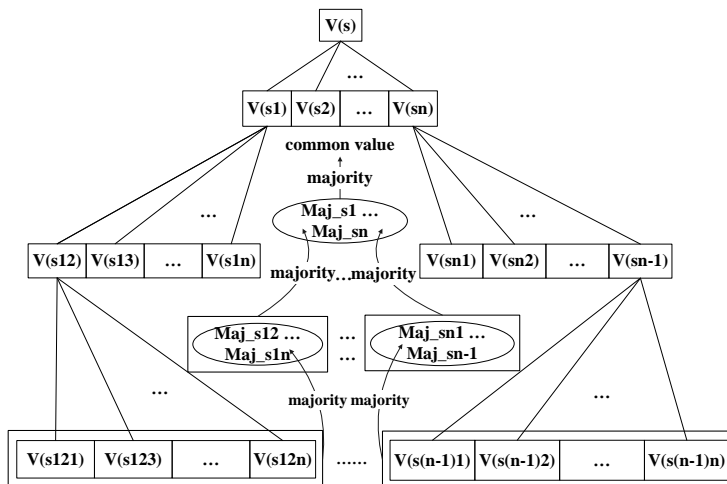


Figure 2. The process of getting common value

same value with $maj(cx)$ to the node $v(x)$ and its subtree are added into the reliable-like network point set (RNPS, recorded as $RNPS(cx)$). In general, judging whether a network node can be added into $RNPS(cx)$ shall meet the following 2 conditions:

- (1) $v(cx) = maj(cx)$
- (2) $cou_maj(cx) \geq (n-(n-1)/3-1)$

Then, the total number of each network node in all RNPS in the second bottom level (expressed as $cou_RNP(z)$) is counted. If $cou_RNP(z)$ of a network node is not less than $(n-(n-1)/3-1)$, this network node is defined as trusted network node.

C. Correction Process of Error Message

After new trusted network node is added, each network node will correct messages from untrusted network node in M-tree to accelerate the election of common value. Subtrees of all nodes in the second bottom level of M-tree will be examined. If the storing value of its child node is different from the value in the majority (expressed as $maj(RNP(cx))$) among values sent by current trusted network node and the sender node of this value is not trusted network node, the value of child node will be replaced as $maj(RNP(cx))$.

D. Election Process of Common Value

After all error messages are replaced, M-tree selects majority values from each subtree in the second bottom level and then selects majority values towards the upper level till the root. Then, the common value of this network node is obtained. The example diagram is shown in Fig. 2.

E. Overall Operational Process of Fault-Tolerant Program

According to the design above, the minimum number of rounds of message exchange is 5. Starting from the 4th round, if no new trusted network node is found in each network node for two consecutive rounds, i.e. all normal network nodes reach a consensus, the operation of the program ends. Fig. 6 shows the details of fault-tolerant program. The main steps is described as below.

Step 1: First, necessary message exchange phase. In the first round of message exchange, the source network node sends its original value $v(s)$ to itself and other network nodes and each network node stores this value in root node of M-tree. Then, go to step 2.

Step 2: In the 2nd round of message exchange, each network node sends root node value in M-tree to itself and all other network nodes and each network node stores the received value in the corresponding node of the second level of M-tree. Then, go to step 3.

Step 3: In the 3rd round of message exchange, each network node sends the storing value in the second level of M-tree to itself and all other network nodes and each network node stores the received value in the corresponding node of the third level of M-tree. Then, go to step 4.

Step 4: After the 3rd round of message exchange, each network node enters decision-making phase and the decision process of trusted network node starts. After the decision process, the condition that whether the cumulative variable null message exchange ($cou_nullmess$) is less than 2 needs to be judge and the initial value of $cou_nullmess$ is 0. If judgment is false, the operation of the program ends. If judgment is true, then judge whether new trusted network node is found. If the result is true, $cou_nullmess$ returns to 0, go to step 5; otherwise, $cou_nullmess$ pluses 1, start the next round of message exchange and the new decision-making phase.

Step 5: After new trusted network node is added, each network node starts the correction process of error message and replace the message sent by untrusted network node in M-tree. Then, go to step 6.

Step 6: Each network node elects common value and the common value of all normal network nodes is obtained. Then start the next round of message exchange and the new decision-making phase.

F. Example of Applying Fault-Tolerant Program

In this section, An example is given to explain how the proposed fault-tolerant program can make all normal network nodes reach an agreement. We assume in a seven-node WSNs, network nodes are marked as a, b, c, d, e, f, g. Node a is defined as source network node. To

check the performance of the proposed program, we design the worst-case scenario, that is Node c and Node e are both error network node. The messaging behavior of the error network nodes is shown in Table 1.

TABLE I. THE MESSAGING BEHAVIOR OF THE ERROR NETWORK NODES

	a	b	d	f	g
a	2	2	2	2	2
c	2	0	0	0	1
e	0	1	0	3	1

The beginning of the program is the necessary message exchange phase, the source network node a sends its initial value to itself and all other network nodes in the first round of the message exchange. Because the source network node a is normal network node, the sending values are the same and the value is 2. Then, each network node stores this value $v(s)$ in the root of M-tree, as shown in Fig. 3. Here, the results of error network nodes do not need to be discussed, thus this example only shows the results of normal network nodes.

- Node a, $V(a) = 2$
- Node b, $V(a) = 2$
- Node d, $V(a) = 2$
- Node f, $V(a) = 2$
- Node g, $V(a) = 2$

Figure 3. The value stored in each normal network node's M-tree

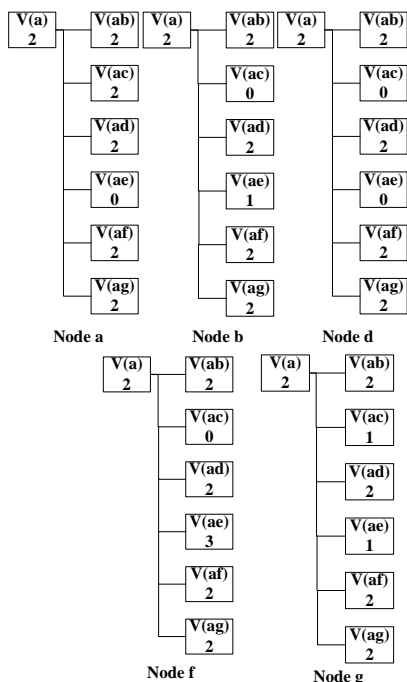


Figure 4. The result of storing values in each normal network node's M-tree after the second round of message exchange

In the 2nd round of message exchange, each network node sends the root value in M-tree to itself and all other network nodes. Similarly, each network node stores the received value in the corresponding node of the second

level of M-tree. The results for the second round of message exchange are shown in Fig. 4.

In the 3rd round of message exchange, each network node sends the storing value in the second level of M-tree to itself and all other network nodes and each network node stores the received value in the corresponding node of the third level of M-tree. The results of all normal network nodes' M-trees are shown in Fig. 5.

After the 3rd round of message exchange, the program accesses the Decision-making phase, and the first process of this phase is the decision process of trusted network. The process specifics of Node d is shown in Fig. 7. For example, the Node a, b, d, f, g should be added in RNPS(ab), that the following conditions are satisfied:

- (1) $v(ab) = \text{maj}(ab)$
- (2) $\text{cou_maj}(ab) \geq (n-(n-1)/3-1)$
- (3) $v(abx) = \text{maj}(ab)\{\text{such as, } v(aba), v(abd), v(abf), v(abg) = \text{maj}(ab) = 2\}$.

Next, the program counts the number ($\text{cou_RNP}(z)$) of each network node occurring in all RNPS in the second bottom level of M-tree and computes whether $\text{cou_RNP}(z)$ is not less than $(n-(n-1)/3-1)$. For example, Node b is appeared in RNPS(ab), RNPS(ac), RNPS(ad), RNPS(af), RNPS(ag). Obviously, $\text{cou_RNP}(b)$ is 5, which is greater than $(n-(n-1)/3-1) = 4$. Hence, Node d is defined as trusted network node. Correspondingly, Node a, d, f, g are also trusted network node.

The program checks whether the cumulative variable null message exchange (cou_nullmess) is less than 2. If the result is false, the program ends. On the contrary, if new trusted network node is found and added, the correction process of error message can be executed, and cou_nullmess returns to 0. If no new trusted network node is found and added, cou_nullmess pluses 1, and the program starts the next round of message exchange and the new decision-making phase. In this example, Node a, b, d, f, g are all new trusted network node and cou_nullmess remains initial state, the value is 0, so the program runs the correction process.

In this process, the values received from the untrusted network nodes in each subtree of the second bottom level of M-tree must be replaced by the majority value of the trusted network node in each subtree. For example, as shown in Fig. 8, Node c and e are untrusted network nodes in the subtree of $v(ab)$. Furthermore, $v(abc)$ and $v(abe)$ is not equal to the majority value of the trusted network node set of $v(ab)\{v(aba), v(abd), v(abf), v(abg)\}$. Thus the value of $v(abc)$ and $v(abe)$ must be replaced by 2 ($\text{maj}(\text{RNP}(ab)) = 2$).

Finally, the majority values from each subtree in the second bottom level of M-tree can be elected. Then the election process can be executed upwards layer by layer till the root. So far, the decision-making phase has been completed, the program starts the new round of message exchange and repeats the decision-making phase. In this example, all the normal network nodes have reached a agreement in the 3th round of message exchange. So the total number of rounds of message exchange used by this example is 5. This value is equal to the minimum design number of rounds.

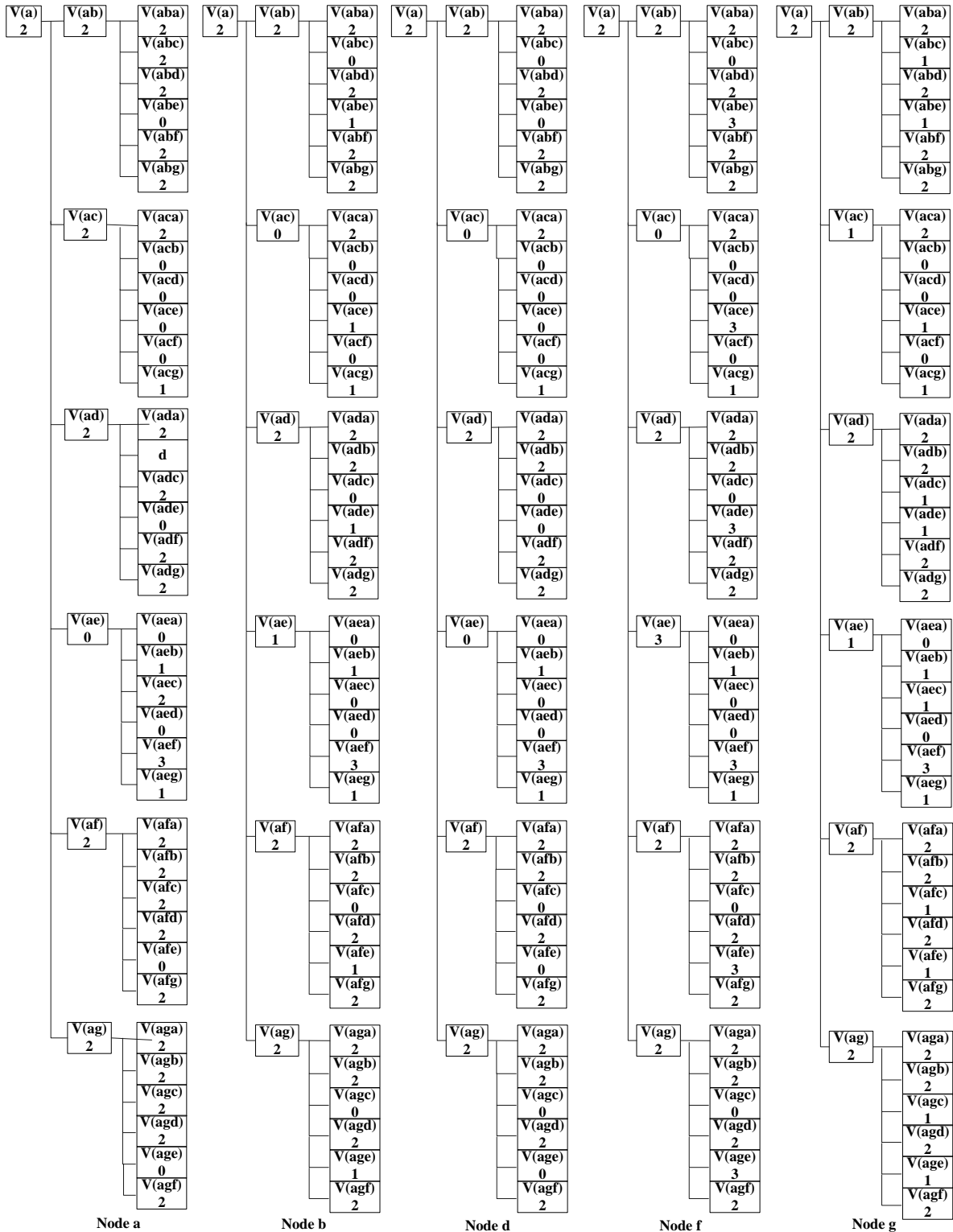


Figure 5. The result of M-tree's storing values in network node a, b, d, f, g after the third round of message exchange

III. SIMULATIONS

To study the performance of the proposed fault-tolerant program and the traditional byzantine fault-tolerant program in the number of rounds of message exchange, NS2 is used as basic simulation platform. The basic setting of simulation scene: 80 network nodes are distributed randomly in 500m × 500m area. The covering

radius of sensor is 100m. There are 2 independent variables in simulation scene: the number of error network nodes {5, 10, 15, 20, 25} and whether error network node is the source network node. There are 10 scenes in total and each scene is simulated for 10 times. Each simulation ends till the completion of fault-tolerant program. The final data is the average value of ten simulations.

```

Byzantine fault-tolerant program

Definition
n: The number of network nodes in WSNs.
v(cx): The node in M-tree.
maj(cx): The value in the majority among values stored in child nodes of node v(cx) in M-tree.
cou_maj(cx): The frequency of values stored in child nodes of node v(cx) in M-tree equal to maj(cx).
RNPS(cx): The reliable-like network point set of node v(cx) in M-tree.
cou_RNP(z): The total number of network node z occurring in all reliable-like network point sets in certain level of M-tree.
RNP(cx): The trusted network node set of node v(cx) in M-tree.
maj(RNP(cx)): The value in the majority among values stored in trusted network node of node v(cx) in M-tree.
cou_nullmess: The total number of rounds of message exchange without new trusted node found, and its initial value is 0.

Necessary message exchange phase
The 1st round
The source network node sends its original value v(s) to itself and other network nodes and each node stores this value in root node of M-tree.
The 2nd round
Each network node sends root node value in M-tree to itself and all other network nodes and each node stores the received values in the second level of M-tree.
The 3rd round
Each network node sends the storing value in the second level of M-tree to itself and all other network nodes and each node stores the received values in the third level of M-tree.

Decision-making phase
//Decision process of trusted network node
foreach all nodes in certain level of M-tree
{
If(v(cx) = maj(cx) and cou_maj(cx) >= (n-(n-1)/3-1)) then
{
Add network node x to RNPS(cx);
foreach child node of v(cx)
{
If (v(cxy) = v(cx)) then
{
Add network node y to RNPS(cx);
}
}
}
}
foreach network node z of this network
{
Count cou_RNP(z) in all RNPS in this level of M-tree;
If (cou_RNP(z) >= (n-(n-1)/3-1)) then
{
Network node z is defined as trusted network node;
}
}
//Judge whether to stop the program
If (cou_nullmess < 2) then
{
If (new trusted network node is added in RNP(cx)), then
{
cou_nullmess = 0;
//Correction process of error message
foreach all nodes in certain level of M-tree
{
foreach child node of v(cx)
{
If (the sender of v(cxy) is not in RNP(cx) and v(cxy) ≠ maj(RNP(cx))), then
{
v(cxy) = maj(RNP(cx));
}
}
}
}
//Election process of common value
Elect majority value level by level starting from the subtree of node in the second bottom level of M-tree and obtain common value;
}
else
{
cou_nullmess = cou_nullmess + 1;
}
//Start the new round of message exchange
Each network node sends the value stored in the bottom level of M-tree to itself and all other network nodes;
Each node stores the received values in the new bottom level of M-tree;
//Start decision-making phase
Goto decision-making phase;
}
    
```

Figure 6. Improved byzantine fault-tolerant program

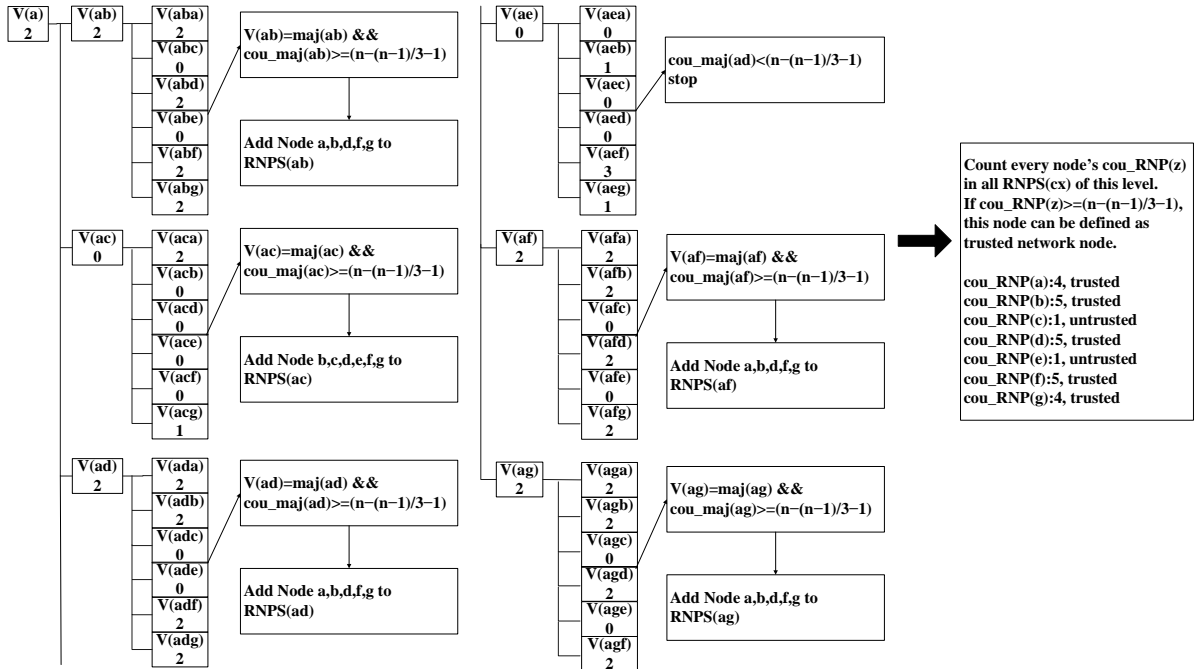


Figure 7. The decision process of trusted network (for Node d)

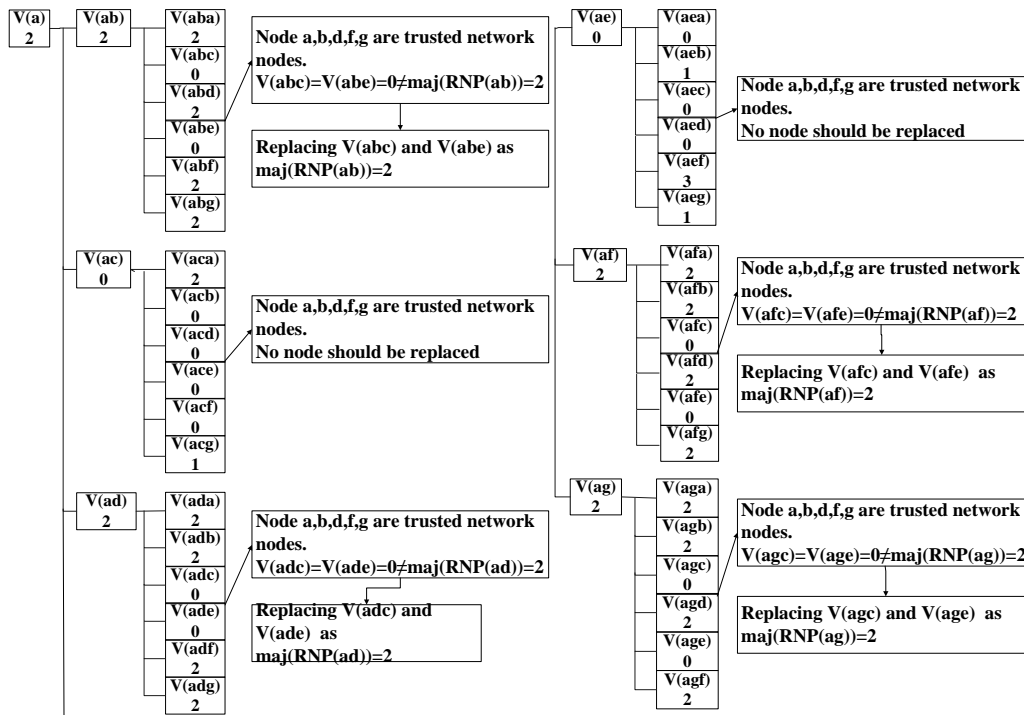


Figure 8. Correction process of error message (for Node d)

Fig. 9 shows the comparison of the number of rounds of message exchange in the proposed program and the traditional byzantine program when the source network node is not error network node. Compared to traditional byzantine program, the proposed program has obvious advantages. When the number of error network nodes is small, the number of rounds of message exchange in the proposed program is always the minimum design number of rounds. It slightly increases when the number of error network nodes approaches the maximum value of

network design. This is because that network nodes with the normal value in the whole network are always in the majority, i.e. this normal value is always in the majority in M-tree of each network node, thus making the final common value equal to this value. It is unnecessary to find error nodes to determine the final common value through at least $(n-1)/3 + 1$ rounds of message exchange in traditional byzantine program, thus reducing the number of rounds of message exchange.

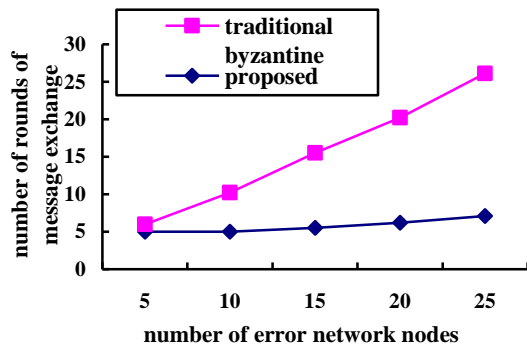


Figure 9. Comparison of the number of rounds of message exchange in the proposed program and traditional byzantine program (source network node is not error network node)

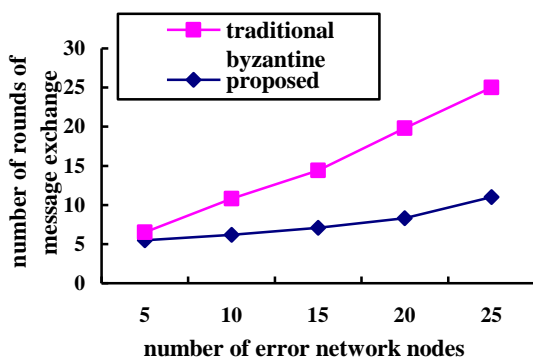


Figure 10. Comparison of the number of rounds of message exchange in the proposed program and traditional byzantine program (source network node is error network node)

Fig. 10 shows the comparison of the number of rounds of message exchange in the proposed program and the traditional byzantine program when the source network node is error network node. Compared to Fig. 9, in the proposed program, the number of rounds of message exchange required when the source network node is error network node is generally higher than that when the source network node is not error network node under the condition of same number of error network nodes, because when the source network node is error network node, the original value $v(s)$ sent to all network nodes will be different, which increases the difficulty of determining common value of normal network nodes. The number of rounds of message exchange increases correspondingly. However, compared to traditional byzantine program, the number of rounds of message exchange required in the proposed program is still small.

IV. CONCLUSIONS

We study and improve the traditional byzantine fault-tolerant program and design a fault-tolerant program applied in WSNs. The proposed program uses normal network nodes in the network as measurement standard of common value, which not only ensures the reliability and correctness of the whole network, but also solves the problem of huge energy consumption and communication overhead caused by large number of rounds of message exchange in traditional byzantine fault-tolerant program. The subsequent simulation proves that the number of

rounds of message exchange in the proposed program decreases greatly compared to traditional byzantine fault-tolerant program whether or not source network node is error network node. In the future research work, the real-time problem of the proposed program will be considered. We will optimize the code and process of the program, thus making the time spend in reaching all normal network nodes a consensus as short as possible.

ACKNOWLEDGMENT

This work is supported by a grant from Shanxi's Department of Education fund (2013JK1160)

REFERENCES

- [1] Y. Zhou, Y. Fang, Y. Zhang, "Securing wireless sensor networks: a survey," *Communications Surveys & Tutorials, IEEE*, vol. 10, no. 3, pp.6-28, 2008.
- [2] I. Bekmezci, F. Alagöz, "Energy efficient, delay sensitive, fault tolerant wireless sensor network for military monitoring." *International Journal of Distributed Sensor Networks*, vol. 5, no. 6, pp.729-747, 2009.
- [3] X. Chen, K. Makki, K. Yen, N. Pissinou, "Sensor network security: a survey." *Communications Surveys & Tutorials, IEEE*, vol. 11, no. 2, pp.52-73, 2009.
- [4] J. Yick, B. Mukherjee, D. Ghosal, "Wireless sensor network survey." *Computer networks*, vol. 52, no. 12, pp.2292-2330, 2008.
- [5] T. Kavitha, D. Sridharan, "Security vulnerabilities in wireless sensor networks: A survey." *Journal of information Assurance and Security*, vol. 5, no. 1, pp.31-44, 2010.
- [6] L. Lamport, R. Shostak, M. Pease, "The Byzantine generals problem." *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 4, no. 3, pp.382-401, 1982.
- [7] M. Correia, N. A. Bessani, P. Veríssimo, "On Byzantine generals with alternative plans." *Journal of Parallel and Distributed Computing*, vol. 68, no. 9, pp.1291-1296, 2008.
- [8] M. B. Kapron, D. Kempe, V. King, J. Saia, V. Sanwalani, "Fast asynchronous byzantine agreement and leader election with full information." *ACM Transactions on Algorithms (TALG)*, vol. 6, no. 4, pp.68, 2010.
- [9] V. King, J. Saia, "Breaking the $O(n^2)$ bit barrier: scalable byzantine agreement with an adaptive adversary." *Journal of the ACM (JACM)*, vol. 58, no. 4, pp.18, 2011.
- [10] M. A. AlZain, B. Soh, E. Pardede, "A Survey on Data Security Issues in Cloud Computing: From Single to Multi-Clouds" *Journal of Software*, vol. 8, no. 5, pp. 1068-1078, 2013.
- [11] L. Zhang, Q. Zhu, A. Chen, "Fast Message Dissemination Tree and Balanced Data Collection Tree for Wireless Sensor Network" *Journal of Software*, Vol. 8, No. 6, pp. 1346-1352, 2013.
- [12] S. S. Wang, Q. K. Yan, C. S. Wang, "Achieving efficient agreement within a dual-failure cloud-computing environment." *Expert Systems with Applications*, vol. 38, no. 1, pp.906-915, 2011.
- [13] L. M. Chiang, "Eventually Byzantine Agreement on CDS-based mobile ad hoc network" *Ad Hoc Networks*, vol. 10, no. 3, pp.388-400, 2012.
- [14] J. Wang, Y. Zhang, J. Wu, "Byzantine Fault Tolerant Network Structure and Algorithm in WMN." *Computer Engineering*, vol. 37, no. 20, pp.83-86, 2011.

- [15] R. Klempous, J. Nikodem, L. Radosz, N. Raus, "Byzantine Algorithms in Wireless Sensors Network" in *Information and Automation, ICIA 2006*, pp.319-324, 2006

Yi Tian received his M.S. degrees in 2011, from Northwest University of Information Technology. He is now an lecturer in Shangluo University. His research interests include computer system architecture, internet of things.

EESA Algorithm in Wireless Sensor Networks

Zhang Pei and Feng Lu

Xi'an University of Architecture and Technology, Xi'an, ShanXi, China

Abstract—Since there are many problems of traditional extended clustering algorithm in wireless sensor network like short extended time, over energy consumption, too many deviated position the of cluster head nodes and so on, this paper proposes the EESA algorithm. The algorithm makes many improvements on the way of dividing clusters, strategy of electing the cluster head and construction method of data relay path, the two aspects of inter-cluster energy balance and energy balance among the cluster are taken into account at the same time. Detailed simulation results are taken in this thesis to compare network lifetime, average residual energy, energy consumption standard deviation of cluster head node and changes of average remaining energy between the EESA algorithm and ACT algorithm, EECA algorithm and MR-LEACH algorithm; the simulation results show that: the proposed algorithm reduces the load of hot regional cluster head, balances the energy consumption of the entire network nodes and extends the networks lifetime of wireless sensor.

Index Terms—Distributed Management; Energy Consumption; Cluster Head; Residual Energy

I. INTRODUCTION

In recent years, wireless sensor network is widely used in network manufacturing remote monitoring and other areas, showing a great value has caused many countries military, industry and academia of great concern. Since 2000, the international community has been found in some of the wireless sensor network research reports; Natural Science Foundation of the United States in 2003 developed a wireless sensor network research program in supporting the underlying theory [1-2]. Natural Science Foundation of the United States, driven by the U.S. University of California, Berkeley, Cornell University and other research institutions began a wireless sensor network theory and key technology research. U.S. Department of Defense and the military departments attach great importance to the wireless sensor networks, regarded as an important area of research, and the establishment of a series of military sensor network research projects. From the emergence of wireless sensor networks so far has been developed from the initial node, network protocol design, developed to a smart group research phase and abroad has become a hot new IT technologies, attracting a large number of scholars have launched a wide range of research, and have made some progress (including a large number of nodes in a large number of platforms and communication protocols). However, has not yet formed a complete system of theory and technology to support the development of this emerging field, there are numerous

issues to be science and technology breakthroughs, the information field is facing a very challenging task [3].

Wireless sensor networks (WSN) has been widely used in civil and military fields, such as smart home, bio-medical, environmental monitoring, machinery manufacturing and space exploration and so on. In recent years, with the sensor node integration and miniaturization, the node's power limited battery energy is mostly used [4]. As nodes are usually located in uninhabited areas or hazardous environments, it is difficult to carry out energy supplies, while renewable energy technology is not yet mature, so by optimizing energy consumption, maximizing network lifetime of wireless sensor networks becomes primary challenge.

In the application of wireless sensor network, the location information is an indispensable part the node data collection; monitoring information without location information is usually meaningless. Determine the event occurs the location or acquiring data of the node position yes wireless sensor network the most basic one of the functions. In order to providing effective position information, the random layout of the sensor nodes in a network deployment after completion is able to determine their own location. Because node exists to limited resources, randomly deployed, communications vulnerable to environmental disturb or even node failure and other characteristics, wireless sensor network's node localization mechanisms must satisfy the self-organization Xing, robustness, energy efficient, distributed computing and other requirements.

Clustering for wireless sensor networks provides an efficient distributed management framework, and its advantage is easing of management, enhanced network scalability, and easy to implement data aggregation. In combination ambush mode, a network connection between the cluster head backbone of the data within the cluster aggregation processing will relay it to the base station [5-7]. However, from the base station closer to the cluster head node relay task often due to overweight and premature deaths, resulting in the "energy hole" appears. LEACH (low-energy adaptive clustering hierarchy) is the first to propose clustering protocol, using equal probability random cluster head selection round robin fashion. But LEACH does not consider the following questions: 1) All cluster heads are to transfer data directly to the base station, no data transfer phase for good optimization; 2) Cluster heads are chosen randomly, can not guarantee its uniform distribution in the network. Since people in the LEACH protocol based on improved to further reduce energy consumption MR-LEACH

(Combination ambush routing with low energy adaptive clustering hierarchy) is a typical improved protocol [8-10]. MR-LEACH based on the residual energy of the node to select the cluster head, while using combination ambush relay transmission mode data to the base station. However, the algorithm does not consider the establishment of relay path all cluster heads load balanced, high-rise cluster head based only on the base station broadcast control packets to determine which relay node [11-13]. This causes some nodes premature deaths due to excessive consumption, thereby affecting the network lifetime. C. Sha other design an energy efficient clustering algorithm EECA (energy efficient clustering algorithm), by optimizing the use of cluster head election strategy and build value function to determine the right way to relay path prolong the network lifetime [14-17]. However, this method can not relay path establishment alleviate hot spots cluster head load "energy hole" appears still affect the network lifetime. W.K. Lai proposed an ACT (Arranging cluster sizes and transmission ranges for wireless sensor networks) protocol, the cluster head by calculating the amount of energy consumed to determine the cluster radius, then the ideal location for election cluster head node [17-19]. Meanwhile, according to the principle of the relay load equal allocation construct data aggregation tree. However, the following deficiencies exist ACT: 1) With time, the cluster head node will gradually deviate from the ideal position, can not continue to maintain the optimality; 2) From the base station closer to the cluster head still bear heavier loads, the "energy hole" the problem is not solved.

In this paper, the inherent characteristics of the wireless sensor networks have the advantages and disadvantages of the above algorithm, an improved energy efficient data aggregation algorithm EESA (Energy-efficient separating algorithm). This paper divided in cluster mode, the cluster head election strategy and data relay path construction method has been improved, taking into account inter-cluster energy balance and energy balance the two aspects of the cluster. New algorithm not only reduces the load on the cluster head hot spots and more balanced energy consumption of the whole network nodes, thereby prolonging the network lifetime.

This paper mainly made in the following areas to expand and innovative work:

(a) Extension of traditional clustering algorithm for wireless sensor networks to extend a short time, energy consumption and the cluster head node is too large too many other issues out of between clusters based on the basic clustering in the way of improvements. Energy consumption according to the network topology and the cluster radius is calculated. Separated from the task of cluster head election strategy point of view has been improved. A single cluster head node to complete the task assigned to two clusters in order to achieve the energy balance.

(b) To further validate the proposed EESA algorithm correctness and validity of the EESA algorithm and ACT,

EECA and MR-LEACH algorithm in the network lifetime, the average residual energy of cluster head node energy consumption standard deviation and average remaining energy changes in other aspects of the simulation were compared in detail experiments using network simulator Omnet + + proceed. MAC layer assumes the ideal and the communication link is correct. Experiments carried out in the steady state. The simulation results show that: EESA algorithm can effectively avoid the energy hole problem and reduce the energy consumption of the entire sensor network, thereby prolonging the network lifetime.

II. ENERGY MODEL

Consider an area of $L \times W$ rectangular sensor network (L and W denote the length and width of the sensing region), N evenly distributed sensor nodes in the area.

A. Hypothesis

(a) Base station and the location of sensor nodes are fixed. All nodes are uniformly distributed density ρ in this region, and through the exchange of information to identify itself and a base station location.

(b) Each node has a unique identifier (ID). The same initial energy of all nodes and transmission are energy controllable. The highest energy level state, the node may communicate directly with the base station.

(c) Links are symmetrical. If the transmission power is known, the node according to the received signals strength to estimate the distance to another node.

(d) All nodes with the same fixed rate sensing and continuously send data to the user, which is equal to the length of each data packet.

B. Energy Model

Energy consumption is calculated using the classical model. A 1-bit message transmission the energy consumption of the distance d :

$$E_t(r, s) = \begin{cases} 1 \times E_{dect} + 1 \times \epsilon_{f_s} \times d^2, & d < d_0 \\ 1 \times E_{dect} + 1 \times \epsilon_{amp} \times d^5, & d \geq d_0 \end{cases} \quad (1)$$

where in, E_{elect} means to transmit or receive data bit energy consumption; f_s and amp denote the sender and the receiver determined by the distance between the amplification different 1-bit data of the energy consumption If the distance is less than a threshold value d_0 , using free-space model (f_s); Otherwise, the multipath model (mp).

Receiving the message the energy consumption:

$$E_r(l) = 1 \times E_{dect} \quad (2)$$

Polymerizing the energy consumed messages:

$$E_b(r, l) = 1 \times E_{agg} \quad (3)$$

where in, E_{agg} polymerized for 1-bit data represents the energy consumed. EMAX represents a threshold residual energy. If a node residual energy is less than this value, it

will give up running cluster head node. According to equation (1) - (3) can be obtained EMAX values:

$$E_{max} = [(\beta \times \gamma E_{dect} + (\lambda - 1)E_{agg} + (1 - \theta)(\lambda + 1)(E_{dect} + \epsilon_{fs}d^2)] \quad (4)$$

where in said number of times to obtain data for each round; θ indicates the aggregation data compression ratio; d represents the current processing node and its next hop distance between nodes; λ represents the number of neighbors of a node.

III. EESA ALGORITHM

Various extension cycle operations by wireless sensor network life cycle saving algorithm are essentially minimize system energy loss, while the energy consumption evenly distributed to each node. In clustering wireless sensor networks, energy imbalance is mainly reflected in two aspects: First, the use of combination ambush transmission mode, different cluster head to the base station due to the different distances ranging from energy consumption, namely the inter-cluster energy imbalance; Second, the cluster head because of the need to complete more than the cluster members work and consume more energy, that energy is not balanced cluster problem. But it is just an ordinary cluster head sensor nodes, limited energy, the cluster head will inevitably lead to premature deaths network lifetime shortened. Therefore, to reduce the energy consumption of the cluster head node, ensure that the entire network node energy balance is the key to extend the network lifetime. Therefore, this paper made some improvements mainly from the energy balance of departure.

A. Improved Clustering Methods

The basic division of cluster clustering algorithm does not consider the energy consumption can be divided into homogeneous and heterogeneous clustering into two categories. EESA energy consumption from the perspective of the clustering method has been improved.

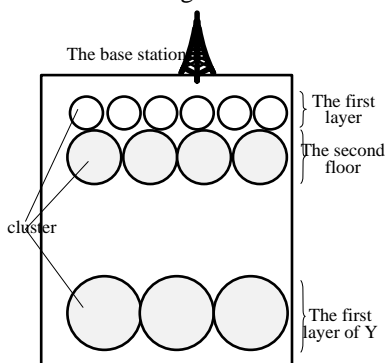


Figure 1. A hierarchical network topology

Y assumes an experimental network has clusters of various sizes layer (1) and the members of each cluster to the cluster head node transfer 1 bit of data. To facilitate the calculation, we will be in two different levels of the transmission distance between clusters as the sum of their radii (except for the first layer). That is, Y to Y-1 layer transmission distance layer $(r_y + r_{y-1})$, Y-1 Y-2 layer to

the distance of the transmission layer $(r_y - 1 + r_{y-2})$, etc. (Figure 2). Since the outermost layer (first layer Y) relay cluster head node is not only responsible for handling this task cluster node transmits data, it is total energy consumption can be expressed as:

$$\pi t_y^2 r E_{elect} + \pi t_y^2 r E_{agg} + (1 - \phi) \pi t_y^2 r [E_{elect} + \epsilon_{fs} (t_y + f_{y-1})^2] \quad (5)$$

The first represents the cluster head node to receive the members of the cluster node to send energy consumption data; second represents the energy consumption for data aggregation; third portion represents the processed data transmitted from the Y layer to the first layer, Y-1 energy consumption. Since the first Y-1 layer only to consider the cluster head node within the cluster node transmits the data, but also on the trunk of the Y layer data processing, their total energy consumption can be expressed as:

$$\pi t_{y-1}^2 r E_{elect} + [\pi t_{y-1}^2 t + (1 - \theta) \pi r_y^2] E_{agg} + [(1 - \theta) \pi t_{y-1}^2 t + (1 - \theta)^2] [E_{elect} + \epsilon_{fs} (t_{y-1} + t_{y-2})^2] \quad (6)$$

Similarly, the first Y-2 layer to simultaneously cluster head node of the cluster and the relay Y-1 of the first layer of the data processing, the total energy consumption can be expressed as:

$$\pi t_{y-2}^2 r E_{elect} + [\pi t_{y-2}^2 t + (1 - \theta) \pi r_y^2] E_{agg} + [(1 - \theta) \pi t_{y-1}^2 t + (1 - \theta)^2] [E_{elect} + \epsilon_{fs} (t_{y-2} + t_{y-3})^2] \quad (7)$$

In this way, the cluster head node of each layer of the total energy consumption can be calculated by the following formula:

$$\left\{ \begin{array}{l} \pi t_y^2 r E_{elect} + \pi t_y^2 r E_{agg} + (1 - \phi) \pi t_y^2 r [E_{elect} + \epsilon_{fs} (t_y + f_{y-1})^2] \\ \pi t_{y-1}^2 r E_{elect} + [\pi t_{y-1}^2 t + (1 - \theta) \pi r_y^2] E_{agg} + [(1 - \theta) \pi t_{y-1}^2 t + (1 - \theta)^2] [E_{elect} + \epsilon_{fs} (t_{y-1} + t_{y-2})^2] \\ \vdots \\ \pi t_{y-2}^2 r E_{elect} + [\pi t_{y-2}^2 t + (1 - \theta) \pi r_y^2] E_{agg} + [(1 - \theta) \pi t_{y-1}^2 t + (1 - \theta)^2] [E_{elect} + \epsilon_{fs} (t_{y-2} + t_{y-3})^2] \end{array} \right. \quad (8)$$

where r_1, r_2, \dots, r_y , respectively from the first layer, the second layer, the Y-radius of the cluster layer; r represents the first layer to the base station transmission distance (see Equation (12)); $E_i (1 \leq i \leq y)$ represents the i layer cluster head node energy consumption.

As each layer is assumed cluster head energy consumption are almost equal, we can use the following

formula to calculate the radius of each layer of the cluster:

$$\begin{cases} E_1 \cong E_2 \cong E_Y \cong E_y \\ t_1 + t_2 + \dots + t_y = \frac{u}{2} \end{cases} \quad (9)$$

According to equation (8) - (9), the calculated radius of all clusters, the cluster division is completed.

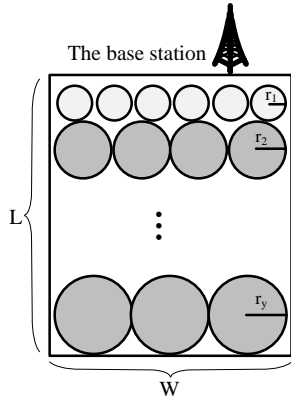


Figure 2. Calculating the radius of FIG Group 2

B. Improved Cluster Head Election Strategy

In order to reduce the load on the cluster head, unlike the previous single cluster head election strategy, EESA elections simultaneously two different nodes - processing nodes and forwarding node as the cluster head node (Figure 3). By separating the task, the task of a single cluster head node is assigned to two finish to reduce cluster head nodes and the cluster members within the energy poor. After the election, combined with the 3.1 calculated cluster radius, the completion of the formation of clusters.

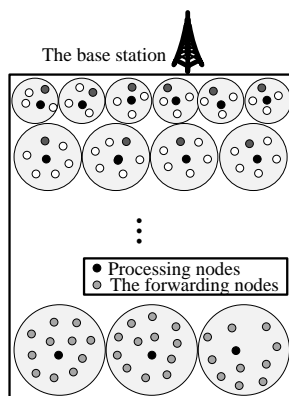


Figure 3. Family head node election

1) Handling and Forwarding Node Election

In the beginning of each round, each node in the communication within a radius of broadcast basic message $E-M_{sg}(IDs, e_{rs}, L(x_s, y_s))$. E_{rs} and $L(x_s, y_s)$ denote the residual energy of the node s and location. All at the broadcast source within a radius of nodes are considered neighbors, and after receiving the broadcast message updates its neighbor table. Each of the remaining energy is greater than $EMAX$ nodes have the

opportunity to participate in the election and become the candidate nodes, assuming that the proportion of the candidate node p . Case in layer j ($1 \leq j \leq Y$), cluster head election process is as follows:

Each candidate node s will be based on the updated information table to calculate the communication of all neighbors within a radius of the average residual energy E_{res} :

$$E_{res} = \sum_{r=1}^{\lambda} \frac{E_{rt}}{\gamma} \quad (10)$$

S and its neighbors average communication distance between nodes can be expressed as:

$$t_s = \sum_{r=1, r \neq m}^{\lambda} \frac{t_{mr}}{\lambda} \quad (11)$$

D_{sm} to represent a node distance between s and m , then:

$$D_{sm} = \sqrt{(x_s - x_m)^2 + (t_r - t_j)^2} \quad (12)$$

Taking into account the residual energy and distances were applied each candidate node equations (13) and (14) to calculate the processing nodes and forwarding node competitive index value:

$$PR_y^x = \sigma \frac{E_{rs}}{E_{res}} + \lambda \frac{1}{t_m} \quad (13)$$

$$PR_y^t = \lambda \frac{E_{rs}}{E_{res}} + \lambda \frac{t_l + t_{l-1}}{r(t, nv)} \quad (14)$$

where μ and the value are determined by the distribution of nodes within a cluster and the residual energy situation decision, $d(s, BS)$ indicates that the node s and the distance between the base station. Then, the candidate node in the communication message broadcasting competition radius:

$$Com_Pr o(ID_s, E_{rs}, CLP_s, CLF_s)$$

Broadcast is completed, all of the candidate nodes are converted to receive state and waiting time T . Set the minimum length of T to ensure that all nodes can be received from the neighbor nodes competing messages. After each candidate node it and the receives index value comparison competition. CLP node with the largest value will be successfully elected to processing nodes, and the maximum value of the CLF has been elected as a forwarding node. If two or CLF CLP equal the highest index value, with the larger residual energy of nodes elected. If a candidate node CLP and CLF also have the highest index value, which will act as both the processing nodes and forwarding nodes both roles.

2) Cluster Formation

According to result of the comparison processing nodes run successfully in its communication radius node broadcasts the campaign success message $S_{uc} - p_{ro}(IDs, E_{rs}, L(x_s, y_s))$, the remaining nodes then waits to receive the message. Once a candidate node receives one or more

of the campaign a success message, it will give up to continue to compete and send join messages $Join_P_{ro}$ (IDs, E_{rs} , $L(x_s, y_s)$) to the maximum received signal strength broadcast nodes. Because only responsible for forwarding node forwards the processed data, so no election broadcast to its entire neighbor success message. It will only win the election messages to join the cluster data packet message, and sends it to the same cluster of processing nodes (forwarding node and the processing node if the same node, this step is omitted). Therefore, compared with the existing algorithms, this paper proposes a new algorithm does not increase the load more messages. Thereafter, the processing node for all cluster members assigned TDMA slots.

3) Data Aggregation Tree Construction

In the data aggregation tree construction process, EESA select from the current forwarding node nearest two lower processing nodes as a candidate relay node and the residual energy of candidate nodes and forwarding this data needs to compare the energy consumption, according to the comparison results to determine relay node. With the first layer and the second layer q, p ($1 \leq p < q \leq y$) as an example, assume that a cluster layer q, u , it is forwarded to the p -layer node selected as a relay node of a processing node. Assume that v and p layers cluster u, w is from two recent clusters. According to equation (1), $CLFu$ packet can be calculated are sent to it and $CLPw, CLPv$ consumed energy difference:

$$\Delta E_t = \varepsilon_{fs} * t_r * (t_v^2 - t_y^2) \tag{15}$$

The formula represents clusters u, k_u amount of data sent. Also, calculate the remaining nodes $CLPv$ and $CLPw$ energy difference:

$$\Delta E = E_v - E_w \tag{16}$$

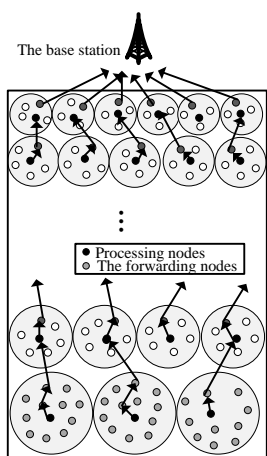


Figure 4. Family first node of transmission mode

Compared with The ΔE and ΔET , if the value is greater than $\Delta E, \Delta ET$, the selected relay node as CPv ; otherwise, select CPw as a relay node. That is, the farther away from the current candidate node forwarding node only if it is much larger than the residual energy from the current candidate node forwarding node closer,

it can be selected. Forwarding node of each cluster in the same manner will find their relay node, the network backbone is formed (Figure 4). This strategy not only considers the residual energy of cluster head nodes, taking into account the balance of the total energy consumption of the network, thus greatly improving the energy utilization.

IV. SIMULATION AND ANALYSIS

Experiment EESA with ACT, EECA and MR-LEACH algorithm in the network lifetime, the average residual energy of cluster head node energy consumption standard deviation and average remaining energy changes and other aspects were compared. All nodes in the sensing range of each other to ensure that the ongoing transmission is not destroyed by other nodes.

A. Simulation Environment

The simulation results by the network emulator O_{nmet}^{++} experimentally derived. MAC layer assumes the ideal and the communication link is correct. Experiments carried out in the steady state, simulation parameters as shown in Table 1.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Node Number	99
Network Range (Square Meter)	(41.52)
Initial Energy Of Nodes (Joule)	2.2
The Candidate Notes Proportion p	0.32
E_{elec} (10^{-5} Joule / Bit)	51
ε_{fs} (10^{-4} Joule / Bit / Square Meter)	10
d_o / Meter	102
E_{agg} (10^{-6} Joule / Bit / Meter / sign)	6
Data gram Length (Bit)	502

B. Experimental Results and Analysis

1) Network Life

Several different algorithms applied network life curves shown in Figure 5. As can be seen from the figure, the use of network EESA has the longest life. MR-LEACH based only on the residual energy of the node to select cluster head, EECA cluster head at the same time in the choice of the reference node residual energy and communication distance and other factors. However, since no combination ambush communication between the two modes of energy consumption imbalance cluster head to make improvements, some cluster head node premature death led to the shortening of the life of the network. ACT according to energy consumption for each cluster head to be adjusted for the size of the cluster, but there are also energy whole problem. Meanwhile, as time progresses, the resulting cluster head node election will gradually deviates from the ideal position so that the cluster head and the cluster members difference between a rapid increase in energy consumption. The EESA calculating the energy consumption among clusters determines the cluster, while using mission critical node separation method to reduce energy consumption, thereby

prolonging the network lifetime. Thus, EESA algorithm can balance between the energy consumption of the node, to improve the lifetime of sensor networks is important.

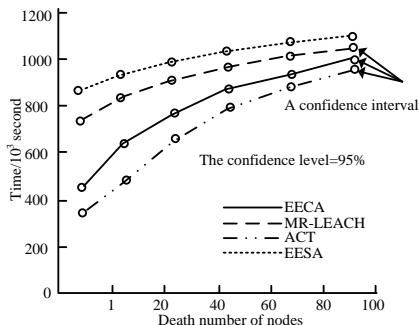


Figure 5. Network Life

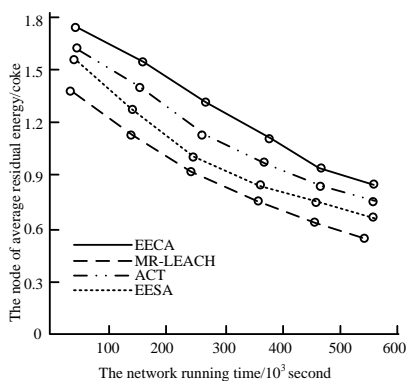


Figure 6. The Average remaining energy

2) Average Remaining Energy

Figure 6 shows the four algorithms average remaining energy of the nodes. Obviously, EESA algorithm is better control of energy consumption of nodes. MR-LEACH, EECA and ACT have adopted a combination ambush communication model, the residual energy difference lies in their cluster head election and data aggregation tree was constructed in different ways. In each one cluster head election, MR-LEACH directly specify the maximum residual energy of cluster head node by the base station to determine the next hop for each cluster head node. EECA cluster head in the election at the same time taking into account the remaining energy and communication distance on two factors, and by constructing an association formed by these two factors, the weights for each cluster head function to select the optimal relay node. ACT according to each cluster head to calculate the amount of energy cluster radius, then the election closest to the ideal location node as the cluster head. Meanwhile, according to the principle of the relay load it has the equal allocation to construct data aggregation tree. Thus, MR-LEACH due to the lack of full consideration of factors which led to excessive energy consumption; EECA no right to make any improvements hot issues, energy hole still exists; while the ACT will be running late because the selected cluster head nodes far from an ideal location resulting increase in energy consumption. EESA improved from three hot issues: First, according to the cluster head to calculate the total energy equal to the radius of the cluster, reaching the inter-cluster energy

balance; Second, while elections two nodes are used to complete the work of a single cluster head, By separating the task slowed key nodes within the cluster energy consumption; Third, by calculating the difference $\Delta E - \Delta ET$ to construct data aggregation tree, to achieve a balance of total energy consumption across the network. Therefore, EESA has the highest energy utility and maximum residual energy.

3) The Cluster Head Node Energy Consumption Standard Deviation

Shown in Figure 7, the cluster head election and bear the load of different cluster head node energy consumption standard deviation is also showing irregular shocks. In the cluster head election process, MR-LEACH only consider the node residual energy, EECA algorithm design also refer to the node's residual energy and communication distance of these two factors. In the combination ambush communication mode, each cluster head due to load a different type of energy consumed. Therefore, MR-LEACH and EECA algorithm curve fluctuations, the cluster head node energy consumption are not balanced. ACT cluster head according to the energy consumption to calculate the cluster radius, to a certain extent, reduce the hot issue. Meanwhile, post-maintenance using cross-layer data transfer mode also allows different levels of cluster head is located has the similar energy consumption. Therefore, ACT algorithms curve is relatively stable. As can be seen from the figure, EESA has the most stable curve algorithm. This is because the EESA using load balancing cluster head idea to reduce hot spots, while the task of separating the cluster head reduces the energy consumption of key nodes within the cluster, avoiding the energy difference between the larger nodes.

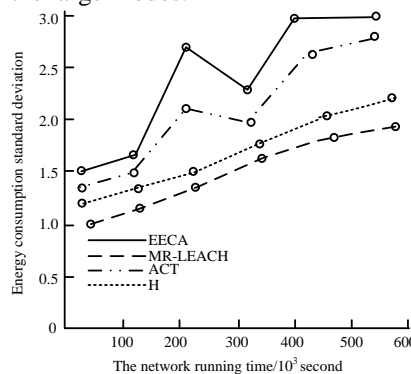


Figure 7. Cluster leader energy consumption standard deviation

4) Average Remaining Energy Changes

Figure 8 shows the average of the remaining nodes in the network the energy curve. In this paper, the average residual energy of nodes to measure the change in the equilibrium level of consumption of several algorithms. The smaller the average remaining energy changes that the more balanced energy consumption. As can be seen from Figure 8, with the other three algorithms, EESA average remaining energy change is small, that is to get more balanced energy consumption. This is because, when using EECA, ACT, and MR-LEACH three algorithms, the "energy hole" problem evident from the

base station closer to the cluster head node often due to overloading of death before the other nodes, resulting in the average residual energy curve shocks. The EESA from both inter-cluster and cluster perspective consider the energy balance, effectively easing the “energy whole” effect and the average change in the minimum residual energy.

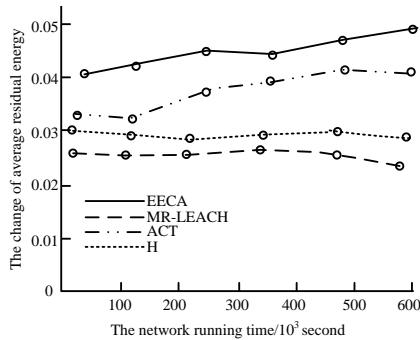


Figure 8. Average remaining energy changes

V. CONCLUSION

Reduce energy consumption of wireless sensor networks is a fundamental problem. In this paper, clustering in wireless sensor network energy consumption imbalance problem, we propose a new and improved algorithm EESA. The algorithm respectively between the clusters and the cluster from the perspective of energy balance considerations, the basic methods and clustering in cluster head election strategy was improved. Energy consumption according to the network topology and the way to solve the clustered combination ambush mode cluster head due to the load caused by the different amount of energy consumption of different problems to achieve a balanced energy consumption among clusters. Meanwhile, the task separation method effectively reduces the energy consumption of key nodes within the cluster, making all nodes in the network tends to balance energy consumption. Simulation results show that the algorithm can effectively improve the energy efficiency of WSN and extend the network lifetime.

REFERENCES

[1] Xin Huang, Xiao Ma, Bangdao Chen, Andrew Markham, Qinghua Wang, Andrew William Roscoe. Human Interactive Secure ID Management in Body Sensor Networks. *Journal of Networks*, Vol 7, No 9 (2012), 1400-1406

[2] Hong Sun; De Florio, V.; Ning Gui; Blondia, C., Towards. 2008. Building Virtual Community for Ambient Assisted Living. 16th Euromicro Conference on Parallel, Distributed and Network-Based Processing.

[3] Kuo-Feng Huang, Shih-Jung Wu, Real-time-service-based Distributed Scheduling Scheme for IEEE 802. 16j Networks. *Journal of Networks*, Vol 8, No 3 (2013), 513-517

[4] Provenzi E Gatta C Fierro M et al. A spatially variant white-patch and gray-world method for color image enhancement driven by local contrast. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30(10) pp. 1757-1770.

[5] Zhou S, Aggarwal G, Chellapa R. Appearance characterization of linear lambrain object, Generalized photometric stereo and illumination- Invariant face recognition. *IEEE Trans on PAMI*, 2007, 29(2) pp. 230-245.

[6] D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procis-si, and F. Vitucci "Sampling Techniques to Accelerate Pat-tern Matching in Network Intrusion Detection Systems", *In Proc. ICC2010*, 2010, pp. 1-5, doi: 10.1109/ICC. 2010. 5501751.

[7] C. Becker, G. Schiele, H. Gubbels, K. Rothermel: BASE - A Micro-broker based Middleware For Pervasive Computing, *In Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications*, pp. 443-451, Fort Worth, USA, March 2003

[8] Muhammad J. Mirza, Nadeem Anjum, Association of Moving Objects across Visual Sensor Networks. *Journal of Multimedia*, Vol 7, No 1 (2012), 2-8

[9] Zhao Liangduan, Zhiyong Yuan, Xiangyun Liao, Weixin Si, Jianhui Zhao. 3D Tracking and Positioning of Surgical Instruments in Virtual Surgery Simulation. *Journal of Multimedia*, Vol 6, No 6 (2011), 502-509

[10] Dietrich, D.; Bruckner, D.; Zucker, G.; Palensky, P. 2010. Communication and Computation in Buildings: A Short Introduction and Overview. *IEEE Transaction on Industrial Electronic* Volume: 57, Issue:11.

[11] Guang Yan, Zhu Yue-Fei, Gu Chun-Xiang, Fei Jin-long, He Xin-Zheng, A Framework for Automated Security Proof and its Application to OAEP. *Journal of Networks*, Vol 8, No 3 (2013), 552-558

[12] R. Berangi, S. Saleem, M. Faulkner, et al. TDD cognitive radio femtocell network (CRFN) operation in FDD downlink spectrum. *IEEE, 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, 2011: 482-486

[13] Aricit, D Ikbaz S, Altunbasak Y. A histogram modification framework and its application for image contrast enhancement. *IEEE Trans on Image Processing*, 2009, 18(9):1921-1935

[14] J. Zhang and M. Zulkernine, "Anomaly based network intrusion detection with unsupervised outlier detection," *in Communications, 2006. ICC '06. IEEE International Conference on*, vol. 5, June 2006, pp. 2388-2393.

[15] Mao X. Nguyen, Quang M. Le, Vu Pham, Trung Tran, Bac H. Le, "Multi-scale Sparse Representation for Robust Face Recognition", *Proceeding(s) of Conference on Knowledge and Systems Engineering (KSE)*, pp. 195-199, 2011.

[16] Yikui Zhai, Junying Gan, Jingwen Li, "Study of occluded robust face recognition approach based on homotopy algorithm and color information fusion", *Signal Processing*, vol. 21, no. 11, pp. 1762-1768, 2011.

[17] Yang M, Zhang Lei, "Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary", *Proceedings of the European Conference on Computer Vision, ECCV*, pp. 448-461, 2010.

[18] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *in SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2000*, pp. 427-438.

[19] Konrad L., Matt W. 2006. MoteTrack: A Robust, Decentralized Approach to RF-Based Location Tracking. *To Appear in Springer Personal and Ubiquitous Computing, Special Issue on Location and Context-Awareness*. ISSN: 1617-4909 (Print) 1617-4917.

Routing Algorithm Based on Delay Rate in Wireless Cognitive Radio Network

Gan Yan and Yuxiang Lv*

College of Physics and Optoelectronics, Taiyuan University of Technology, Taiyuan 030024, Shanxi Province, China

*Corresponding Author, Email: yg3713@126.com, lyx823@126.com

Qiyin Wang

Shanxi Datong Electric Power Supply Company, Datong 037008, Shanxi Province, China

Email: rui_zhaoli@163.com

Yishuang Geng

Center of Wireless Information Network Studies (CWINS), Worcester Polytechnic Institute, Worcester, MA, 01609, USA

Email: yugeng@wpi.edu

Abstract—To reduce the end-to-end average delay of algorithm in wireless network, this paper proposes the real-time routing algorithm in spectrum network. It is analyzed that the dynamic changes of the radio network model and routing algorithm in spectrum network. Through using Markov state transition and adjusting the router with scaling factor, the high-quality resources in the network can be obtained and fully utilized, and then these can reduce the transmission time latency rate and timely adjust the route. After that the tendency of spectrum network and specific real-time algorithm are given. Finally, by using the network simulation NS-2, simulation experiments are used to estimate the performance test. Experimental results show that compared with the traditional algorithm, the proposed algorithm can obtain a lower end-to-end average delay and improves network throughput and the steady and reliability of the link connection.

Keywords—Real-time Routing; Protocol System; Bandwidth Capacity; Proportion

I. INTRODUCTION

Cognitive radio is an intelligent wireless communication system. It is able to perceive the external environment and uses artificial intelligence technology to learn from environment, and makes its internal state adapt to the statistical changes of received wireless signal by changing some operating parameters (such as transmit power, carrier frequency and modulation technology, etc.) so as to realize high reliable communication at any time and in any place and the effective use of spectrum resources [1].

With the rapid growth of the number of wireless subscribers and the rapid development of wireless communication technology, wireless spectrum resources become increasingly scarce [2]. In order to improve the spectrum utilization, Dr. Joseph Mitola first proposed the concept of cognitive radio; the report of Federal Communications Commission (FCC) also gives a

definition of cognitive radio committed to solve the problem of spectrum scarcity, and the core idea is to make it have learning ability and interact information with the surrounding environment to percept and utilize the available spectrum in the space, and to limit and reduce conflicts [3-5]. It is remarkable for its flexibility, intelligence, reconfigurable properties. It can learn from the environment, sense the external environment and change intentionally certain operating parameters (such as transmission power, carrier frequency and modulation techniques, etc.) in real time to make the internal state adapt to the statistical variations of the received wireless, thus it can achieve highly reliable communication at all times and places and efficiently use of the limited radio spectrum resources on heterogeneous network environment [6-9]. The people in cognitive radio network research have focused on the physical layer, MAC layer (Media Access Control MAC) key technologies and network layer routing protocol [10]. Among them, the routing protocol is an important part of the cognitive radio networks, but so far do not have a representative routing protocol become important issues in cognitive radio network research for optimizing the routing algorithm [11-13].

In recent years, more and more attentions have been given to cognitive radio routing algorithm being, and some routing algorithms have been proposed. Literature MA Hui-Sheng and other proposed a single transceiver demand routing protocol, where do not need for using the control channel between nodes, and the gain and cost causing by node's channel switching have been balanced, but there is no the specific solution for the problem of "deafness" while switching node [14-17]. The algorithm given by Chun LS based on the node local's sensing information establishes the spectrum map, puts forward transmission indicators of opportunity link and uses the coordination mechanism to maintain throughput of the statistical quality of service (QOS) in the multi-channel

transmission. How KC proposes a routing algorithm with opportunity differentiated services; through a combination of transmit power control and opportunistic routing to obtain differentiated services, minimum delay and the most stable routing.

Bogliolo Alessandro and the others will solve the maximum flow problem of FF based on energy harvesting wireless sensor networks. Because when FF looks for a maximum flow path, it takes into account the capacity problem and establishes the connection between capacity and energy harvesting rate, it makes the algorithm well used for energy harvesting wireless sensor networks, but due to the arbitrariness of augmented link choice of FF itself, it improves the calculation complexity and it can not guarantee convergence to the maximum of flow. ZHU Jia and the others improve the energy detection performance in fading environment through the use of joint spectrum sensing method of multiple cognitive users. In the joint spectrum perception, the cognitive users send local perception results to the fusion center and fusion centers use the specific fusion rule to combine all the sensory information to make final decision about the presence or absence of authorized users [18-22]. SUN Chunhua and the others adopt the method of clustering cooperation awareness to improve the detection performance under fading channel, but they don't give how to make clustering for cognitive user [22-25]. GUO Chen and the others use cluster nodes to make spectrum perception and send sensory information, but because other cognitive users are not involved in perception, the algorithm's performance does not improve during the period of channel fading. In particular, when the channel is perfect, due to the shrinking of users' participation in cooperation, perceived performance can reduce instead.

The above is the algorithm of cognitive radio routing. Although there are some improvements on the reliability of delay, throughput and link, still it is necessary to study the routing algorithm. This kind of routing algorithm is aimed at finding the minimum end-to-end delay path. Traditional route is mostly based on the shortest hops, however, in cognitive radio network, and the state information of each link is different. The information includes bandwidth of each channel between chains, available probability and channels, so the traditional routing algorithm is not suitable in cognitive radio networks. This paper proposes the Dynamic Spectrum Variation real-time routing algorithm (DSVR); the main idea are as follows: When better (greater bandwidth or higher utilization) spectrum in cognitive radio networks is idle due to the exit of primary user, cognitive users can sense it through spectrum, and timely update the table; if there is a user using a better routing (such as: lower transmission delay) in the routing table, cognitive users take the initiative to withdraw current route and adjust to a better route. While the traditional routing protocols will not take the initiative to adjust the route in the case of cognitive users have routing to use; unless the existing route can not be used, cognitive users will find a new route. Finally, the following theoretical analysis and

simulation prove the correctness of the proposed algorithm, and we can see that on the indicator of average end-to-end delay DSVR is significantly improved than traditional routing.

This paper mainly discusses the work of expanding and innovations in the following areas:

As to the defects of higher average end-to-end delay for the radio network in traditional routing protocols, firstly, the author proposes Dynamic Spectrum Variation real-time (DSVR), which analysis the radio network model and the routing algorithm in detail. Secondly, in the process of spectrum dynamic change, the Markov state transition is used to adjust the routing timely and the scaling factor α and the TTT are introduced to balance the effect on the network in the process of routing, and then the Markov model is used to evaluate the average end-to-end delay of the entire network. To a certain extent, this method makes fully use of resources with high quality in the network and it can obtain a lower transmission delay. Finally, gives specific steps on Dynamic Spectrum Variation real-time routing algorithm. The algorithm in the process of dynamic Spectrum Variation uses the Markov state transfer, and timely adjusts the routing to make full use of quality resources on network to a some extent, and achieve lower transmission delay.

In order to further verify the algorithm's correctness and effectiveness of lower transmission delay based on Dynamic Spectrum Variation real-time, the author makes a detailed experimental simulation in the radio network model, and experimental simulation results show that: DSVR algorithm has a lower average end-to-end delay compared with traditional routing algorithm in the case of channel's availability probability greater than 0.5, and has a lower average end-to-end delay, which ensures the stability and reliability of the throughput the link.

Cognitive radio networks using cognitive radio technology show some characteristics different from the traditional network due to its unique spectrum reusability and huge coverage:

Allocating radio resources in the multi-system coexisting conditions. The link between the users needs to carry out an effective control and management, at the same time to meet the delay and bandwidth requirements, and to realize a data transmission scheduling.

The system should have the capability of multi-channel support. If it is needed the central controller should be able to multiple adjacent channel aggregation processing to improve system performance, and support users to use and occupy a wider coverage. It can the user indicate which channel groups can be polymerized for use in some control frame, so the user can adopt multichannel mode.

Coexistence problems faced by the system. The coexistence problems include two levels: the first is the interference on the primary user system; the second is the coexistence problems of cognitive network entities in the overlapping areas or partially overlapping. To avoid interference to the primary user, distributed spectrum senses, measures and detects algorithms, and manages spectrum. All the specific functions of cognitive radio

technology must be considered. In the reality, the plurality of cognitive radio cell with great coverage is likely to occur partially overlap, in the worst case it is even completely overlapped. Consequent self-interference problems can not be resolved, it will seriously affect cognitive radio network.

Since the cognitive radio network having a dynamic, flexible, intelligent features, and thus the requirements for the network protocol is also relatively high, the protocol with asynchronous real-time characteristics, must be adaptive in the availability of the terminal changes, changes in the wireless environment the dynamic changes in the spectrum resources, network topology changes. Therefore, in the design of cognitive radio network protocols, follow these guidelines:

The agreement should be designed to fully reflect the characteristics of the cognitive radio technology. The common communication protocol architecture is hierarchical, in the cognitive radio network design, it will mainly consider the physical layer and media access control (MAC) layer and the network layer. In the specific design process, and will draw on the existing physical layer, MAC layer and the network layer protocol hierarchy, on this basis, by adding a function of the characteristics of the cognitive radio module.

Protocol architecture design should be combined with the results of algorithms and network architecture design systematically considering. Due to the design of cognitive radio network protocols and the network structure is closely related to the algorithm and the network structure is closely related to, among complement each other and influence each other. So in the process of designing network protocol, a preliminary framework should be established, and then combines with the results of algorithm design and network architecture design constantly revised, and finally completes the design of network protocols.

Protocol architecture design should, as far as possible, consider the compatibility, that is, considering the coexistence with other systems. The current communication pattern is coexistence of multiple systems, so when cognitive radio protocol architecture is designed should, it is really needed to take full account of the coexistence problem with other systems.

II. PROPOSED SCHEME

A. Radio Network Model

Cognitive radio network consists of N cognitive nodes, M link and m orthogonal channels within two adjacent nodes; each of the cognitive nodes has the same transmission range and interference range, and can access available channel set through the spectrum. The channel set dynamic changes over time and space. Cognitive radio network can be abstracted in Figure 1, where $G < V(G)$, $E(G) > G$, $V(G) = \{V_1, V_2, \dots, V_n\}$ means the cognitive nodes set in the network; $E(G) = \{e_1, e_2, \dots, e_m\}$ means the available channel set between two adjacent nodes. Just as shown in Figure 1, the source node and the destination node, respectively, represented by s and t; a, b,

c, d means the intermediate node, the connection $e_i \in \{1, 2, \dots, m\}$ between the nodes between nodes means that the available channel set.

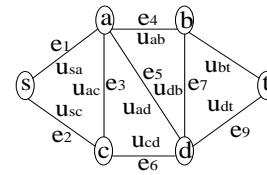


Figure 1. Topology of cognitive radio network

Assuming that the probability of primary user occupying the channel is b_k , then we can see:

$$d_k = t_{on}^k / (t_{on}^k + t_{off}^k)$$

t_{on}^k means statistical average time taken by the primary user in channel K with an active state; t_{off}^k statistical average time taken by the primary user in channel K with an inactive state. Then the probability of two adjacent nodes v_i and v_j communicating together in channel (channel availability) is as follows:

$$EI \frac{\partial^4 u(x,t)}{\partial x^4} + \frac{m_s}{L} \cdot \frac{\partial^2 u(x,t)}{\partial t^2} = 0 \tag{1}$$

$$a_{ij}^k = 1 - d_k$$

According to the Shannon formula, the capacity c_{ij}^k of nodes in the channel is as follows:

$$c_{ij}^k = w_{ij}^k \log_2(1 + snr_{ij}^k) \tag{2}$$

$$u(x,t) = X(x)e^{i\omega t}$$

w_{ij}^k is the bandwidth between two adjacent nodes and inter-channel; snr_{ij}^k is signal-to-noise ratio of nodes.

Given packet length as L, the data transmission delay between nodes and the channel is as following:

$$t_{ij} = \frac{L}{C_{ij}^k} \tag{3}$$

$$\begin{cases} T u_{ik} + T_f U_f = T_e I_f \\ T s n_{ik} + U_f = T_e \end{cases}$$

In the network topology Figure 1, the transmission delay between nodes represented by the value of the edge weights u_{ij} .

B. Routing Algorithm

According to the above instructions of the cognitive radio routing algorithm, in order to make it easier to describe the system, we establish a Markov link-state system model. n routes have been abstracted in the system; state 0 indicates no routing available; state 1 indicates that the route1 is available; P_0 indicates probability that there is no route can be used; P_1 indicates probability that route 1 cannot be used; P_2 indicates probability that route 2 cannot be used; p_n indicates the

probability that no route can be used. The probability that no route can be used is defined as p , which is the joint probability of the probability that the route between nodes can not be used. Routing table sorts the routing with priority decreasing. as following: if the route 1 is optimal (minimum transmission delay), routing 2 is sub-optimal and routing n is the worst. The nodes select routers according to the priority. System state transition diagram is shown in Figure 2:

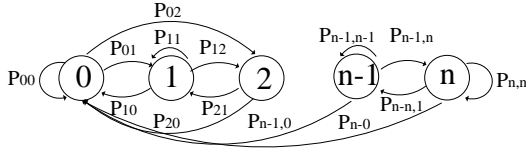


Figure 2. State transition diagram of cognitive radio network routing

From the above state transition diagram the state transition matrix composed by the state transition probability P_{ij} can be drawn as the follows:

$$P = \begin{bmatrix} P_{00} & P_{01} & \dots & P_{0j} & \dots \\ P_{10} & P_{11} & \dots & P_{1j} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ P_{i0} & P_{i1} & \dots & P_{ij} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

P_{ij} in the state transition matrix means the transition probability from state i to state j .

According to the traditional routing algorithms state transition matrix can be drawn as following:

$$P = \begin{bmatrix} p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \\ p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \\ p_1 p_2 \dots p_n & (1 - p_1)p_2 & 1 - p_2 & \dots & p_1 p_2 \dots (1 - p_n) \\ \dots & \dots & \dots & \dots & \dots \\ p_1 p_2 \dots p_n & (1 - p_1)p_n & p_1(1 - p_2)p_n & \dots & 1 - p_n \end{bmatrix} \quad (4)$$

The state transition of traditional routing algorithms can be described as follows: In Figure 2, if the system is in state 0, the system will select the route with the highest priority, and jump to the state 1; if the state 1 is not available, the system jump to state 2, 3..... n ; these will stop when the system has been found the current network route with the highest priority. If the system's current state i ($i < n$) are available caused by spectrum changes in the network environment due to the in the, the system continue to remain in the state (state can continue to use the case) according to the traditional routing algorithm. Take state 1 and state 2 as example: if the system is in state 2, and the state 1 at this time is available, the system can only stay in state 2, and the state transition probability is $P_{22}=1-P_2$. If the state 2 can not continue to use, the system will jump to state 1; transition probability of state 2 and state 1 is $P_{21}=1-P_2$. From the state transition matrix we can see that in the case of the original state available, the system can not use the current network of quality resources, make the high-quality resources in an idle state, and not make full use of network resources.

According DSVR algorithm the author proposed the following state transition matrix can be drawn:

$$P = \begin{bmatrix} p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \\ p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \\ p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \\ \dots & \dots & \dots & \dots & \dots \\ p_1 p_2 \dots p_n & 1 - p_1 & p_1(1 - p_2) & \dots & p_1 p_2 \dots (1 - p_n) \end{bmatrix} \quad (5)$$

Assuming that the system is in this state, if state i is available, from DSVR algorithms the system will jump from state j to state i (even if in the case states 1 and 2 of a state can continue to use). The author will describe DSVR algorithm by the state transition from states 1 to state 2. If the system is in state 2, and state 1 available, in accordance with the DSVR algorithm, the system will exit state 2 to the state 1 (even in the case of state 2 can be used); the state transition probability is $P_{21}=(1-P_1)$. Compared with the state transition probability of traditional routing algorithms, the probability of backing to state 1 has been increased. So, compared with traditional routing algorithms DSVR algorithm can better make full use of the high-quality resources in the network to improve the overall network performance.

Hereinafter, the author makes mathematical verification on state transition diagram, just as shown in Figure 2; set the state transfer Picture as C , which is the non-empty subset of the state space I ; as to any state i , when $i \in c$ and, $p_{ik} = 0$, which is a random closed set and all the state in random closed set are interconnected, so it is irreducible closed set; from state transition Figure 2 is easy to know that for each state, the cycle is equal to 1, so the state transition figure is the Markov chain of non-periodic irreducible closed set, in which the probability distribution $\{\pi_j, j \in I\}$ is steadily distributed, and satisfies:

$$\begin{cases} P_j^* = \pi_i P_{i,j}^* \\ \pi_j = \sum_{i \in I} \pi_i P_{ij}, \\ \sum_{j \in I} \pi_j = 1, \pi_j \geq 0 \end{cases} \quad (6)$$

P_{ij} means the state transition probabilities; $\pi_0, \pi_2, \dots, \pi_j$ means the stationary distribution of P_0, P_1, \dots, P_j . From the equation (7), (8), (9), (10), (11) end-to-end average delay E can be drawn:

$$\begin{cases} E = \sum_{i=0}^n R_i \pi_i \\ \sum_{j \in I} \pi_j = 1, \pi_j \geq 0 \\ \pi_j = \sum_{i \in I} \pi_i P_{ij} \end{cases} \quad (7)$$

$$P = \begin{bmatrix} P_{01} & P_{02} & \dots & P_{0n} & \dots \\ P_{03} & P_{04} & \dots & P_{1n} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ P_{i0} & P_{in} & \dots & P_{ik} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \quad (8)$$

$$C_{i,k} = \frac{D_{i,k}}{R[p(i), p(k)]} \tag{9}$$

Among them, π_j indicates the stationary distribution backing to the j route; T_i means the transmission delay (T_i is the accumulated transmission delay of the i route between nodes) of the i route.

The DSVR algorithm has less propagation delay than the cognitive users, which are using the current router after the network routing table is updated, and can timely adjust the router. This makes a difference on the stability of communication and overhead of the network between the pair of nodes. For example: the losing of packet, re-transmission of packet, and the delay causing by switching channel. Therefore, this paper introduces the scale factor α and switch trigger time to balance these effects.

$$T = \begin{bmatrix} T_{00} & T_{01} & \dots & T_{0j} & \dots \\ T_{10} & T_{11} & \dots & T_{1j} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ T_{i0} & T_{i1} & \dots & T_{ij} & \dots \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix} \tag{10}$$

$$com = |T_{1,j}^*| \times |T_{2,j}^*| \times \dots \times |T_{k,j}^*| \tag{11}$$

$$T' \leq T(1-\alpha)$$

In equation (9), T represents the transmission delay before adjustment; T' means the transmission delay after adjustment. When the above conditions have been satisfied, the route can be switched. The introduction of the scale factor not only can make effective control on routing adjustment, balance the influence caused by route switching, and ensure the quality of communication between nodes, but also can prevent frequent adjustment of the routing. TTT is equivalent to observations of a time window. In cognitive radio networks spectrum is dynamic and changes with time. When on the period of TTT, the system is only make detection on the routing, and not makes adjustment on routes; when observation time is over TTT, it makes route adjustment after satisfying (9). This also effectively prevents the over spectrum of route changes caused by the dynamic changes of cognitive radio networks spectrum; this makes the routing adjustment too frequent and shuffled adjustment between routes, generating a ping-pong efficiency; this affect the stability of the inter-node communication and increase network overhead.

The specific steps on algorithm DSVR are as follows:

a to d is the traditional routing algorithms; e to g is the DSVR algorithm which makes improvements based on traditional routing algorithm,

a) Vector $[\alpha_{ij}^k, w_{ij}^k]$ indicates the characteristics between the two nodes; α_{ij}^k, w_{ij}^k are the available probability and bandwidth of the two nodes in the channel k . In the construct cognitive radio network $G < V$

(G), $E(G) > G$, the source node and the destination node are separately represented by s and t .

b) From equation (2), (3) to obtain the channel capacity between two nodes as c_{ij}^k and t_{ij} .

c) Finding out all the link of the destination nodes in the source nodes, and obtaining the transmission delay of each link by superimposition between two nodes.

d) Arranging all the links in the c from small to large (fast row), and selecting $\min\{Ti\}$ from it. Using formula (4), (7) to draw the average network delay.

e) When d is complete, sending the message of Hello every 2s, making routing maintenance, and timely updating the routing table.

f) When e is completed, if $T' \leq (1-\alpha) \min\{T_j\}$ and the observation time is greater than TTT, it is the updated routing table $\min\{T_j\}$. Using formula (5), (7) to get the average network delay.

g) If it is not satisfied f, it will back to the e.

Analysis of algorithm time complexity and space complexity:

The time and space complexity of traditional routing algorithms can be expressed as:

$$T = nO(\log n), s = O(n)$$

where: n is the number of link in the network, which the source node s can reach the destination node t .

The time and space complexity of the algorithm DSVR can be expressed as:

$$T = mO(\log m), s = o(m)$$

where: m is the link number after the routing table in the network updated and satisfying the formula (8), so $m < n$. Therefore, the time and space complexity of the algorithm DSVR are less than the time and space complexity of the traditional routing algorithm.

TABLE I. SIMULATION PARAMETER SETTINGS

Simulation Setup	Simulation parameters
Region-wide	1200
Data Type	CBR
Multiplexing method	TDD
Packet size	1000 B_{yte}
MAC Layer	803.12DCF
Wireless transmission range	255m
Signal-to-noise ratio	17db
Transmission cycle	2s
TTT	5s
α	0.11

III. SIMULATION AND ANALYSIS

A. Simulation Parameter Settings

For performance analysis of the DSVR algorithm, the author uses common network simulation NS-2 to evaluate the performance. Simulation environment is as follows: N cognitive nodes randomly distributed within a range of 1200m*1200m; the physical layer and the MAC layer of the wireless transmission adopt the IEEE802.11 protocol; the wireless transmission is ideal, i.e. no error and delay. Five groups of different nodes are randomly selected from the collection of nodes in the simulation

environment; each set has two nodes, respectively, as the source node and the destination node. In the case of available probability, the number of available channels and the number of nodes in different channel, the system make evaluation on the average end-to-end transmission delay. Other specific simulation parameters are shown in Table 1.

B. Analysis of Simulation Results

As can be seen from Figure 3, when the channel availability is relatively low, the size of average delay obtained by different types of methods are almost the same; when the channel availability is greater than 0.5, the size of average delay adopting DSVR is significantly less than the average delay using traditional routing. Moreover, with the increase in channel availability, this trend is more and more obvious. The reasons for this phenomenon are as follows: based on the Markov state transition matrix the greater the channel available probability, the greater the steady-state probability of selecting lower transmission delay path, so under the same parameters of the number of nodes, channel bandwidth and the number of channels the between nodes, the DSVR routing's average delay is significantly lower than traditional routing. Meanwhile, it can be seen from Figure 3, the probability of channel availability is from 0.5 to between 0.6; DSVR routing average delay is lower than the conventional route by 10%, and this is consistent with the equation (6), which indicate that at this time the routes adjustment occurs. The vector $[a_{ij}^k, w_{ij}^k]$ represents the characteristics between the two nodes; a_{ij}^k, w_{ij}^k are the available probability and bandwidth of the two nodes on channel k. In construct cognitive radio network $G < V(G), E(G) > G$, the source node and the destination node are respectively presented as s and t. b) the channel capacity and t_{ij} between two nodes can be obtained through equation (2) and (3). c) Identifying all node links which the source node can reach the destination and overlying two nodes to get the transmission delay of each link. d) Arranging all the links according to the ascending mode (using fast row) and selecting $\min \{T_i\}$. Using Equation (4) and (7) derived o get average network delay. e) When d finished, Hello message is sent once every 2 seconds for routing maintenance and updating routing tables. f) When e complete, if and the observation time is more than TTT, the $\min \{T_j\}$ in updated routing table is appeared. The average network delay is got by using formula (5) and (7). After the network routing table updated and satisfied the link number in formula (8), the $m < n$. It can be seen that, compared with traditional routing algorithm, the time complexity and space complexity of DSVR algorithm time are much less, and its average delay is significantly reduced.

From Figure 3 it can be concluded that in the case of the parameters of DSVR routing and the traditional routing algorithm are the same, the advantage of DSVR routing is mainly reflected in the channel's probability

above 0.5. So, we set the channel's probability are $P_i = 0.6$. As can be seen from Figure 4, when the channel's probabilities are the same, with the number of channels available increased in the network, the source node and the destination node has more routes to choose from. According to DSVR routing algorithm, the probability of router adjusted to the lower transmission delay is greater, so the average delay of DSVR cognitive routing algorithm is significantly reduced compared to the average delay of traditional router.

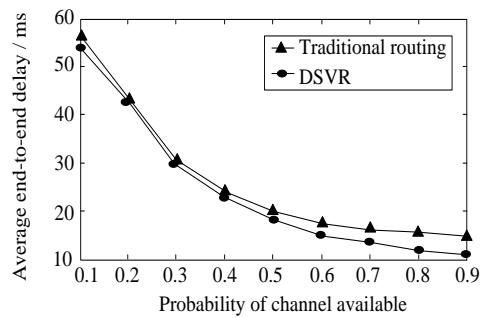


Figure 3. Changes of average delay with the channel's probability

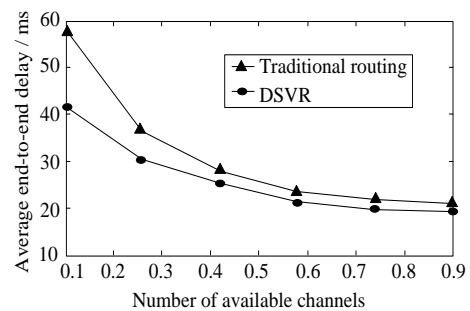


Figure 4. Changes of average delay with the number of available channels

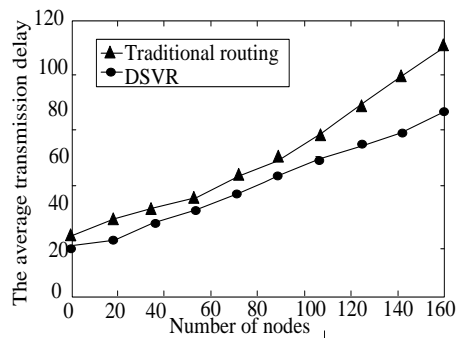


Figure 5. Changes of average delay with the number of nodes

As can be seen in Figure 5, with increase of the number of nodes in the network, the average delay gradually increased. But after using DSVR the average delay is significantly lower than the traditional route, and growth of the average delay is slower than traditional route. The causes of this phenomenon are as follows: with the increase nodes' number, the size of the entire network becomes large, the number of hops of the source node and the destination node increase, so the average end-to-end delay increases; But with the increase of the number of nodes, available route in the network also

increase, while the system can timely adjust the route by adopting DSVR, and make use of route with a smaller transmission delay, which can make some compensation to the transmission delay caused by the increase in the number of hops. Therefore the average delay adopting DSVR route is lower than adopting traditional routing.

IV. CONCLUSIONS

This paper presents a kind of DSVR algorithm, which mainly takes the impact of the route caused by dynamic spectrum in cognitive radio network into account. According to the size of the route's transmission delay, the system timely adjust the routing, introduce the scale factor a and switch trigger time (TTT) to balance the impact on the network in the process of adjusting the routing, and then use a Markov model to evaluate the average end-to-end delay of the entire network. Finally, after the analysis of simulation and comparison, the author concludes that compared with traditional routing algorithm DSVR algorithm has a lower average end-to-end delay under the circumstances channel's probability greater than 0.5. The next study is to improve the routing algorithm and all aspects of performance.

REFERENCES

- [1] Y. Geng, J. He, H. Deng and K. Pahlavan, Modeling the Effect of Human Body on TOA Ranging for Indoor Human Tracking with Wrist Mounted Sensor, *16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Atlantic City, NJ, Jun. 2013.
- [2] Yang Chungang, Sheng Min, Li Jiandong, et al.. Energy-Aware Joint Power and Rate Control in Overlay Cognitive Radio Networks: A Nash Bargaining Perspective. *In: International Conference on Intelligent Networking and Collaborative Systems, Bucharest*, pp: 520-524, Sept. 2012.
- [3] R. Berangi, S. Saleem, M. Faulkner, et al. TDD cognitive radio femtocell network (CRFN) operation in FDD downlink spectrum. *IEEE, 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, 2011 pp. 482-486.
- [4] D. Xu, Z. Y. Feng, Y. Z. Li, et al. Fair Channel allocation and power control for uplink and downlink cognitive radio networks. *IEEE., Workshop on mobile computing and emerging communication networks*, 2011 pp. 591-596.
- [5] S. Li, Y. Geng, J. He, K. Pahlavan, Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, *2012 5th International Conference on Biomedical Engineering and Informatics (BMEI)*, Chongqing, China Oct. 2012 pp. 721-725.
- [6] Neamatollahi P., Taheri H., Naghibzadeh M., Yaghmaee M. H., "DESC: Distributed Energy Efficient Scheme to Cluster Wireless Sensor Networks," *Proc. The 9th IFIP TC 6 International Conference 2011*, pp. 234-246, Jun. 2011.
- [7] Ning Xu, Aiping Huang, Ting-Wei Hou, Hsiao-Hwa Chen, "Coverage and Connectivity Guaranteed Topology Control Algorithm for Cluster-based Wireless Sensor Networks," *Wireless Communications and Mobile Computing*, vol. 12, no. 1, pp. 23-32, Jan. 2012.
- [8] J. He, Y. Geng and K. Pahlavan, Modeling Indoor TOA Ranging Error for Body Mounted Sensors, *2012 IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, Sydney, Australia Sep. 2012 pp. 682-686.
- [9] Zhao Liangduan, Zhiyong Yuan, Xiangyun Liao, Weixin Si, Jianhui Zhao. 3D Tracking and Positioning of Surgical Instruments in Virtual Surgery Simulation. *Journal of Multimedia*, Vol 6, No 6 (2011), 502-509
- [10] O. Olabiyi, A. Annamalai. Efficient Performance Evaluation of Cooperative Non-regenerative Relay Networks. *In: IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, U. S. A.*, pp. 797-801, Jan. 2012.
- [11] W. Q. Yao, Y. Wang, T. Wang. Joint optimization for downlink resource allocation in cognitive radio cellular networks. *IEEE., 8th Annual IEEE consumer communications and networking conference*, 2011 pp. 664-668
- [12] S. H. Tang, M. C. Chen, Y. S. Sun, et al. A spectral efficient and fair user-centric spectrum allocation approach for downlink transmissions. *IEEE., Globecom.*, 2011 pp. 1-6
- [13] K. Ruttik, K. Koufos, R. Jantir. Model for computing aggregate interference from secondary cellular network in presence of correlated shadow fading. *IEEE, 22nd International symposium on personal, indoor and mobile radio communications*, 2011 pp. 433-437
- [14] J. Naereddine, J. Riihijarvi, P. Mahonen. Transmit power control for secondary use in environments with correlated shadowing. *IEEE, ICC2011 Proceedings*, 2011 pp. 1-6
- [15] W. Ahmed, J. Gao, S. Saleem, et al. An access technique for secondary network in downlink channels. *IEEE, 22nd International Symposium on Personal, Indoor and Mobile Radio Communications*, 2011 pp. 423-427
- [16] D. L. Sun, X. N. Zhu, Z. M. Zeng, et al. Downlink power control in cognitive femtocell networks. *IEEE., International conference on wireless communications and signal processing*, 2011 pp. 1-5
- [17] N. Omidvar, B. H. Khalaj. A game theoretic approach for power allocation in the downlink of cognitive radio networks. *IEEE, 16th CAMAD*, 2011 pp. 158-162
- [18] Wright J, Yang A, Ganesh A, et al., "Robust face recognition via sparse representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, 2009.
- [19] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Toward a practical face recognition: Robust registration and illumination via sparse representation", *In Proceeding(s) of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 597-604, 2009.
- [20] D. Donoho, "Compressed Sensing", *IEEE Transactions on Information Theory*, vol. 52, pp. 1289-1306, 2006.
- [21] Q. Zhang, B. Li, "Discriminative K-SVD for dictionary learning in face recognition", *Proceeding(s) of Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2691-2698, 2010.
- [22] Mao X. Nguyen, Quang M. Le, Vu Pham, Trung Tran, Bac H. Le, "Multi-scale Sparse Representation for Robust Face Recognition", *Proceeding(s) of Conference on Knowledge and Systems Engineering (KSE)*, pp. 195-199, 2011.
- [23] Yikui Zhai, Junying Gan, Jingwen Li, "Study of occluded robust face recognition approach based on homotopy algorithm and color information fusion", *Signal Processing*, vol. 21, no. 11, pp. 1762-1768, 2011.
- [24] Yang M, Zhang Lei, "Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary", *Proceedings of the European Conference on Computer Vision, ECCV*, pp. 448-461, 2010.
- [25] Guopeng Zhang, Kun Yang, Qingsong Hu, et al.. Bargaining Game Theoretic Framework for Stimulating Cooperation in Wireless Cooperative Multicast Networks.

IEEE Communications Letters, vol. 16, no. 2, pp. 208-211, Feb. 2012.

Gan Yan was born in Hebei province in China on 24th May, 1989. He received the bachelor degree of applied physics from Taiyuan University of Technology (China) in 2012, currently, he is a Master student of integrated circuit engineering in Taiyuan University of Technology (China).

Energy Hole Solution Algorithm in Wireless Sensor Network

Lu Yuting

Department of Electromechanical Engineering, Changzhou Textile Garment Institute, 213164, China

Wang Weiyang

Florida State University Tallahassee, Florida, United States, 32304

Abstract—Since the sensor nodes near the Sink take more communication load leading to excessive energy consumption and short its life cycle, this paper proposes a new energy hole solution algorithm. The algorithm adds some long chain to the Sink node to relieve energy hole of reducing data forwarding number around the Sink node, so as to prolong the lifecycle of the network. Firstly, The algorithm carries out the analysis of energy consumption to equidistance transmission network and puts forward adopting tactics of small world to alleviate energy hole and analyzes the position and number of long chain' influence on energy consumption and the network life cycle. Finally, the thesis carries out the simulation experiment. The experimental results show that this algorithm can significantly improve the network lifetime and easy to implement in practice.

Index Terms—Network Model; Energy; Equidistance Transmission; Cycle

I. INTRODUCTION

In sensor networks, how to effectively improve the network lifetime is a core research question. Sensory data is usually transmitted in the form of multiple hops to the Sink node, such failure of a sensor section near the Sink lead to formation of energy hole around the Sink.

The advent of energy hole makes data of sensor could not be served to Sink node, so the network life cycle is come to the end, and at this point, a large number of surplus energy is wasted in the network. How to design network effectively, make full use of the surplus energy to prolong the network life cycle, is a significant research topic. In order to solve the energy Empty problem, people put forward many methods. Point because need to transfer the data from other nodes and take more communication load, so these nodes easily exhausted their energy prematurely. Different with the existing research, this article is based on ideas of small world, through adding some long chain to the Sink node, reduce the number of data forwarding around the Sink node to relieve energy cavity, so as to prolong the lifecycle of the network, this method not only can significantly improve the network life cycle, and is easy to implement in practice [1-2]. We theoretically analyzed the number of the long chain and location of network energy consumption and energy cavity effect. Long chain in a

network can be a long chain of wireless or wired long chain. Wireless long chain arranged flexibly, however, may increase the network energy consumption; Cable long chain once the layout is very difficult to change, but in many application environments are difficult to deploy cable long chain, however, it is beneficial to reduce the total energy consumption network. Considering the convenience of practical application, this paper mainly adopts the wireless long chain, puts forward a new method based on small world to effectively relieve the energy hole.

The experiments results of Data capacity improvement of wireless sensor networks using non-uniform sensor distribution show that when the life of a sensor network ended, the total residual energy is over 90%. It is an important research topic on how to effectively design network and make full use of the surplus energy to extend the life cycle of the network. To solve the problem of energy holes, a number of methods have been proposed:

Lian J [3] studies the energy consumption in sensor networks, they found that through the experiment, for a static Sink, large-scale of uniform distribution networks, when the network at the end of the life cycle, there are still as much as 90% of the remaining energy. Therefore, the static model composed of uniform distribution of isomorphic node cannot make full use of the energy of the network. It also indicates the actual network life cycle can be further extended.

To do this, they put forward a kind of sensor nodes' uniform distribution strategy, near the Sink area is decorated more sensor nodes, the experimental results show that the new strategy can significantly improve the network lifetime. Wu X also uses uneven node layout strategy to relieve the energy hole [4].

They theoretically discusses the strategy that in the no uniform node distribution of circular network, if the node continues to send data to the Sink node, the energy empty phenomenon will not be able to avoid, and when the number of nodes meet a certain relationship, subprime balanced energy consumption can be implemented in a network. They offer a no uniform node distribution strategy and the corresponding routing algorithms are used to implement the subprime balanced energy consumption. Non-uniform distribution of nodes is an

effective way to solve energy hole, but it is often difficult to implement in practice [5].

Song Chao etc. based on the improved model, to realize energy-saving purpose by adjusting the data transmission distance of each ring node [6]. They prove that search optimal transmission distance of each area is a multi-objective optimization problem, and put forward a distributed algorithm based on ant colony optimization, all area according to its adaptive distribution to explore the approximate optimal of transmission distance, implementation network life extension. This method often require complicated calculations, also does not apply to sensor

Li J [7] studied energy space of more to one sensor network, they established an analysis model. Based on this model, analysis the validity of various existing technologies for alleviating energy hole, but they haven't come up with new methods to solve the energy hole. Olariu S [8] theoretically analyzed uneven energy consumption of sensor network; the author assumes that a circular uniform distribution in the network node, the node continues to transmit data to the Sink. In this paper, we use the energy consumption model is $E = d_a + c$, where $d(d \leq t_x)$ is the transmission distance, $\partial \geq 2$ is energy attenuation coefficient, c is a normal number, t_x is the maximum transmission radius of sensor nodes. The author proved that if the network is classified in the ring with same width, energy consumption on the routing come to a minimum. Also, they found that when the $\partial > 2$ by adjusting the width of the ring hole can be used to avoid energy, and $\partial = 2$, the network is unable to avoid energy cavity formation.

Jarry [9] a puts forward a hybrid routing strategy, adopt the method of linear programming and balance the regional energy consumption to prolong the life of network, the author assumes that all sensor nodes can transfer data directly to the Sink, namely distance from any sensor to the Sink are less than the maximum transmission distance, node pass data to the Sink based on a residual energy function directly or more jump way. Perillom m [10] analysis and points out that more jump way of in the same communication radius, node close to the base station is more likely to run out of energy; Instead, when nodes and base station use direct one hop communication, node far from the base station is death earlier due to large energy consumption of nodes. To this end, a power control strategy was proposed to balance the energy consumption between each node, sum up the maximum of the network life cycle down to a linear programming problem. The algorithm is not applicable for large-scale sensor networks.

Dasguptak k etc [11], proposed maximization algorithm by node responsibility assignment to the network lifetime, the algorithm contains two types of nodes, the sensor nodes and the forwarding nodes. Hierarchy is used by Hou y t [12], etc to balance energy consumption, the author proposes a dual system of wireless sensor network (WSN), the nodes are divided into groups, each group includes a funnel (AFN) node,

the node is responsible for the all nodes data's collection of this group and transfer data to the Sink, the AFN node and the Sink form the network system of layer 2.

Wang w, etc using mobile relay to prolong the lifecycle of the network, the authors found that the mobile relay need only mobile within two jump range can significantly improve the life cycle of the network. However, in many real networks, especially under the bad environment condition, it is difficult to effectively implement the mobile relay. Luo J etc [13], uses the mobile Sink for data collection. As the result of the Sink's movement, changing the node around it, so that you can avoid formation of energy hole around the Sink. However, in many applications, the Sink is not suitable for mobile, especially in the enemy's area.

Helmy a etc [14], introduces the logical link in the wireless sensor network to form a wireless network of small world effect, also applies to validate the small world network with spatial attributes of wireless sensor network (WSN). Research of Sharma G etc. shows that by adding a small amount of long chain into the network, can significantly improve in homogeneity energy consumption, improve the network lifetime. However, their work did not consider the problem of energy hole.

Liu A etc [15], using non-uniform node layout strategy to relieve the energy hole, the core of strategy is to get close to Sink the cluster radius is lesser, and the far Sink cluster radius is larger. Liu Tao put forward energy empty avoid mechanism based on non-uniform distribution of the node, namely according to the node energy consumption level to prepare different initial energy reserves for each node.

The solution to the energy hole can be summarized as two main kinds: (a) by the uneven distribution of nodes to solve energy hole, which is more close to the Sink and more decorate more sensor nodes, studies have shown that this method can significantly improve the network life cycle, but the no uniform node distribution strategy is difficult to achieve in practice, because in most cases the nodes are random, the distribution of local node density is difficult to control. (b) Using linear programming or optimize method to search for sensor nodes such as transmission distance to prolong the network life time, so as to solve the energy hole, this kind of method usually require different nodes using different communication radius, some method also ask each node has the ability to communicate directly with Sink, the complexity of the calculation and node mobility are more sensitive and constrains them in actual application.

According to the situation above, this paper proposes a new energy hole solution algorithm, not only can significantly improve the network life cycle, and it is easy to implement in practice.

This paper mainly makes the development and innovative work in the following aspects:

(1) Aimed at the sensor nodes near the Sink taking more communication load caused by excessive energy consumption and shorten its life cycle, therefore this paper proposes a new energy hole solution algorithm. In sensor networks, it is easy to form energy hole around the

Sink. Energy Hole makes the data undeliverable to Sink, and then the life cycle of network is ended, while there are lots of residual energy in the network.

This paper analyzes the network Model and energy model, from the Angle of energy make theory analysis to sensor network adopting the tactics of small world. First of all, analysis network energy consumption of isometric transmission, then puts forward that adopting the tactics of small world to alleviate energy hole, the position of the long chain are analyzed theoretically and the quantity of energy consumption and the influence of the network life cycle, discussed the small world network 's way of implementation.

(2) In order to further validate r the algorithm's correctness and effectiveness of a new energy empty proposes by this paper, experimental verification of the theoretical analysis result above. In simulation experiment, the parameter $\partial, \partial_1, \partial_2$ and ∂_3 take typical values, the method is suitable for the small-scale network, also applies to large-scale sensor networks, has a good scalability; It is most favorable to arrange long chain within the ring c_2 , when the r1 increases to a certain value, Tl/T basic remains the same, close to the λ value, when the $\lambda = \lambda l_j$, nodes in ring c_1 and c_2 have the same average energy consumption, at this time, the network has the maximum life cycle; Cost of network building and node deployment is small, simple communication mechanism, environmental impact of the system is small, and easy to implement. Experimental simulation results show that this algorithm can significantly improve the network lifetime and easy to implement in practice.

II. MODEL

This article assumes that all sensor nodes are uniformly distributed in the circular area with a radius R, sink, set in the center of the network. The network was divided into annular region by k center circle, the radius of the center circle respectively $r_1, r_2, r_3, \dots, r_k$, and satisfy $0 < r_1 < r_2 < r_3 < \dots < r_k = R$, particularly $r_0 = 0$. When $1 < i < k$, for any circle c_i is the annular region separated by two concentric circles with radius r_{i-1} and r_i . Circle's maximum width is the biggest transmission radius of sensor nodes t_x . Assume that sensor nodes in circle c_i adopt transmission distance r_{i-1}, r_{i-r} pass the data to c_{i-1} in the sensor nodes. Each node collects data within the scope of its induction and passes to sink. Assuming that the network size is large, to complete the data transmission of multiple hops. Node within the ring c_i not only need to pass its own data, but also passed the data from c_{i-1} to c_i . Structure of the network is shown in figure 1.

A typical sensor node usually consists of three parts: sensing unit, processing unit and transceiver unit. Assume that all sensor nodes have the same initial energy C, Sink node's energy is infinite. The network uses a

optimized sleep scheduling agreement, free sensor nodes into sleep mode, so energy consumption of the sensor nodes in the free state are ignored. Due to the energy consumption of sensor network is mainly caused by data transmission, so the model only considering the energy consumption of the data transmission and receive energy consumption of the data, the formula for energy consumption.

$$C_{tx} = \partial_1 + \partial_2 r^a, C_{rx} = \partial_3$$

Among them, the C_{tx} as the number data of transmission unit's energy consumption, C_{tx} for energy consumption of receiving unit number of data, ∂_1, ∂_2 and ∂_3 is normal, d is the data transmission distance, $2 \leq \partial \leq 7$.

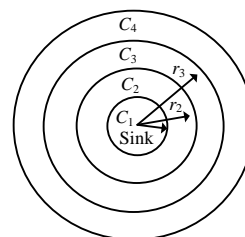


Figure 1. Network model

III. A NEW ENERGY HOLE SOLUTION ALGORITHM

From the Angle of energy, the tactics of small world is adopted to make theory analysis in sensor network. First of all, analysis to isometric transmission of network energy consumption, then puts forward that adopting the tactics of small world to alleviate energy hole, the position of the long chain are analyzed theoretically and the quantity of energy consumption and the influence of the network life cycle, discussed the small world network 's way of implementation.

A. Energy Consumption of Equidistance Transmission

In much of a sensor network, all sensor nodes of sensory data are transmitted to the Sink node through multiple hops ways, this led to the uneven energy consumption of sensor nodes near the Sink forward because the need for more data from other nodes and early exhausted their energy, leading to a energy hole around the Sink. We assume that each sensor node produce b unit data per unit time, since the node is randomly distributed evenly, so the whole network node density is uniform, that is

$$q = \frac{R}{M}$$

Among them, R is the total number of sensor nodes in the network; M is the area of the network area.

According to our network model and energy model, the total energy consumption of ring c_i is

$$M_{total} = M_{rx1} + M_{tx1} = q\pi r_k^2 c (\partial_1 + \partial_2 r_1^a) + \partial_3 q\pi (r_k^2 - r_1^2) c \quad (1)$$

Perillo M theoretically analyzed the uneven energy consumption of sensor network, and proves that if the network is classified having the same width with the ring,

energy consumption on the routing come to minimum. Therefore, our model also uses the same width of the circle, so $r_i = r_{i-1}$ ($1 \leq i \leq k$), (1) into

$$M_{total} = q\pi c \left(r_k^2 \partial_1 + r_k^2 \partial_2 r_1^a + r_1^2 (k^2 - 1) \partial_3 \right)$$

The average energy consumption of each node in the ring c_i is

$$M_1 = \frac{M_{total}}{q\pi r_1^2} = \frac{q\pi c \left(r_k^2 \partial_1 + r_k^2 \partial_2 r_1^a + r_1^2 (k^2 - 1) \partial_3 \right)}{q\pi r_1^2} \quad (2)$$

$$= c \left(k^2 \partial_1 + k^2 \partial_2 r_1^a + (k^2 - 1) \partial_3 \right)$$

Assuming sensor node in circular ring c_i uses the transmission distance $r_i - r_{i-1}$ pass the data to the sensor nodes in r_{i-1} , the number of nodes in ring i is $q\pi(r_i^2 - r_{i-1}^2)$, the receiving data is $q\pi b(r_k^2 - r_i^2)$, the amount of data needed to transferred is $q\pi b(r_k^2 - r_i^2)$, if the total energy consumption is M_{total} , the average energy consumption of each node is M_i , there are

$$M_{total} = q\pi c \left(r_k^2 - r_{i-1}^2 \right) \left(\partial_1 + \partial_2 r_1^a \right) + q\pi c \left(r_k^2 - r_i^2 \right) \partial_3$$

$$= q\pi c r_1^2 \left(k^2 - i^2 + 2i - 1 \right) \left(\partial_1 + \partial_2 r_1^a \right)$$

$$= q\pi c r_1^2 \left(k^2 - i^2 \right) \alpha_3$$

$$M_i = \frac{M_{total}}{q\pi \left(r_i^2 - r_{i-1}^2 \right)}$$

$$= \frac{q\pi c \left(r_k^2 - r_{i-1}^2 \right) \left(\partial_1 + \partial_2 r_1^a \right) + q\pi c \left(r_k^2 - r_i^2 \right) \partial_3}{q\pi r_1^2 \left(2i - 1 \right)}$$

$$= \frac{q\pi c r_1^2 \left(k^2 - i^2 + 2i - 1 \right) \left(\partial_1 + \partial_2 r_1^a \right) + \left(k^2 - i^2 \right) \partial_3}{2i - 1}$$

so, $\frac{M_i}{M_1} = \frac{q\pi c r_1^2 \left(k^2 - (i-1) \right) \left(\partial_1 + \partial_2 r_1^a \right) + \left(k^2 - i^2 \right) \partial_3}{(2i-1) \left(k^2 \partial_1 + k^2 \partial_2 r_1^a + (k^2 - 1) \partial_3 \right)}$

Thus, we get

$$\frac{k^2 - i^2}{k^2 (2i - 1)} \leq \frac{M_i}{M_1} \leq \frac{k^2 - (i-1)^2}{(2i-1)(k^2 - 1)} \quad (3)$$

Maximum ring number k take 10, 20 and 30, respectively, according to the type (3), get range of $\frac{M_i}{M_1}$ value in table 1.

As can be seen from table 1, the ring C_i 's node energy consumption is much larger than energy consumption of other ring, their energy depletion is empty, the main reasons for the formation and the maximum number of ring to change does not affect the law above. Therefore, we can add small amount of long chain, sent parts of data of the other ring directly to the Sink, and reduces the energy consumption of the ring C_i , prolong the life cycle of the network.

B. Energy Consumption Analysis of Wireless Networks based on the World

Because node energy consumption and transmission distance α in the direct ratio of 2 ~ 6 times, therefore, the presence of long chain can increase the energy consumption of the corresponding node. Through the analysis of the front, we found as far away from the Sink, the sharp decline in energy consumption of sensor nodes, energy cavity mainly appeared in the circle, near the Sink, especially the first circle C_1 . Therefore, through connecting long-range, increase far away from Sink's node energy consumption and reduce near the Sink's node energy consumption, so as to realize the energy equilibrium, prolong the life cycle of the network. However, how much long chain should be increased, where to put the extended chain most favorable, these long chains have much impact on the network life cycle? The following analysis.

Considered there is long chain of certain ratio $(1 - \lambda_i)$ within the ring c_i ($2 \leq i \leq k$), through these long chains, the corresponding node can communicate directly with the Sink, and these nodes within the data don't need a ring C_1 node forwarding, namely the amount of data from the I to the k ring produced need the amount of data node receiving and forwarding in ring C_1 for $q\pi c \lambda_i (r_k^2 - r_{i-1}^2)$, node in ring C_1 should forwarded the amount of producing data from 1 to $I-1$ ring $q\pi c r_{i-1}^2$, node in ring C_1 needs to receive also from 2 to $i-1$ ring's produced the amount of data $q\pi c (r_{i-1}^2 - r_1^2)$, as a result, the total energy consumption of the ring C_1 .

$$M_{total} = M_{tdi} + M_{ertl}$$

$$= q\pi c \lambda_i \left(r_k^2 - r_{i-1}^2 \right) \left(\partial_1 + \partial_2 r_1^a \right) +$$

$$q\pi c r_{i-1}^2 \left(\partial_1 + \partial_2 r_1^a \right) +$$

$$q\pi c \lambda_i \left(r_k^2 - r_{i-1}^2 \right) \partial_3 + q\pi c \left(r_{i-1}^2 - r_1^2 \right) \partial_3$$

The average energy consumption of each node in the ring C_1 is

$$M_{t1} = \frac{M_{total}}{q\pi r_1^2}$$

$$\left(q\pi c \left(\lambda_i r_k^2 - \lambda_i r_{i-1}^2 + r_{i-1}^2 \right) \left(\alpha_1 + \partial_2 r_1^a \right) + \right.$$

$$\left. \partial_3 \left(\lambda_i k^2 - \lambda_i (i-2)^2 + (i-1)^2 - 1 \right) - \left(\partial_1 + \partial_2 r_1^a + \partial_3 \right) \right)$$

$$= \frac{\left(\lambda_i k^2 - \lambda_i (i-1)^2 + (i-1)^2 - 1 \right) \left(\partial_1 + \partial_2 r_1^a + \partial_3 \right)}{k^2 \left(\partial_1 + \partial_2 r_1^a + \partial_3 \right)}$$

$$= \lambda_i + \frac{(1 - \lambda_i)(i-1)^2 - 1}{k^2}$$

In the presence of long chain, node energy consumption reduced in the ring C_1 :

That is

$$\lambda_i + \frac{(1-\lambda_i)(i-1)^2 - 1}{k^2} \leq \frac{M_{ii}}{M_1} \leq \lambda_i + \frac{(1-\lambda_i)(i-1)^2}{k^2} \quad (4)$$

The number of long chain in sensor network is limited by energy consumption and achieve, unfavorable overmuch. Set a fixed number of long chain in a network, denoted by n, these long chain appear within the ring C_i , ring in the C_i , total number for T_i , the ratio of long chain for $(1-\lambda_i)$, then

$$T = (1-\lambda_i)T_i = (1-\lambda_i)q\pi(r_i^2 - r_{i-1}^2) = q\pi r_1^2 (1-\lambda_i)(2i-1)$$

so,

$$\lambda_i = 1 - \frac{T}{q\pi r_1^2 (2i-1)} \quad (5)$$

It can be seen from type (5) that λ_i increases with the increase of the I . According to the type (4) and type (5), when the long chain appeared in the ring far away from Sink, the average energy consumption E11 within ring C_1 node will increase, namely E11 increases with the increase of the I . As a result, most good long chain arrangement within the ring C_2 , which is beneficial to prolong the lifecycle of the network, at the same time, because the long chain length is shorter, and easy to implement.

TABLE I. THE CHANGE TREND OF $\frac{M_i}{M_1}$ OF VALUES

Number of rings i	Maximum number of rings $k=10$		Maximum number of rings $k=20$		Maximum number of rings $k=30$	
	upper limi	Lower limit	upper limit	Lower limit	upper limit	Lower limit
1	1	1	1	1	1	1
2	0.334	0.334	0.335	0.331	0.335	0.331
3	0.196	0.188	0.201	0.196	0.120	0.197
4	0.133	0.124	0.142	0.138	0.138	0.138
5	0.092	0.081	0.108	0.105	0.108	0.106

Ring C_2 node's receives the energy consumption for $\partial q\pi c(r_k^2 - r_2^2)$, make in ring C_2 totally has N_2 sensor nodes, so there are λt_2 to transfer data to a sensor node within the C_1 , sending the energy consumption for $q\pi c(r_k^2 - r_1^2)\lambda(\partial_1 + \partial_2 r_1^a)$, $(1-\lambda)T_2$ sensor node transfer data directly to the Sink through long-range connection, for energy consumption $q\pi c(r_k^2 - r_1^2)(1-\lambda)(\partial_1 + \partial_2 r_2^a)$, so the ring C_2 'S total energy consumption:

$$\begin{aligned} M_{2total} &= M_{ix2i} + M_{rx2i} \\ &= q\pi c(r_k^2 - r_1^2)\lambda(\partial_1 + \partial_2 r_1^a) + \\ &= q\pi c(r_k^2 - r_1^2)(1-\lambda)(\partial_1 + \partial_2 r_2^a) + \\ &= \partial_3 q\pi c(r_k^2 - r_2^2) \end{aligned} \quad (6)$$

Each node's average energy consumption within the ring C_2

$$\begin{aligned} M_{2i} &= \frac{M_{2i/total}}{q\pi(r_2^2 - r_1^2)} \\ &= (q\pi c((\lambda r_k^2 - \lambda r_1^2)(\partial_1 + \partial_2 r_1^a) + \\ &= (r_k^2 - r_1^2)(1-\lambda)(\partial_1 + \partial_2 r_2^a) + \partial_3 r_1^2(k^2 - 4))/ \\ &= (3q\pi r_1^2) \\ &= (c((\lambda k^2 - \lambda)(\partial_1 + \partial_2 r_1^a) + \\ &= (k^2 - 1)(1-\lambda)(\partial_1 + \partial_2 r_2^a) + \partial_3(k^2 - 4))/3 \end{aligned}$$

In order to alleviate energy hole and prolong the network life cycle, need nodes in ring C_1 and C_2 have the same life, namely, their energy consumption are the same, therefore, make $M_{ii} = M_{2i}$, get it

$$\lambda = \frac{(k^2 - 4)(\partial_1 + \partial_3) + (2^a k^2 - 2^a - 3)\partial_2 r_1^a}{3(k^2 - 1)(\partial_1 + \partial_3) + (k^2 - 1)(2^a + 2)\partial_2 r_1^a} \quad (7)$$

For a specific network, the related parameters are taken in type (6) λ value can be calculated out.

C. Life Cycle and Long Chain of the Network

First calculate the isometric transmission network and small world network life cycle, and compare them, and then describes the implementation method of the long chain in the small world network.

According to the type (2), get the isometric transmission network life cycle, e for the node energy:

$$T = \frac{r}{M_1} = \frac{m}{c(k^2 \partial_1 + k^2 \partial_2 r_1^a + (k^2 - 1)\partial_3)}$$

When in ring C_2 , there are $(1-\lambda)T_2$ long chain, according to the type (4), we have

$$\lambda - \frac{\lambda}{k^2} \leq \frac{M_{ii}}{M_1} \leq \lambda + \frac{(1-\lambda)}{k^2}$$

So, the life cycle of small world network

$$\frac{mk^2}{M_1[\lambda(k^2 - 1) + 1]} \leq T_i = \frac{e}{M_{ii}} \leq \frac{mk^2}{M_1\lambda(k^2 - 1)}$$

That is

$$\frac{k^2}{\lambda(k^2 - 1) + 1} T \leq T_i \leq \frac{k^2}{\lambda(k^2 - 1)} T \quad (8)$$

According to type (8), after joining part of the long chain within the ring C_2 , at least improve network life cycle Times, for a specific network, the related parameters are taken in type (7) with type (8)

$\frac{k^2}{[\lambda(k^2 - 1) + 1]}$ value can be calculated out.

We achieve long-range connection through two transmission radius of sensor nodes decorated within the ring C_2 . A maximum transmission radius is r_1 , only

transmit the data to the nodes within C_1 , another maximum transmission range is $2r_1$, data can be directly to the Sink, Sink as the origin of coordinates, the circular area is divided into four quadrants. Long-range connection initiated by the Sink, at regular intervals, Sink randomly select has a larger residual energy from each quadrant $(1-\lambda)T_2/4$ nodes to establish a long-range connections, namely uses the communication radius $2r_1$ of these nodes to transfer data directly to the Sink. Other nodes within the C_2 can only through nodes of C_1 transmitting data to the Sink. When the remaining energy of a node inside the C_2 below average remaining energy of nodes in the C_1 , it cannot be selected as a long-range connected nodes, thus preventing nodes within the C_2 earlier run out of energy than nodes within C_1 .

As can be seen from table 1, ring C_1 node energy depletion is empty, the main reasons for the formation, therefore, it can add a small amount of long chain, parts of the other ring data is sent directly to the Sink, and reduces the energy consumption of the ring C_1 , prolong the life cycle of the network. According to the type (4) and (5), when the long chain appeared in the far away from the Sink ring, ring C_1 node within the average energy consumption of M_{il} will increase, namely M_{il} increases with the increase of number of ring I. As a result, most good long chain arrangement within the ring C_2 , which is beneficial to prolong the lifecycle of the network, at the same time, because the long chain length is shorter, and easy to realize. According to the type (8), after joining part of the long chain within the ring C_2 , network lifetime at least improve $k^2/[\lambda(k^2-1)+1]$ times, if λ takes 0.5 approximation, the network life cycle is about 2 times. Through the analysis above we can see that in the ring C_2 to join part of the long chain can effectively improve the network life cycle, and compared with other related methods, it is very easy to implement in practice.

IV. EXPERIMENTAL SIMULATION AND ANALYSIS

On the theoretical analysis above from the perspective of the experimental results for validation, and analyze the maximum ring number of k and ring width r_1 influence on network life cycle, through comparing with equidistance transmission network, shows that our method can significantly improve the life cycle of the network. Due to the energy hole problem is mainly focus on the network lifetime, therefore, the simulation experiment mainly examination, analyzed the network life cycle.

In a simulated experiment, the parameter $\partial, \partial_1, \partial_2$ and ∂_3 value: $\partial=2$, $\partial_1=46 \times 10^{-9} j/bit$, $\partial_2=12 \times 10^{-13} j/bit/m_2$, $\partial_3=136 \times 10^{-9} j/bit$. node initial energy $e=100J$, per unit time to produce the amount of data that $b=104 bit/s$, maximum transmission

radius for r_1 , ordinary nodes maximum transmission radius of the long chain node for $2r_1$, the simulation node number is respectively 1000, 2200, 4000, 6000 nodes (respectively corresponding to different values of k) evenly distributed sensor nodes. We assume that the MAC layer is the ideal; there is no conflict and additional energy consumption caused by data retransmission. In C_2 , there are $(1-\lambda)T_2$ sensor nodes and Sink communication directly, λ value according to the type (6).

Sensor area is divided into several such wide circle, circle is equal to the width of r_1 , the largest circle radius is fairly r_k , k for the total number of network circle. T_l/T for small world network and isometric transmission ratio of the life cycle of the network, reflects the degree of improving the life cycle. After a small amount of long chain can be seen from the figure 2, when r_1 must have little impact on life cycle to improve degree of network scale, it shows that our method is not only suitable for small networks, and also applies to large-scale sensor networks, has a good scalability. This conclusion is accordance with type (7).

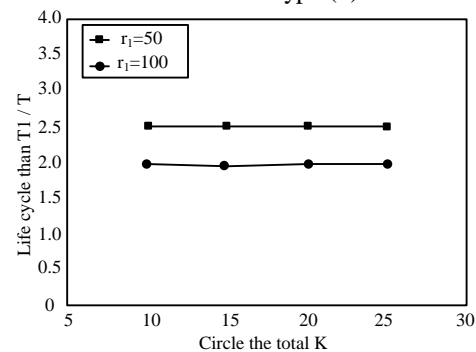


Figure 2. The influence of network scale

As can be seen from the network energy model, data from one node to another node, the energy consumption is proportional to r_1

Power ∂ According to this network model, when the k value is sure, data transmission, hop is certain. Thus, k at constant values, different r_1 value reflects the single hop transmission distance (or ring width) influence on network life cycle. As can be seen in figure 3, single hop transmission distance has a great influence to r_1 's increase in the network life cycle, T_l/T is gradually slow down with the increase of r_1 . This is mainly because with the increase of r_1 , long chain transmission will consume more energy, increasing the total energy consumption of the network. It also illustrates the most good long chain arrangement within the ring C_2 ; especially in a single hop transmission distance is large. Further analysis shows that when the r_1 increases a certain value, the T_l/T basic remains the same, close to the value of λ .

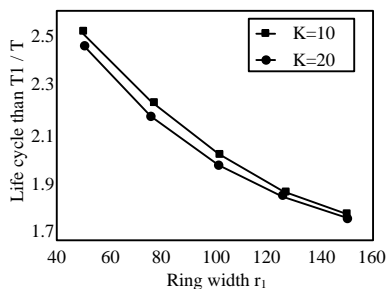


Figure 3. Influence of single hop transmission distance

Figure 4 reflects the number of long chain $(1-\lambda)T_2$ influence on network life cycle. The $k=10$, r_1 , respectively take 80 m and 100 m. It can be seen that add long chain can significantly improve the network life cycle, and with the increase of λ , gradually improve the network life cycle, when λ reaches a certain value of λ_{ij} , however, with the increase of λ , instead, gradually reduce the network life cycle. This is because the presence of long chain will be node energy consumption, increase ring C_2 when $\lambda < \lambda_{ij}$, node failure occur within the ring C_2 , along with the rising of the λ , more data to ring C_1 , therefore, to extend the life cycle of the node in the ring C_2 . When $\lambda \geq \lambda_{ij}$, node failure occurs within the ring C_1 , along with the rising of the λ , as more data to ring C_1 , ring C_1 node within the life cycle becomes shorter. When $\lambda \geq \lambda_{ij}$, ring of C_1 and C_2 nodes have the same average energy consumption, at this point, the network has the maximum life cycle.

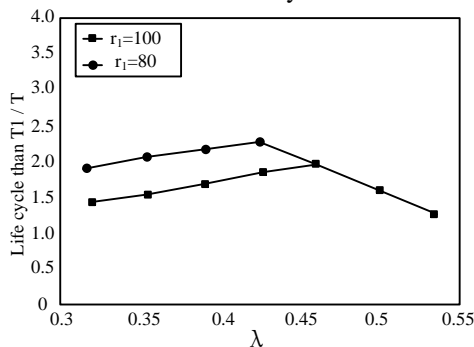


Figure 4. Influence of the number of long chain

Figure 5 reflected nodes energy consumption increases in the ring C_2 due to the presence of long chain. Which $k=10$, r_1 take 80 m and 80 m respectively, according to the figure 4, λ respectively take $0.45(r_1=80)$ and $0.50(r_1=100)$, M_{2l}/M_2 represents in the ring C_2 , the ratio between presence long chain and there is no long chain node energy consumption. Ring C_2 data through long-range connection sent directly to the Sink, due to the increase of transmission distance, lead to the increase of the ring C_2 nodes energy consumption. On the other hand, reduced ring C_1 data's volume, reduce the energy

consumption of the ring C_1 node. Through the analysis of the front, we know that energy hole is mainly within the ring C_1 , therefore, to add long chain extending the network lifetime, however, it is increased with the cost of ring C_2 node energy consumption.

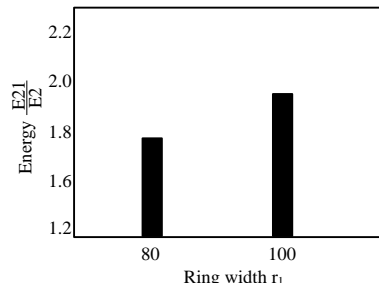


Figure 5. Ring C2 node energy consumption when there is a long chain

Addressed by the uneven distribution of node energy hole, but because in most cases, the node distribution is random, local node density is difficult to control; the no uniform node distribution strategy in practice is difficult to achieve. Through the ant colony optimization algorithm to determine the transmission radius of each node, the complexity of the calculation of node mobility is more sensitive and constrains their practicality. Therefore, we compared with “Using mobile relays to prolong the lifetime of wireless sensor networks”. By joining in the ring of C_2 and C_1 a mobile relay to achieve the purpose of ease energy hole. Figure 6 reflects the influence of the two methods of network life cycle. Among them, the T_l/T_{mr} represents the proposed method and the ratio of the life cycle of the above methods, $r_1=80m$, node density $p=4$, k respectively take 5, 7, 9, 11, 13 and 15.

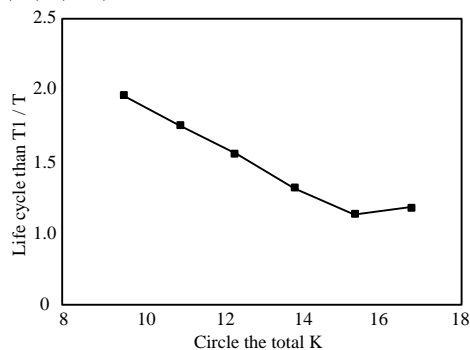


Figure 6. Comparison of network life cycle

It can be seen from the figure 6, the maximum number of ring k is small, our algorithm is superior to the literature of “Using mobile relays to prolong the lifetime of wireless sensor networks”. In addition, his method increases the cost in the construction of the network also bring larger network transmission delay, and cannot be used in many applications of mobile relay, for example, the complex terrain environment, therefore, its application has great limitation. And the proposed method not only have small cost of network building and

node deployment, simple communication mechanism, environmental impact on the system is small, and easy to implement, has the very good practical value.

V. CONCLUSION

In sensor networks, it is easy to form energy hole around the Sink. Energy hole allows data could not be served to Sink and make the end of the network life cycle, then still have a lot of the remaining energy in the network. In order to solve the problem of energy hole, people put forward many methods. Different from existing research, this paper studies and proposes a solution based on small world energy hole algorithm, by adding a small amount of long chain in the network, decrease data forwarding near the Sink node to prolong the life cycle of the network. This paper analyzes the different positions of the long chain influence on network life cycle, and also analyzed theoretically the best proportion of long chain and its influence on network life cycle. The proposed method can improve the network life cycle, and it is easy to implement in practice.

REFERENCES

- [1] S. H. Tang, M. C. Chen, Y. S. Sun, et al. A spectral efficient and fair user-centric spectrum allocation approach for downlink transmissions. *IEEE, Globecom.*, 2011 pp. 1-6.
- [2] Jeong Cheol, Kim Hyung-Myung, Song Hyoung-Kyu, et al.. Relay Precoding for Non-Regenerative MIMO Relay Systems with Partial CSI in the Presence of Interferers. *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, Apr. 2012, pp. 1521-1531.
- [3] Zhiyong Zhang, Muthucumaru Maheswaran, Guest Editorial. *Journal of Multimedia*, Vol 7, No 4 (2012), pp. 277-278.
- [4] Tang X, Xu J. Optimizing life time for continuous data aggregate on with precision guarantees in wireless sensor networks. *IEEE/ACM Transactions on Networking*, 2008, 16 (4) pp. 904- 917.
- [5] S. Li, Y. Geng, J. He, K. Pahlavan, Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. 2012 pp. 721-725.
- [6] Yang Yu, Krishnamachari B. Energy-latency trade-offs for data gathering in wireless sensor networks // *Proceedings of the INFOCOM 2004*. Prasanna V K, 2004, 1 pp. 7-11.
- [7] Lian J, Naik K, Agnew G. Data capacity improvement of wireless sensor networks using non-uniform sensor distribution. *International Journal of Distributed Sensor Networks*, 2006, 2(2) pp. 121-145
- [8] Chiu Eddy, Lau Vincent. Cellular Multiuser Two-Way MIMO AF Relaying via Signal Space Alignment: Minimum Weighted SINR Maximization. *IEEE Transactions on Signal Processing*, vol. 60, no. 9, Sep. 2012, pp. 4864-4873.
- [9] Y. Geng, J. He, H. Deng and K. Pahlavan, Modeling the Effect of Human Body on TOA Ranging for Indoor Human Tracking with Wrist Mounted Sensor, *16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Atlantic City, NJ, Jun. 2013.
- [10] Chen G, Li C F, Ye M, Wu Jie. An unequal cluster-based routing strategy in wireless sensor networks. *Wireless Networks (JS)*, 2009, 15(2) pp. 193- 207.
- [11] Wei-Ying Kung, Chang-Su Kim and C.-C. Jay Kuo, "Spatial and Temporal Error Concealment Techniques for Video Transmission Over Noisy Channels", *IEEE Transactions on Circuits and System for Video Technology*, vol. 16, no. 7, pp. 789-802, 2006.
- [12] Y. Zhao, D. Tian, M. M. Hannukasela, M. Gabbouj, Spatial Error concealment Based on Directional Decision and Intra Prediction, *IEEE International Symposium on Circuits and Systems*, Volume 3, pp. 2899-2902, 2005.
- [13] Dong-Eok Kim, Dong-Choon Lee, "Fault Diagnosis of Three-phase PWM Inverters Using Wavelet and SVM," *IEEE International Symposium on Industrial Electronics*, 2008, pp. 329-334.
- [14] Jie Gao, Vorobyov S. A., Hai Jiang, et al.. Sum-Rate Maximization with Minimum Power Consumption for MIMO DF Two-Way Relaying Part II: Network Optimization. *IEEE Transactions on Signal Processing*, vol. 61, no. 14, Jul. 2013, pp. 3578-3591.
- [15] Stoica I, Morris R, Liben-Nowell D, et al. Chord: a scalable peer-to-peer lookup protocol for internet applications. *Networking, IEEE/ACM Transactions on*, 2012, 11(1) pp. 17-32.

Yuting LU, born in 1980, graduated from Jiangsu Normal University, and now is a lecture. His main research area is computer network.

Weiyang WANG, born in 1987, graduated from the University of Iowa, and now is a doctoral candidate in Learning and Cognition Program at the Florida State University. His main research area is high technology and learning process.

Identification Method of Attack Path Based on Immune Intrusion Detection

Huang Wenhua

School of Telecommunications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, China

Yishuang Geng

Center of Wireless Information Network Studies (CWINS), Worcester Polytechnic Institute, Worcester, MA, 01609, USA

Email: yugeng@wpi.edu

Abstract—This thesis takes researches on the immune intrusion detection and IP trace back technology. To find out the network data features of the real-time analyses, the distributed immune intrusion detection system and the packet marking theory are used; to guide the dynamically processing of path signs technology, the immune intrusion detection system is used; what's more, to dynamically adaptive different methods of characteristics of network data, the path signs technology is adopted. After that, the attack paths can be quickly identified to provide path information for feature detector on attack path in the immune intrusion detection system. Experiment results show that this scheme can quickly reconstruct the attack path information, and the performance on the aspects of the convergence is with efficiency rate and false positive rate, which is superior to the current probabilistic packet marking algorithm and can provide characteristic path information for immune intrusion detection system.

Index Terms—IP Packet Marking; IP Trace Back; Adaptive Mechanism; Artificial Immune

I. INTRODUCTION

Computer network security technology is a multi-disciplinary, multi-disciplinary comprehensive discipline, including traditional firewall technology, access control technology, encryption technology, intrusion detection technology, IP trace back technology and so on [1]. Network security technology research includes dual nature, that is, offensive and defensive, in which the immune intrusion detection and IP trace back technology represent the technology of both ends of the network security technology. Although it is not ripe yet at this stage, as an important direction of development network security technology, it received sustained attention by experts and scholars. The thought of immune intrusion detection system derives from the recognition and treatment of the "non-self" material in biological immune system [2-4]; this system does not rely on a large number of signatures to determine whether invaded or not, but the characteristics of normal network flow are modeled; once the current network characteristics are not within the normal range, the system will consider that the

potential attacks are discovered. So the immune intrusion detection system has a good dynamically adaptive capacity and high sensitivity for unknown attacks, and it is very suitable for the current changing network environment [5]. On the other hand, in order to find the perpetrators of cyber attacks and to punish them, and making the deterrent effect on the potential attackers, attack source tracing technology has steadily developed [6].

All along, the attack source tracing technology has always been researched as a stand-alone network security technology, but it is difficult to apply the results of their research. Among the reason, firstly, it is the flaw when the IP protocol is designed, because the current tracking technology just relies on changing the IP protocol, which is difficult to increase their effectiveness; secondly, it is the tracking strategy for relatively fixed algorithm adopted by tracing algorithm, and in order to improve this strategy, the artificial immune intrusion detection technology has been used as a typical anomaly-based intrusion detection system study since proposed in 1994 [7-9]. This technology does not rely on signatures, but has the ability to detect new, potential network attacks to protect the host against invasion. Since then, the American University of Memphis proposes a multi-agent detection system based on artificial immune intrusion, in which the intrusion detection system is no longer confined to the local host and the each agent can co-process to improve the robustness of the system, but there is a problem of not fully covered [10]. Gonzalez firstly proposed using the real value to represent the network data flow characteristics, in which the extracted features can be easy for people to understand and be easier to classify data [11]. In 2010, Dr. Jamie Twycross proposed the concept of second-generation artificial immune, which tries to combine the innate immunity and adaptive immunity and proposed the model to assess and test the immune algorithm, but the model is still difficult for practical application [12]. There are some problems of current immune intrusion detection system like inadequate autologous collection, overburdened local intrusion detection systems and so on, but it has a highly

sensitive for unknown new attacks, a rapid response capability for the known attacks, and a strong change characteristic to adapt the network characteristics, which has the application value [13-14].

Probability p ; when the attacking the host sends a large number of attack packets to the victims host, the victim host can refractor the attack path through the receiving information on data packets [15-16], but the method presents the problem of the weakest chain [17], and the reconstruction requires a lot of data packets with a \log 2000 or so, Trace back technique began to receive the academic attention [18]. In the same year, Stone R. proposed a technology through the recursion method to trace back the attack source, but this technology needs to get a router management authority, and at the same time only able to launch the attack when attacking, which has application value. Then, Burch H. proposed a link test tracking strategy method based on denial of service, but this method itself is a huge burden to the network [19]. Thereafter, Bellovin proposed IP trace back technology based on ICMP protocol, which applies a new "iTrace Report" to trace back attack source, and through the router sending these messages to help the victim host identifying attack paths, but this method will make the overload router load, which affects its normal performance, and the ICMP report is easy to be filtered by the security policy [20]. The current researched IP trace back technology mainly focuses on the improvement of marking method to the probabilistic packet, which is proposed by the Savage. Routers are required to mark its own address information for each passing packet with a fixedw efficiency high false alarm rate and poor immunity. Although the passive tracking algorithm represented by packets marking has been conducting improvement, there are various restrictions for used alone packet marking algorithm or improved packet marking algorithm due to the inherent flaws and fixed format of IP protocol, while the "iTrace" method based on ICMP protocol implementation provides a good idea, and the implementation based on router needs to be improved.

In general, all the existing intrusion detection systems focus on the discovery and prevention of attacks. Although most network-based attacks can be detected, all of them can not provide a real source tracking of attack. In order to avoid exposure the usual approach of the attacker firstly break a system, and then use it as a platform to use the approach of network jump (H () P) to attack another system. Because it is difficult to track the invasion, many cases are attacked after many jumps before reaching the real target. Additionally, currently tracing algorithm do not with dynamic adaptability. Packet marking algorithm is not based on packet characteristics, or the characteristic dynamic detection and tracking of data traffic, but always uses a fixed markup policy, which led to a large number of normal data packets are marked. On the other hand, when the host is tracing attacked it seems that the approach is inefficient.

In addition, the current tracing algorithm does not have the dynamic adaptability. Packet marking algorithm is not based on packet characteristics or the detection and tracking of dynamic characteristics of the data flow, but always adopts a fixed marking policy, which marks a lot of normal data packets, but inefficient when the attack host is tracing.

Innovations in this thesis:

(1) Many researches for immune intrusion detection and combined technology of IP trace back are studied. By using the distributed immune intrusion detection system and the packet marking theory, network data features of the real-time analyses in immune intrusion detection system is used to guide the dynamically processing of path signs technology, and then the path signs technology can dynamically adaptive different methods of characteristics of network data.

(2) Through consulting the distributed processing ideas of IDRA and DDoS, combined with a distributed immune intrusion detection systems theory, this design implements of a path signs technology based on distributed open-immune intrusion detection system. Technology implementation relies on immune response server located in switched networks. On one hand, this server can make immune processing to the passing attack packets; on the other hand, it can make path marking to the suspected potential attacks, so the IP trace back algorithm is dynamic, adaptability and resources saving.

(3) Model and algorithm of attacking path signs based on distributed open-immune intrusion detection can further mark passing the data packet path to provide technical foundation for immune feature detector aimed at the single attack in the immune intrusion detection system. Thus this can deal with the attack packet in the attacking path and avoid the situation that victim host suffers the big attack packet and then refuses to serve. Experimental results show that the proposed scheme is better than the traditional tracking algorithms in convergence efficiency, false alarm rate and so on. Although supports of the immune response server are needed, the server is not located for the path signs, and its main function is an immune node to process the attack packets in switched network.

II. PROPOSED SCHEME

The combination of distributed immune intrusion detection system and IP trace back technology will place distributed open-immune response servers in the key path location in the switched network to serve the local immune intrusion detection system. On one hand, network data features of real-time analysis in immune intrusion detection system can be used to unfold the IP trace back algorithm, and then the algorithm has a dynamic adaptability, which can quickly find out the specific attack path. On the other hand, since the characteristics of a single attack in the shape-space are relatively continuous, just as shown in Figure 2-1, and the subsequent packets will be marked with a specific path signs, so that the immune intrusion detection system can attack data for a specific path selection method with a

positive culture to generate immunoassay device, and the path makes response to processing attack packets of the subsequent.

For the convenience of discussion and research of the algorithm and model, the prerequisites are as follows:

- a. The data packet may be lost or out of order during transmission.
- b. The data packets path of the same source hosts and the destination host are basically stable.
- c. The attacker can generate any desired packets.
- d. The router is credible, but limited resources.

The design of model and algorithm of the attack path signs based on distributed open-immune intrusion detection is mainly divided into the following three steps:

1) Modeling and building the overall framework to determine the location of the immune response server.

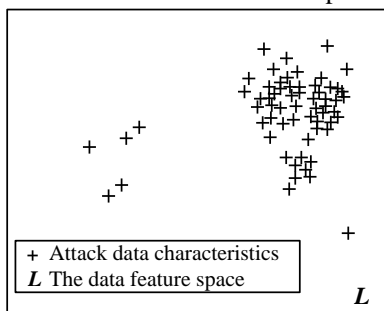


Figure 1. Mapping of the single attack features in shape-space

- 2) Determining the steps of the module and specific work.
- 3) Redefining IPv4 packet header and using packet to mark.
- 4) Designing algorithms in each module.

A. System Model and Module Design

Technical models building and core concept of the path sign based on immune intrusion detection system are locating the analysis of the immune intrusion detection system and response module into the transmission network, and after the real-time analysis of network data, the attack source trace back algorithm module are dynamically used according to the data characteristics of decision-making.

By referring to the “Distributed Intrusion Detection System and Distribution open intrusion detection and response framework” IDRA technology [12] and DDoS distributed processing thinking [13], an open track-type immune response server is located in the critical path in the transmission network (such as network border routers) shown in Figure 2-2; the server does not belong to a separate intrusion detection system, but can provide services for any immune intrusion detection system and work together; the server match the passing packets with the immune detector, which is used to determine whether the packet needs to respond, to treatment or trace back attack path, and then cooperate with the existing various routers to realize route signs; tracing algorithm is converged quickly and promptly traced to the desired position.

In addition, the set of this server can release a heavy relay on the route resources of previous various algorithms, which makes the router is only in collaborative work, rather than proactive, high loading trace back status, so the network status can be better guaranteed.

Attack source tracing algorithm based on immune Intrusion detection system can be divided into tracking / responding server algorithms, router algorithm and path reconstruction algorithm. The module design is shown in Figure 2-3:

Tracing algorithm only trace back unusual packet. Though to achieve synergies algorithm needed by routers is seemingly relatively complex, the vast majority of normal packets transmitting in the network will not be processed. Compared to other tracing algorithm it can reduce the load of routers and network, however, immune response server needs to have better performance, and can identify attacks path to provide the path packet information for the immune intrusion detection systems.

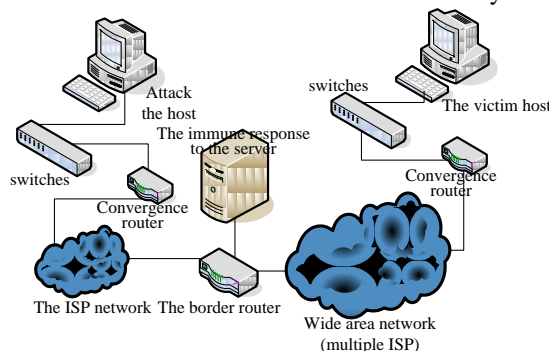


Figure 2. Location diagram of the immune response server

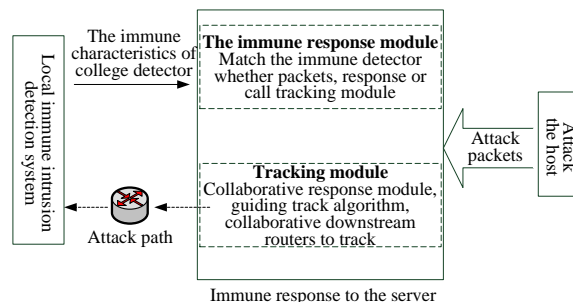


Figure 3. Server module of immune response

Overall model steps based on immune distributed open systems are described as following:

- 1) Immune response server uses immune memory detector and feature detector between nodes to analyze passing network packets. As for packet that not be matched by these two detector, it will be normally forwarded, and the packets matching feature detector are deleted, while for the packets which is matched with the memory detector will take tracking algorithms.
- 2) Immune response server applies path signs F to the victim host immune, and at the same time submit the packet information judged as abnormal.
- 3) Immune response server sends 64 tracking packets to the victim host for the downstream routers to mark and the packet is the sign of the applied path.

4) Downstream router identifies the packet itself should be labeled and records the IP into the two packets.

5) Victim host verifies the packets and reconstruct the path from the immune response server to its own point.

6) Victim host informs the immune response server that the reconstruction is completed, and accepts the attack marked and traced by the improved probabilistic packet out of the "control zone".

7) Victim host informs immune response server among the reconstructed paths that the packets marked as F not be marked repeat.

8) Thereafter, the immune response servers mark all the packets through themselves and sending to the victim host as path marker F, so that the local immune intrusion detection system can generate the immune feature detectors for this path.

9) Immune intrusion detection system of the victim host distinguishes and marks packets as autologous (normal) or non-autologous (attack) data.

10) Immune intrusion detection system of the victim host generates the immune feature detector to non-autologous attack data through negative selection algorithm.

11) The generated attack immune feature detector is sent back to feature library of immune response server, and recognizes it in its life cycle and deals with the subsequent attack packets. After the life cycle of immune feature detector is over, it will be placed into the memory abnormality detector for detection to find the potential attacks.

Wherein, 1-8 steps are designed for the path signs algorithm based on immune intrusion detection system and they are the key points of this paper.

B. Algorithm Design

1) Immune Response Algorithm

The algorithm is used in the immune response module of immune response server. Algorithm uses the "memory" immune detector and "characteristic" immune detector in the immune nodes to compare with the passing packets, in where the "memory" immune detector is the further heritable variation of "characteristic" immune detector after its life cycle is over, while the "characteristic" immune detector comes from the training and submitting of the private immune intrusion detection system of attack host. Once the packet is matched with the "characteristic" immune detector, it considers that the packet belongs to attack behavior, which can be deleted; if it matched with the "memory" immune detector, it considers that the current packet is similar with the previous attack behavior, which belongs to the potential attack and can be traced.

Of course, in order to further improve the recognition abilities of the immune response nodes to the potential attack, "memory" detector can accept this immune response server via local data features to establish a long-term normal model, and after that via the "positive selection" algorithm to generate the "non-self" immune characteristics, which can be processed by the specific situations of the dealing performance of the immune

response server and feature modeling. This thesis will not discuss further here.

Immune response algorithm is as follows:

Step1 Processing the next packet and controlling the packet in regional location 1

Step2 Extracting the features in the packet

Step3 Matching the features of the packet and immune features detector submitted in the immune intrusion detection system with the same destination IP, which successfully delete the packet and back to Step 1, otherwise the progress continues

Step4 Matching the characteristics of the packet and the set of the memory detector; if fail, back to Step 5, or to Step 7

Step5 Checking to find out whether exists the path sign F with the same destination IP; if exist, F will be marked into the path and be forwarded; if not, it will be directly forwarded.

Step6 Back to Step 1

Step7 the packet information is recorded and then submitting it to the tracking module for processing path signs via controlling the tracking algorithm.

Step8. Back to Step 1

2) Controlling Tracing Algorithm

The algorithm is used in the tracking module, which can be called by the immune response module. For safety, controllability and feasibility considerations, routers within the control area are not active sign address, but it will generate a particular packet for the cooperating of the downstream routers by the immune response server. It can be called as "path tracing packets", and it is recognized and marked its address in turns by router. Since there are only 16-space in each "path tracing packets" any host to victim host, so there are only 64 "path tracing data packets" are need for the immune response serve can carry address information, there are two packets for each route to mark itself address, and via the current monitoring of network packets it is learned that there are less than 32 routes in the path from. In order to enable the downstream router determining its own IP address to which the trace data packet marker, a contrast marker is needed in the tracing packet. Because there is no need for compared market after the IP is marked, the compared market and marked IP can share a head, that is, head marking segment. Controlling and tracking algorithm uses the TTL decrement feature to compare with the TTL via setting the marking segment TTL, which determines whether it need to be marked or not.

The specific algorithm is as follows:

Step1. Applying random path marker F to the immune intrusion detection system with the tracing destination IP address and submitting the abnormal packet information which second causes this tracing algorithm

Step2. Waiting for the path marker; if it is timeout or reject, it comes to the end; if it is successful, it will be continued

Step3. Generating 64 path tracing packets, in which one set of two, and in total there are 32 group; the TTL compared segment of two packets' IP header in each set are the same, in which from top to bottom 32 groups are

decreasing from 33 to 1; the IP offsets of two packets in each set are different, in which one is for 0 and the other is for 1; the tracing marker of the all 64 packets is set as 1, in which the distance between segments set as 0; the part of data in the packet are loaded into the path marker F.

Step4. Sending the 64 packets to the destination IP

3) Router Collaborative Algorithm

Collaborative algorithm on downstream routers is based on the different state of the data packets makes different movements. Firstly, the routes determine the controlling zone bits. If the bit is equal to 0 (not considered the forgery case), the packet do not enter the control area and can't control the tracing algorithm to track; in order to compensate the deficiency of the tracing algorithm in this segment (usually the first few jump in full path and no more than 5 jumps), the router with algorithm marked by the improved probabilistic packet marks the packets [14], and the tracing algorithm will be conducted when it is through by the first immune response server and can be sent to the victim host, thus the victim host will conduct the entire path.

Once the packet P are into the control area, collaborative routers judges P via tracking markers as a normal data packet or the tracing packet needed to be co-processed; when P is the tracking packet, compared the read TTL marked segment of the read P with the TTL of P, if it is different and its distance is not as 0, which indicates that the package has been marked by the upstream router, and it can be forwarded by adding 1 distance; otherwise, it will be directly forwarded.

When the read TTL marked segment of the read P and the TTL are the same, the router via IP offsets marks the first 16-bit or the last 16-bit of their own IP address into the compared field TTL, that is, the mark field. Since the IP is marked as the mark field, the meaning of the mark field changes from TTL comparison to IP marker. Meanwhile, the odd-even check of P mark the IP location via the filled markers. At this point, the router completes its coordination algorithms.

It is can be known from the above description that the value of the mark field is less than 32, and it should be as the comparing value of TTL value than the right; in order to prevent conflict, 10 hexadecimal value of the 16 random markers allocated by the victim host should be larger than 32.

Routers collaborative algorithm is as follows:

Step1. Dealing with the next data packet

Step2. If the packet control area bit is as 0, the packet is traced by the improved probabilistic packet marking algorithm and forwarded to Step 1; otherwise, the process is continuing.

Step3. If tracing mark bit of the packet is 0, the packet is forwarded; otherwise the process is continuing.

Step4. If the distance between the fields of the packet is not as 0, the distance between the fields will plus 1 and then the packet is forwarded; otherwise the progress is continuing.

Step5. If the compared field TTL of the packet is different with TTL, the packet is forwarded; otherwise the progress is continuing.

Step6. If the offset field of the IP in the packet is 0, the router mark its high 16-bit of into the marked IP field of the packet, and then the distance between the fields is set as 1 and the IP verifying field is filled in the verifying bit and be forwarded and then it is returned to Step 1.

Step7. Routers mark its low 16-bit of IP into the marked IP field of the packet, and the distance field is set as 1; the IP verifying field is filled in the verifying bit and be forwarded and then it is returned to Step 1.

After the victim host receiving 64 tracks packet, firstly it verifies if the random mark in the data field is sent by itself or not, quickly builds this path according to the contents of the packet and sends a confirmed report to the immune response server to inform that its tracking is complete.

When the immune response server receives the confirmed report, it is indicates that the controlling and tracking algorithm is complete. Thereafter, the immune response server can mark the generated path of the pass by probabilistic packet out of the controlling area and inform the victim host to conduct full path. More importantly, after that, the immune response server write the previous obtained 16-bit markers through the packet sending to the victim host, in which the victim host knows where these packets are derived from, so it is convenient to deal with or the specific immune detector is trained.

Since there is one or several immune response servers in a path, if all of them mark the random marker to the packet, it is only can be retained by the reserve closest to the victim host. In order to solve this problem, the victim host is needed to conduct a complete path tree according to the received packet locally; once several immune response servers are found in a path, the victim host sends packet to the middle of the servers and not repeats the furthest random markers. The method can effectively prevent the situation that all of the immune response servers don't mark the packets in the path caused by forged random markers by the attacker, makes the longest path tracing and has a more obvious characteristic of marking packet. In addition, when a random marker is no longer used in a certain period of time, the victim host empties all the information of this marker, and re-allocates this marker out.

III. ALGORITHMS ANALYSIS AND EXPERIMENTS

A. Algorithm Analysis

In the current mainstream tracking algorithm, the performance and behavior of router cause various shortcomings of tracking algorithm, such as heavy refactoring packets, high false alarm rate and other issues. Since the router needs to forward a lot of data, it is not suitable for a large number of computing and storing and the router can only operate the header of the IP packet. All of these make all the routers address mark limited in the header; what's more, marking a router address requires large amounts of data packet (header option is not practical). While the presence of The immune response server can solve or improve the above problem: based on this algorithm the normal packet will not be

path marked, but the network characteristics are real-time analyzed by the immune response server, so the “trace route packets” will be actively generated and forwarded to the downstream for router to identify and mark. Therefore, many problems can be solved like the slow convergence speed of probabilistic packet, complex router algorithm, without dynamic tracking strategy and the “weakest link”. While the memory detector in the immune response server can be evaluated according to changes of network characteristics, making the tracing algorithm has adaptivity to the network characteristics. The main indicators of evaluating the convergence rate, false alarm rate and immunity of an attack source tracing technology. An excellent IP trace back algorithm should try to make fast convergence, low false positives and fending off the attacker’s malicious interference. The convergence, the false alarm rate and immunity of the IP trace back algorithm will be theoretically analyzed based on the immune intrusion detection system as following.

1) Convergence The convergence of the algorithm needs to satisfy the convergence both of the path tracing algorithm in “control area” and the path tracing algorithm out of “control area”. In the “control area”, the convergence begins when the algorithm sends the 64-th data packets. The trace back algorithm begins when the feature similarity between a packet with the normal packet is small, i.e., the affinity of a “autologous Detector” in the tracking response server is less than a given threshold. Out of the “control area”, if the improved marking algorithm of probabilistic packet is adopted, there is less than 5 hops between the attacker and the first boundary router hops, and then the probability can be marked as $p = 1/5$. For an attack path of 5 hops out of “control area”, there are only a few dozen packets are required to conduct. On the other hand, the router out of the “control area” should belong to the unique ISP, and through log files recorded by the ISP, the packet sender can be easily found out.

2) False Alarm Rate The unique random marker of a 16-bit is used in the algorithm to track the path from the tracking response reserve to the victim host, as long as there are less than $2^{16} \times 32$ tracking response servers in the attack path, i.e. the 65504, the situation of false positives or false negatives will not exist. But once the hosts are suffered more than 65,504 tracking response servers’ attacks, the victim host can’t allocate the random markers to complete the trace, and then the situation of omissions is appeared. According to the practical attack experience, the attacking hosts are usually no more than a few hundred units even the DDOS attacks. The proposed program can completely avoid the appearance of false paths.

3) Immunity IP trace back algorithm based on immune intrusion detection does not use common data packet as the path information carrier, but use “path tracing data package” generated by the trace response server to forward the IP, which increases difficulties for the attacker to forge the path tracing information. The path marking algorithm based on the immune can mark the TTL value of the packet in the immune response host

corresponding to the records of the victim host. Thus if the attacker want to interfere a path refactoring it should be satisfied that the “random identifier” and the corresponding “TTL hops” are the same with the records of the victim host. It is very difficult for an attacker, and the probability of successfully forge a “path tracing data packets” is about $\frac{1}{2^{16} \times 32}$; even it is forged successful, the attacker is difficult to know. Therefore, anti-interference performance of this algorithm is really strong.

Through the using of simulations like NS-2, C-language and so on under Linux, the core parts of the algorithm are used to make experiment analysis, which will observe the algorithm’s performance of convergence time, false alarm rate and so on.

B. Convergence Experiments

In experiment 1 under the attack of a single path, the convergence time of the tracking algorithm is compared with different situation of the distance between the attack host and the victim host from 1 to 30. If the immune response server is located in the position of the five hops, the improved probabilistic packet out of the control area is adopted for tracking. The results are shown in Figure 4.

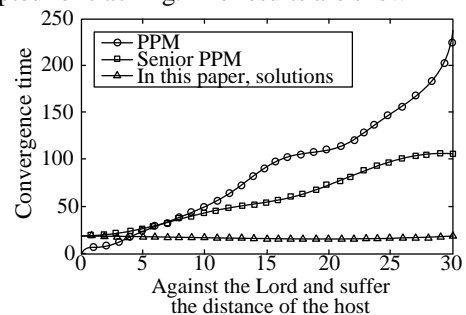


Figure 4. Distance and convergence of attack path

From the experimental results it can be seen that the existence of the immune response server can assist the system to mark the attack path; when the proposed marking algorithm of immune path are attacking, the convergence time is reduced to constant level compared to the other two algorithms; as long as the distance between the first immune response server in the attack path and the victim host does not exceeds 32 hops, the attack path can be quickly found. Although the convergence algorithm out of the control area is still using the marking improved program in probabilistic packet, it can be seen from the experimental results that the algorithm can converge quickly because of relatively small number of hops. The two kinds of marking schemes in probabilistic packet are in direct proportion of the number of hops on the convergence time; when the distance of the between the attack host victim and host is increased, the issue of weakest chain will stand out slowly; although the senior probabilistic packet is greatly improved on the convergence time compared to the marking scheme of basic probabilistic packet, it is really hard to apply in practice for it need to know the network topology.

C. Experimental of False Positives Rate

In experiment 2 there is a simulation on the false positive rate condition to three path marking algorithms in the multi attack paths. According to the actual attack situation, even if large-scale distributed denial services are attacking, the attack path is not more than 1000, so the experiment makes situation simulation within 1000 attack paths. The results are shown in Figure 5.

As can be seen from the experimental results, because there is no way to be taken against the problem of false positives, false positives situation based on PPM is very serious; when several attack paths are attacking, it is completely unable to reconstruct the attack path correctly. There is a great improvement of the advanced PPM on the aspect of false positives, but when the attack paths increase, the false alarm rate is rising rapidly. In the proposed immune-based path marking algorithm, because the immune response server is connected with the victim host and the marker of a unique path is determined, there is no false positives within the control area and false positive rate out of the control area is discussed according to different used methods. Low false positive rate makes the higher information entropy of the tracing path information. It has a reference value for the pertinence treatment of the subsequent immune intrusion detection system.

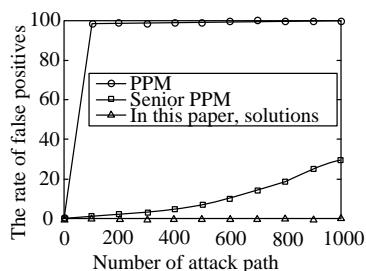


Figure 5. Experiment on the false positive with several attack paths

D. Experimental Conclusions

Through the above discussion and experiments it can be seen that the path marking technology based on immune intrusion detection has a better efficiency in searching the source of attack packets compared with other existing probabilistic packet parking. It can find out the source of the attack packets in a relatively fixed time, and the authenticity and credibility of the path information can be guaranteed; compared with the other two tracking algorithms it has more practical value.

IV. CONCLUSION

By reference to distributed processing ideas of IDRA and DDOS, combined with a distributed immune intrusion detection systems theory this thesis designs the path marking technology based on distributed open-immune intrusion detection system. The implementation of this technology relies on the located immune response server in switched networks. On the one hand, the server can make immune processing to the passing attack packets; on the other hand, it can launch path marking against the suspected potential attacks, and thus the making IP trace back algorithm has a dynamic

and adaptability, which save resources. In addition, this path marking technology can further mark the passing packets to provide technological base for the immune intrusion detection system to generate immune feature detector after a single attack and to process attack packets in the attack path, which can avoid denial of service condition caused by the great regularity attack packet aggregation on victim host.

Experimental results show that the proposed scheme is superior to the traditional tracking algorithms on the aspects of convergence efficiency, false positive rate and so on. Although the immune response server is needed to support the processing, the server is not only located for the path marker, but for processing attack packets as an immune node in the switched network.

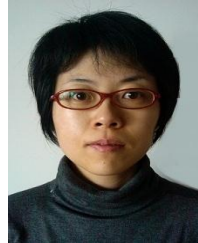
ACKNOWLEDGMENT

The paper is Supported by Research on the method for dynamic risk security assessment of IP network, Special Scientific Research plan of the department of education Shaanxi Province, 11JK0920, 2011.7 And the evaluation for dynamic risk of network security, Natural science fund of Shaanxi Province, 2009MJ8002-3, 2009.7

REFERENCES

- [1] Forrest S, Perelson A S, Allen L. Self-nonsel Self Discrimination in a Computer: *In Proceedings of IEEE Society Symposium on Research in Security and Privacy and Privacy, 1994. Massachusetts, USA, 1994* pp. 202-212.
- [2] Gonzalez F, Dasgupta D, Nino L F. A Randomized Real-valued Negative Selection Algorithm: *In Proceedings of Second International Conference on Artificial Immune Systems, 2003. Edinburgh, UK, 2003* pp. 261-272.
- [3] Twycross J. Stochastic and Deterministic Multiscale Models for Systems Biology: an Auxin-transport Case Study. *BMC Systems Biology*, 2010, 12(9) pp. 29-41.
- [4] Savage S, Wetherall D, Karlin A, et al. Practical Network Support for IP Traceback: Proceedings of the 2000 ACM SIGCOMM Conference, 2000. *Stockholm, Sweden, ACM, 2000* pp. 118-128.
- [5] Zhang Yongguang. Feature De duction and Ensemble Design of Intrusion Systems. *Computers & Security*, 2004, 20(8) pp. 1360-1361
- [6] Lee W K. Feature Selection of Intrusion Data Using a Hybrid Genetic Algorithm Approach. *Wireless Networks*, 2007, 13(6) pp. 459-460.
- [7] Zhang Yongguang. Intrusion De tectio n Techniques for Mobile Wireless Networks. *Wireless Networks*, 2003, 38(9) pp. 1869-1871.
- [8] Yu L, Liu H. Efficient Feature Selection via Analysis of Relevancy and Redundancy. *Journal of Machine Learning Research*, 2004(5) pp. 1205-1224.
- [9] Hu W M, Hu M, Maybank S. Adaboost based algorithm for network intrusion detection. *IEEE Transactions on Systems, Man and Cybernetic, Part B: Cybernetics*, 2008, 38: (2) pp. 577-583.
- [10] Khan L, Awad M, Thuraisingham B. A new intrusion detection system using support vector machines and hierarchical clustering. *The VLDB Journal*, 2007, 16 pp. 507-521.
- [11] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines.

- Expert Systems with Applications*, August 2009, 31(2) pp. 231-240.
- [12] Palomo E J, Dominguez E, Luque R M, et al. A new GHSOM model applied to network security. *Lecture Notes in Computer Science Springer*, 2008, 5168 pp. 680-689.
- [13] L. Carettoni, C. Merloni, and S. Zanero, "Studying Bluetooth Malware Propagation: The BlueBag Project," *IEEE Security and Privacy*, vol. 5, no. 2, pp. 17-25, Mar. 2007
- [14] C. Gao and J. Liu, "Modeling and Restraining Mobile Virus Propagation," *IEEE Trans. Mobile Computing, Early access article*, no. 99, 2012.
- [15] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A. -L. Barabasi, "Understanding the Spreading Patterns of Mobile Phone Viruses," *Science*, vol. 324, no. 22, pp. 1071-1076, May 2009.
- [16] J. M. Heffernan, R. J. Smith, and L. M. Wahl, "Perspectives on the Basic Reproductive Ratio," *J. Royal Soc. Interface*, vol. 2, no. 4, pp. 281-293, Sept. 2005.
- [17] A. Bose and K. G. Shin, "On Mobile Viruses Exploiting Messaging and Bluetooth Services," *Proc. Conf. Securecomm and Workshops*, pp. 1-10, 2006
- [18] J. Hollis, Focus on London 2010: Population and Migration, O. f. N. Statistics, 2010.
- [19] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A. -L. Barabasi, "Understanding the Spreading Patterns of Mobile Phone Viruses," *Science*, vol. 324, no. 22, pp. 1071-1076, May 2009.
- [20] J. Su and K. K. W. Chan, "A Preliminary Investigation of Worm Infections in a Bluetooth Environment," *Proc. ACM Workshop Recurring Malcode (WORM)*, 2006.



Huang Wen-hua (1980.5-) female, Master degree, lecture, Her interests are in network security & information security, especially security evaluation.

Online Order Priority Evaluation Based on Hybrid Harmony Search Algorithm of Optimized Support Vector Machines

Zhao Yuanyuan

Department of Management, Guangzhou Vocational College of Technology & Business, Guangzhou Guangdong, China

Chen Qian

College of Science and Engineering, City University of Hong Kong, Hong Kong, China

Abstract—To make production plan, online order priority evaluation is the current priority weakness of online order evaluation model. This thesis proposes an online order priority evaluation model based on hybrid harmony search algorithm of optimized support vector machine (HHS-SVM). Firstly, an online order priority evaluation index system is build, and then support vector machine is adopted to build an online order priority evaluation model; secondly, harmony search algorithm is used to optimize the parameters of support vector machine; what's more, in the parameter optimization process the foraging behavior of AFSA is introduced to improve the ability and convergence speed of algorithm escaping from the local optimal solution; finally, simulation test is used to test the performance of the model. The simulation results show that HHS-SVM improves the accuracy of the online order priority evaluation relative to the comparison model. What's more, the online order priority evaluation model is feasible and effective.

Index Terms—Online Order Priority; Supply Chain; Support Vector Machine; Harmony Search Algorithm

I. INTRODUCTION

Currently, global resources and the environment have become increasingly prominent, the world increasing emphasis on sustainable development and recycling economy, have introduced legislation to improve and extend the consciousness and responsibility of enterprises to reuse waste materials [1]. The closed-loop supply chain mode which is traditional supply chain plus reverse supply chain has become a enterprise and academic research focus [2]. In the production enterprises, often appear simultaneously into multiple online orders to deal with, or online order processing is not completed, add to the mix of new online orders, while the production capacity is limited, in a moment of online orders production tasks exceeds production ability, you can only make a choice in a number of lesser, which needs to know the online order priority. Some companies will process the online order according to the FIFO principle, from the surface, it is fair to all customers, but it will bring a range of troubles to companies and customers.

Treat all customers equally will extend the average online order processing time, not even in time to meet some of the customer service requirements, and thus lose an important long-term customer online orders. When online order backlog, you should take the appropriate method to sort online order processing, rules may involve delivery, amount, cost, profit, quality, etc., derived by some algorithm integrated priority online orders, resulting in ranking between online orders or online order, to help executives make the right decisions. In production capacity and material inventory capabilities limited circumstances, help companies try to arrange the relative importance of production online orders priority to optimize resource utilization, improve production efficiency. Sort, also known as classification, is a set of data or things in certain online order or online order rules. Sorting method based on a quantitative linear weighting method, ABC cost method, fuzzy comprehensive evaluation, neural network algorithm and AHP, etc.

In the current fiercely market competitive circumstances, quick release and implementation of the task online orders, provide customers with accurate, real-time online order information, and enable enterprises quickly respond to customer delivery requirements is more important. Online order processing capabilities are an important part of customer service, and determine of online order production priority is the most important aspect, so how to evaluate single production priority scientifically and rationally, and improve service efficiency become an important topic in the study [3]. Supply chain disruption management was firstly proposed by Clausen, aimed at resolving the airlines to respond to emergencies areas and get a good application [4]. Currently the supply chain for emergency research focuses on positive aspects : the literature [5-10] investigate the demand caused by unexpected events, promotional investment or cost sensitivity coefficient and market size changes were studied by adjusting the original contract to coordinate the supply chain; literature [7] based on literature [8] extended the study object to one supplier and a leading position in the composition of multiple retailer supply chain coordination problems,

propose use amended quantity discount contract to coordinate disruption event; literature [9-11] introduces the time factor, demand and prices will be set to a function of time, revenue sharing contract research coordinated response to emergencies; literature [12] investigated the event when the interference caused by production costs, market size and price-sensitive coefficient simultaneously disturbance, build two supply chain game model, the results show that the period of stability of the contract on production plan has some robustness, but when the disturbance exceeds a certain limit, you need to adjust the production plan and design a new contract to coordinate the supply chain; literature [13-16] study the market demand when unexpected events lead to changes in the distribution, the original revenue sharing contract is no longer coordinate the supply chain, but through the amendment of the contract to coordinate the interests among the members.

For production online orders priority evaluation issues, many scholars have done a lot of research, proposed online orders priority evaluation model [17-19] based on the theory of constraints, linear programming theory, strategic theory, AHP and entropy weight method and so on. These models are considered from different perspectives priority online orders, their advantages and disadvantages are present, such as the analytic hierarchy process is simple, easy to implement, but the results of the evaluation is subjective; relationship between linear programming assuming online order priority and influencing factors is a linear relationship, but in fact corporate online orders priority evaluation issues affected by many factors, is a nonlinear problem, so the scope of application of these models is limited. In recent years, Tang Lichun [20] and other people put forward an online order production priority evaluation nonlinear model based on the RBF neural network, neural network has a strong self-organizing and nonlinear approximation capability, increased online order evaluation accuracy of production priority evaluation and make the results more scientific [21]. However, neural network is based on empirical risk minimization principle and the "large sample" the theory of machine learning algorithms, however, online order history data is a small sample size problem, when you can not meet the "large sample" requirement, so neural networks prone to "over-fitting" phenomenon, while there is difficult to overcome their own shortcomings, such as the complexity of network structure [22], etc. Support vector machine (SVM) better overcome the defects of the neural network over-fitting, generalization ability is superior, and provides a new research idea for solving problem of online order production priority. SVM in practical applications, its performance is closely related to parameter selection, the current major genetic algorithm, particle swarm algorithm to optimize the SVM parameters, these algorithms have advantages and disadvantages, such as genetic algorithms for parameter optimization, demand is set different genetic operators, complex operation; particle swarm algorithm, while having a faster convergence speed, but also easy to fall into local minima [23]. Harmony Search

(HS) algorithm is developed in recent years as a heuristic global search algorithm, which has better optimization precision and the ability to escape from local optima, has been used in engineering practice [24].

To solve the problem like the company's own production capacity, profitability and customer relations and other issues, faced with numerous demands orders enterprises should identify critical sequence of completing the orders, and meet each customer's need within their abilities. Overall, the previous studies on this issue are focused on decision analysis methods and analytic hierarchy process method. These methods require policy makers having in-depth understanding of index system of preferred order, the general terms of business operations, market conditions and customers. For the various indicators in order selection system, policy makers need to be given different weights to obtain a decision matrix, and then the final calculation is processed to get the order of priorities; after that, program development and scheduling are performed. But such a decision-making process requires not only tedious calculations, but need a lot of subjective factors. Not using the existing success stories tends to make biased results of the decisions.

The innovation of this paper is as follows:

(a) In online order to improve the accuracy of the online order priority evaluation, present a hybrid harmony search (HHS) algorithm optimizing SVM parameters of online order priority evaluation model (HHS-SVM), while online order backlog, need to do sort processing for online orders, this paper adopt online order evaluation model to scheduling online order processing sequence, the other nodes companies in the supply chain sharing information are used to construct the index system process.

(b) In the supply chain environment, supply chain node enterprises to establish coordination mechanisms between upstream and downstream, and thus the implementation of information sharing. In this environment, the processing of customer online orders sequentially arranged, not only consider their own master data, the choice of delivery, profit limits, the importance of the customer, the amount of online orders, production costs, quality requirements, online order material satisfaction factors such as the agile enterprise production management system evaluation online order priority targets, but also from other nodes to extract business information provided valuable part, take advantage of the value of information sharing, and using genetic algorithms to solve programming by analyzing the generated image and calculation results to determine the solution obtained by this method converges, and through simulation testing to test the effectiveness and superiority of HHS-SVM.

II. ONLINE ORDER PRIORITY EVALUATION MODEL FRAME FOR IHS-SVM

A. Online order Priority Evaluation Model Frame

Online orders priority evaluation is a dynamic, non-linear decision-making process, many factors affect

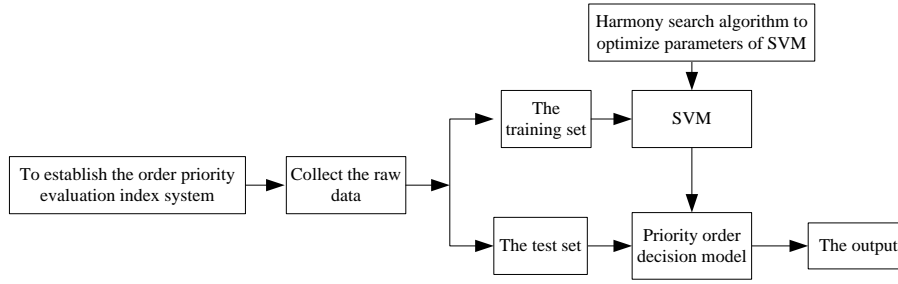


Figure 1. Online order priority model framework of IHS-SVM

the online order priority evaluation, each factor affect the results of the online order priority evaluation with varying degrees, the online order priority evaluation between input and output is a kind of complex nonlinear relationship, and it is difficult to establish a reasonable and precise mathematical expression. Set online orders priority evaluation indicators as $\{x_1, x_2, \dots, x_m\}$, then the mathematical model of online order priority evaluation is

$$y = f(x_1, x_2, \dots, x_m) \quad (1)$$

In formula, $f()$ represents an evaluation function.

Establish online order priority evaluation model, in essence, is to find a best adapted $f()$, you can accurately portray online order priority evaluation system whose relationship between input and output is complex nonlinear relationship. Online order priority evaluation model based on HHS-SVM framework is shown in Figure 1. First, establish the correct online order priority index system, and then use HHS algorithm to optimize SVM parameters, get the optimal SVM parameters, and finally create an online order priority evaluation model, and evaluate the performance of the model.

B. The Centralized Decision-making Model under Disruption Event

In the manufacturer leading of centralized decision situations, in online order to facilitate the discussion, consider the closed-loop supply chain manufacturers optimal countermeasures. The goal centralization decision at this time is to maximize the benefit of closed-loop supply chain. Therefore, interference with the incident, the closed-loop supply chain profit function is:

$$\begin{aligned} \pi_{scd}(Q) = & ((D + \Delta D - Q) / k_1 - c_0 + (\eta + \Delta \eta) \\ & (\Delta_0 - \Delta c_m - c_r))Q - C(\eta + \Delta \eta)^2 \\ & - \lambda_1(Q - \bar{Q})^+ - \lambda_2(\bar{Q} - Q)^+ \end{aligned} \quad (2)$$

Here $(x)^+ = \max\{x, 0\}$. Similar to the proof of literature [16], lemma 1 can be obtained.

Lemma 1 Suppose under condition of centralized decision-making, the optimal solution of formula(2) is Q^* , when market scale D , remanufacturing cost c_m , and recovery rate η are changed and fulfill $\Delta D > k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r)$, there is $Q^* \geq \bar{Q}$; when $\Delta D < k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r)$, there is $Q^* \leq \bar{Q}$.

From Lemma 1, when the market demand size increases which caused by the interference event, and the remanufacturing cost and recoveries rate and the amount of change to meet the above conditions, the manufacturer want to maximize profits of the closed-loop supply chain, should increase production to meet market demand, and vice versa, its regularity realistic scenarios.

Learn from lemma 1, the following situations can be considered.

(1) When $\Delta D < k_1(\eta\Delta c_m - \Delta\eta(\Delta_0 - \Delta c_m - c_r))$, get $Q^* \leq \bar{Q}$ from lemma 1, so $\pi_{scd}(Q)$ can get the best explain in region $[0, \bar{Q}]$. Now there is:

$$\begin{aligned} \pi_{scd}(Q) = & ((D + \Delta D - Q) / k_1 - c_0 + (\eta + \Delta \eta) \\ & (\Delta_0 - \Delta c_m - c_r))Q - C(\eta + \Delta \eta)^2 - \lambda_2(\bar{Q} - Q) \end{aligned} \quad (3)$$

According to the first class optimality condition $\partial\pi_{scd}(Q) / \partial Q = 0$, can know that the optimal value of Q is:

$$Q^{**} = (D - c_0k_1 + k_1(\eta + \Delta\eta)(\Delta_0 - \Delta c_m - c_r) + k_1\lambda_2 + \Delta D) / 2 \quad (4)$$

1) When $\Delta D < k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) - k_1\lambda_2$, from $k_1, \lambda_2 > 0$, ΔD obviously fulfill $\Delta D < k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r)$, now $Q^{**} \leq \bar{Q}$, as $\pi_{scd}(Q)$ is strict concave function, get the only best explain during region $[0, \bar{Q}]$ for $\pi_{scd}(Q)$ is Q^* , which means:

$$\begin{aligned} Q^* = & Q^{**} = \bar{Q} + (k_1\eta\Delta c_m + k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) \\ & + k_1\lambda_2 + \Delta D) / 2 \end{aligned}$$

Opposite optimal retail price is:

$$p^* = \bar{p} + (\Delta D + k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) - k_1\lambda_2) / 2k_1.$$

2) When $\Delta D \geq k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) - k_1\lambda_2$, get $Q^{**} \geq \bar{Q}$. Besides, $\pi_{scd}(Q)$ in region $[0, \bar{Q}]$ is strict concave function about Q , so $\pi_{scd}(Q)$ get the maximum value at the end point, compare with $\pi_{scd}(0) < \pi_{scd}(\bar{Q})$, so the optimal result is $Q^* = \bar{Q}$. At this time, the best price for closed-loop supply chain is:

$$p^* = \bar{p} + (\Delta D - k_1\eta c_r) / k_1.$$

(2) When $\Delta D > k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r)$, get $Q^* \geq \bar{Q}$ from lemma 1, so $\pi_{scd}(Q)$ get the maximum value in $[\bar{Q}, +\infty)$, the current optimal benefit is:

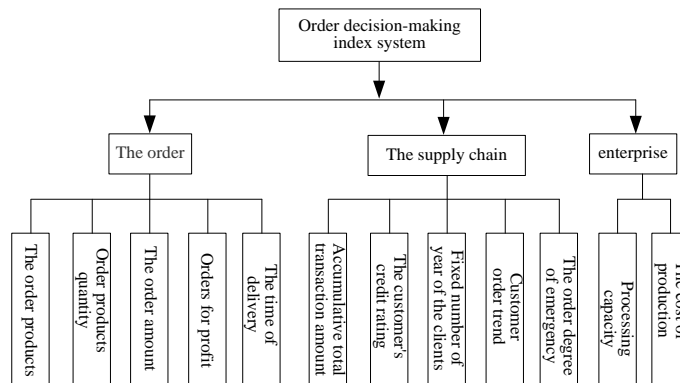


Figure 2. Online orders priority evaluation index system

$$\pi_{scd}(Q) = ((D + (\Delta D - Q) / k_1 - c_0 + (\eta + \Delta \eta) (\Delta_0 - \Delta c_m - c_r))Q - C(\eta + \Delta \eta)^2 - \lambda_1(Q - \bar{Q}))$$

Similar to the discussion on (1), get the following conclusions:

1) When $k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) < \Delta D < k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) + k_1\lambda_1$, the optimal output for closed-loop supply chain is $Q^* = \bar{Q}$, optimal retail price is:

$$p^* = \bar{p} + (\Delta D - k_1\eta c_r) / k_1$$

2) When $\Delta D \geq k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) + k_1\lambda_1$, optimal output and optimal retail price of supply chain is:

$$Q^* = \bar{Q} + (k_1\eta\Delta c_m + k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) - k_1\lambda_1 + \Delta D) / 2$$

$$p^* = \bar{p} + (\Delta D + k_1\eta\Delta c_m - k_1\Delta\eta(\Delta_0 - \Delta c_m - c_r) + k_1\lambda_1) / 2k_1$$

In summary, we can get the following Theorem 1.

C. Online orders Priority Evaluation Index System

The first step to establish online orders priority evaluation model is to build a whole set of index system, and index system is scientific and reasonable or not is directly related to the scientific and practical of evaluation model, but the online order priority is affected by many factors, such as production capacity, online order profits, customer online order trends. This paper establishes the online order priority index system which shown in Figure 2 through system analysis and expert commentary, and with reference to relevant literature and research.

III. HHS OPTIMIZE SVM ONLINE ORDER PRIORITY EVALUATION MODEL

HS algorithm simulates that musicians do musical creation with their own memories, by repeatedly adjust the pitch of each instrument in the orchestra, and ultimately achieve a wonderful harmonies state process [12]. In HS, the harmony memory storage feasible solution vector, harmony memory size determines the number of feasible solutions, harmony memory retention rate is selected from the newly generated solution probability, and tone adjustment probabilities are generated new solutions for the probability of disturbance. In online order to find better SVM parameters, in the HS algorithm, combines AFSA foraging behavior, and enhance the ability to escape from local optima and

convergence speed, resulting in a hybrid harmony search algorithm (HHS). Steps for online orders priority evaluation of HHS-SVM are:

Step 1: according to expert systems and related research and production enterprise real-world conditions, establish online order priority evaluation index system.

Step 2: According to the corresponding data which index system collected, according to experts on the online order priority evaluation index system for comprehensive analysis of the expected output, and gives the corresponding quantized value, larger value indicates a higher priority for the online order, $y \in (0,1)$, thereby construct an online order priority evaluation sample of HS-SVM modeling.

Step3: divide the sample into designated set and training set, the training set for the HS-SVM learning, the establishment of online order priority evaluation model, the test set is used to verify the priority online order established evaluation model performance.

Step 4: normalize index, dimensionless index difference eliminate the adverse effects on the online order priority evaluation samples normalization:

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

In this formula, x'_i means the data after normalization, and x_{\max} , x_{\min} separately means maximum value and minimum value.

Step 5: Set parameter range of SVM. According to the related literature, the range of the parameter C determined as [1, 1000], the parameter σ range is set to [0.1, 100], and set the related parameters for HS algorithm: harmony memory size (HMS), harmony memory considering rate (HMCR), pitch adjustment probability (PAR), maximum number of iterations NI.

Step 6: Harmony memory initialization. A number of HMS initial solutions randomly generated and stored in the harmony memory (HM). HM is treated as a matrix:

$$HM = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^{HMS} \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 & f(X^1) \\ x_1^2 & x_2^2 & \cdots & x_n^2 & f(X^2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_1^{HMS} & x_2^{HMS} & \cdots & x_n^{HMS} & f(X^{HMS}) \end{bmatrix} \tag{6}$$

In this formula, x_j^i means $i(i=1,2,\dots,HMS)$ of component No $j(j=1,2,\dots,N)$ for harmony vector, $x_{jL} \leq x_j^i \leq x_{jU}$, in this formula x_{jL} and x_{jU} separately stands for the under and up boundary of j , $f(X^m)$ is target function value.

This paper takes the RMSE between output value y_i of HS-SVM and expect output y_i as target function value, and there is:

$$f(X^m) = RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

In this formula, n represents the number of samples in the training set.

Step7: Based on HMCR, PAR, take extemporaneous creation on vector $x'=[x'_1, x'_2, \dots, x'_N]$, generate a new harmony vector. If harmony memory consideration rate HMCR is fulfilled, and primary adjusting rate PAR can not be fulfilled, then execute harmony memory consideration. Refer to harmony memory consideration, it means select one x_j^i from harmony memory bank $\{x_j^1, x_j^2, \dots, x_j^{HMS}\}$, $j=1,2,\dots,N$, constitute vector x' ; If harmony memory consideration rate HMCR is fulfilled, and primary adjusting rate PAR is also fulfilled, then adjust the primary.

HHS algorithm and basic harmony algorithm are different in the following two aspects:

(1) In harmony memory base, in the current optimal solution implemented on the basis of pitch adjustment operation, so you can give full play harmonies memory guiding role in the optimal solution, and improve the convergence speed.

(2) In the pitch adjustment operation, introduce AFSA foraging thought. Treat the optimal harmony vector x^{best} in the harmony memory base as the current state of artificial fish, and calculate the objective function; when optimization, random remove x^{best} a candidate solution which has been selected, and randomly select a solution has been selected as substitute candidate solutions, obtain harmony vector x^j . If the objective function value of harmony vector x^j is bigger than the objective function value of harmony vector x^{best} , then let the resulting new harmony vector $x' = x^j$; contrary, again randomly select a solution has been selected as a substitute candidate solutions, build status x^j , to determine whether meet the requirements, repeated a few times, if you still can not meet the requirements, then let the last harmony vector x^j which based on foraging behavior to be the newly generate harmony vector x' , the introduction of fish feeding ideas is good to do fine search in optimum harmony vector field.

Step 8: According to equation (11) which generate a new individual X' , and calculating a new individual objective function value, if the individual memory than

the worst harmony vector x^{worst} , then replace x^{worst} with the replacement X' .

Step 9: iteration $k = k + 1$, if k is bigger than the maximum number of iterations, then select the harmony vector of optimal objective function value in harmony memory bank, it means to find the optimal SVM parameters (C, σ) , otherwise go to step (7) to continue optimization.

Step 10: training set input SVM, according to the most optimal parameters (C, σ) , to establish the optimal online order priority evaluation model.

Step 11: Enter the test set priority online order to establish the optimal evaluation model, get the online order priority value.

IV. EXPERIMENTAL RESULTS

A. Data Sources

To test the performance of online order priority evaluation model of HHS-SVM. Set a Shanghai company whose main business is apparel products as an example, online order priority evaluation indicators include online order quantity (x_1), online order amount (x_2), online order profit (x_3), online order delivery time (x_4), online order accumulated transaction amount (x_5), customer credit rating (x_6), customer online order trend (x_7), customer cooperation time (x_8), online order urgency degree(x_9), and production cost(x_{10}), total 10 indicators (x_{10}), collected the company's 100 online orders, of which the first 70 online orders data as the training set, after 30 as the test set. Specific data is shown in Table 1. In the environment of PIV 3.0GHZ CPU, 2G RAM, and Windows XP, do simulation through MATLAB 2009a toolbox.

TABLE I. NORMALIZED ONLINE ORDER HISTORY DATA

Number	x_1	x_2	x_3	x_4	...	x_{10}	y
1	0.202	0.202	0.027	0.491	...	0.058	0.354
2	0.405	0.405	0.003	0.895	...	0.033	0.191
3	0.180	0.180	0.117	0.216	...	0.184	0.404
4	0.311	0.311	0.009	0.906	...	0.058	0.574
5	0.259	0.259	0.042	0.513	...	0.052	0.037
6	0.476	0.476	0.008	0.478	...	0.011	0.393
7	0.134	0.134	0.014	0.592	...	0.004	0.976
8	0.129	0.129	0.008	0.888	...	0.018	0.043
9	0.028	0.028	0.285	0.471	...	0.049	0.071
...
100	0.886	0.886	0.653	0.337	0.234	0.453	0.359

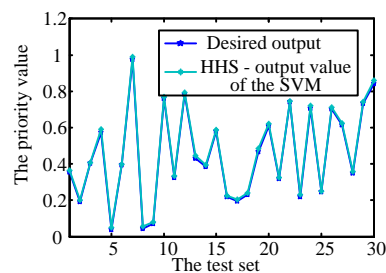


Figure 3. HHS-SVM online order priority evaluation results

B. Model Implementation

Put training set into SVM for learning, and use HHS algorithm to optimize SVM parameters, get the optimal

parameters $C = 175.15$, $\sigma = 19.228$, finally use $C = 175.15$, $\sigma = 19.228$ to establish the optimal online order priority evaluation model, and test testing set, the results shown in Figure 3. From Figure 3 shows, HHS-SVM actual output is very close to the model output, precision of evaluation result is very high, it shows that, HHS-SVM is an evaluation of high precision, reliable results online orders priority evaluation model.

C. Compare with Other Algorithms Optimized SVM Model Performance

To allow HHS-SVM results become more convincing, select genetic algorithm optimization SVM (GA-SVM), particle swarm optimization SVM (PSO-SVM) and basic harmony search algorithm (HS-SVM) to do comparing experiments, using both rms error (RMSE) and the mean relative error (MAPE) as a measure of performance of the model, and their values are shown in Table 2. The correlation curve of actual output and the model output for GA-SVM, PSO-SVM and HS-SVM are shown in Figure 4-6.

TABLE II. COMPARISON OF EVALUATION ERRORS BETWEEN VARIOUS MODELS

Evaluation Model	RMSE	MAPE (%)
GA-SVM	0.064	10.41
PSO-SVM	0.057	9.75
HS-SVM	0.023	3.41
HHS-SVM	0.019	3.08

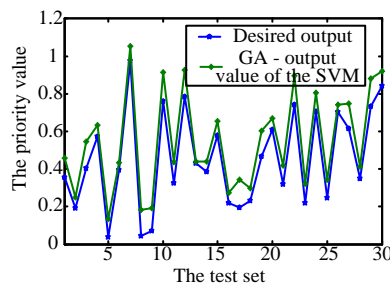


Figure 4. Online order priority evaluation results for GA-SVM

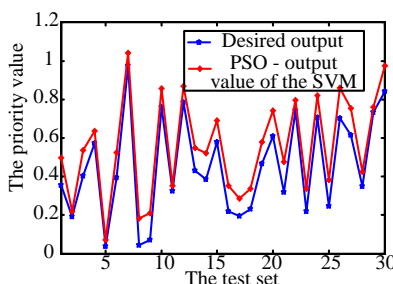


Figure 5. Order priority evaluation results for PSO-SVM

After comparing and analyzing the simulation results on table 2 and Figure 4-6, overall performance of HHS-SVM is better than comparison model, and can get the following conclusions:

(1) Evaluation accuracy of HS-SVM is slightly better than GA-SVM, PSO, which indicates HS algorithm has better global search ability than GA, PSO, get better SVM parameters C , σ , can reduce the error rate of online

order optimization evaluation, effectively improve the evaluation accuracy of online order priority evaluation.

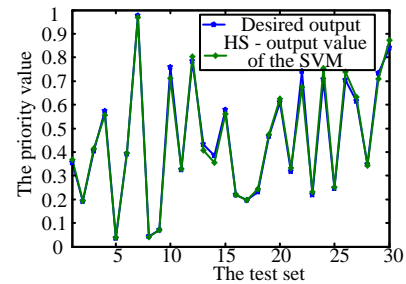


Figure 6. Order priority evaluation results for HS-SVM

(3) The evaluation accuracy of HHS-SVM is the highest, evaluation error is much smaller than the comparison model HS-SVM, which indicates HHS algorithm introduces the foraging behavior of artificial fish swarm algorithm based on HS algorithm, it can further improve the capacity that algorithm jumps out of local optima, thus improving the quality of the optimization algorithm. Apply it to SVM parameter optimization, you can get a better SVM parameters, and work with small samples, nonlinear approximation ability of SVM, and the obtained result of the online order priority evaluation is completely the same as the evaluation result from expert. The comparative result shows that HHS-SVM model can more accurately describe the complex non-linear relationship between online order priority value and its affect factors, so the evaluation result for online order priority value is more accurate.

V. CONCLUSIONS

To the problem of online order priority evaluation in Supply Chain, propose a nonlinear online order priority evaluation model based on HHS-SVM. The model introduces the foraging behavior of artificial fish swarm algorithm based on HS algorithm to improve the algorithm optimization ability, and combined with the results of SVM applying to online orders priority evaluation problem to get the answer. The result shows that, the evaluation accuracy of HHS-SVM is obviously better than other online order priority evaluation method, well ensured the objectivity of online order priority evaluation results, and the evaluation results will help enterprises to enhance their competitiveness.

ACKNOWLEDGMENT

This work is supported by 2013 guangdong province higher vocational education teaching reform project from management class teaching steering committee (YGL2013096)

REFERENCES

[1] Wu Chong, David Barnes, Duska Rosenberg, et al. An analytic network process-mixed integer multi-objective programming model for partner selection in agile supply chains. *Production Planning & Control*, 2009, 20 (3) pp. 254-275.

- [2] Ren J, Yusuf Y Y, Burns N D. A decision-support framework for agile enterprise partnering. *International Journal of Advanced Manufacturing Technology*, 2009, 41 (1-2) pp. 180-192.
- [3] Mahdavi IM. Global-best harmony search. *Applied Mathematics and Computation*, 2008, 198(2) pp. 643-656.
- [4] Bojarski, A. et al. Life cycle assessment coupled with process simulation under uncertainty for reduced environmental impact: application to phosphoric acid production. *Industrial & Engineering Chemistry Research*, 05 November 2008, vol. 47
- [5] Cholette S., K. Venkat. The Energy and Carbon Intensity of Wine Distribution: A Study of Logistical Options for Delivering Wine to Consumers. *Journal of Cleaner Production*, (2009)17(16), 1401-1413.
- [6] B. Sundarakani, R. Souza, M. ZGoh etc. Modeling carbon footprints across the supply chain. *Int. Production Economics*, 2010(128) pp. 43-50.
- [7] Nhan N, Insoo K. An enhanced cooperative spectrum sensing scheme based on evidence theory and reliability source evaluation in cognitive radio context. *IEEE Communications Letters*, 2009, 13(7) pp. 492-494, 760.
- [8] YANG S, NAGAMACH M, LEE S. Rule-based inference model for the Kansei engineer systems. *International Journal of Industrial Ergonomics*, 1999, 24(5) pp. 459-471.
- [9] XU R, WUNSCH DC. Clustering. New York: Wiley-IEEE Press, 2008.
- [10] KYUNGME C, CHANGRIM J. A systematic approach to the Kansei factors of tactile sense regarding the surface roughness. *Applied Ergonomics*, 2007, 38(1) pp. 53-63.
- [11] GUHA S, MUNAGALA K. Exceeding expectations and clustering uncertain data: Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, *Rhode Island, 2009. USA: ACM*, 2009, pp. 269-278.
- [12] GRIMSAETH K. Kansei Engineering: Linking Emotions and Product Features. Norwegian: *Norwegian University of Science and Technology*, 2005.
- [13] Yuexiang HUANG, Chun-Hsien CHEN, Li Pheng Khoo. Kansei clustering for emotional design using a combined design structure matrix. *International Journal of Industrial Ergonomics*, 2012, 42(5) pp. 416-427.
- [14] STEWARD D V. Systems Analysis and Management: Structure, Strategy and Design. *New York: Petrocelli*, 1981.
- [15] EPPINGER S D. Mode-I based Approaches to Man-aging Concurrent Engineering. *Journal of Engineering Design*, 1991, 2(4) pp. 283-290.
- [16] SMITH R P, EPPINGER S D. A Predictive Model of Sequential Iteration in Engineering Design. *Management Science*, 1997, 43(8) pp. 1104-1120.

Framework and Modeling Method for Heterogeneous Systems Information Integration Base on Semantic Gateway

Xianwang Li*, Yuchuan Song*, Ping Yan, and Xuehai Chen

State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing, China

*Corresponding author, Email: lxw023@163.com, syc@cqu.edu.cn, yp@cqu.edu.cn, cxh@cqu.edu.cn

Abstract—To realize the integration of heterogeneous systems with greater flexibility and more convenience, a framework for heterogeneous systems information integration based on semantic gateway is constructed firstly, and then the integration process and several relevant technologies are discussed. On the basis of above researches, a modeling method for heterogeneous systems information integration is put forward. According to the integration requirements and traditional information models, the method is able to rapidly generate the information integration models, which provide supports to the implementation and operation of semantic gateway. Finally the research results have been verified in a manufacturing enterprise.

Index Terms—Modeling Method; Information Integration; Semantic Gateway

I. INTRODUCTION

The application of information systems in manufacturing enterprises have greatly improved the work efficiency and the market competitiveness. However, these systems are mostly dispersed and heterogeneous, and difficult to achieve information sharing and integrating between each other, which bring great difficulties to the collaboration between different departments within the enterprise, as well as the cooperation between enterprises. Therefore, there exist urgent requirements for the integration of heterogeneous systems. On the other hand, to adapt to the environment change more agilely, the information systems of manufacturing enterprises are required to have the ability of rapidly reconfiguration and integration on demand.

With the development of interoperation technologies, such as the appearance of COM, Web Service, MOM, the middleware-based integration technology was put forward and became dominant gradually. In this case, if the amount of systems involved is n , it just needs n times operations because the integration of these systems is conducted not between each other but with the middleware separately. In this way, the integrating is greatly simplified. At the same time, the middleware has the ability to transform data format and convert protocols. As a result, the problem of Heterogeneity is solved and the flexibility is improved.

The middleware-based integration is the research focus. Many researchers pay great attention to this area and lots of solutions were proposed separately. For instance, ref [3] discussed the data integration in banking statement management system. Ref [4] proposed a collaborative and integrated platform to support distributed manufacturing system. Ref [5] introduced a kind of heterogeneous data integration middleware.

The technologies mentioned above have solved the problems of interoperation and greatly improved the flexibility. But most of them lack effective practice. On the other hand, the research mainly concerned technology, ignoring information content.

(3) Integration in the information content layer

In this aspect, how to describe, analyze and map the relevant information is discussed. The research mainly consists of the following three aspects: a) information format; b) information semantics; c) information modeling.

The purpose of information format research is to make Information exchange standards, in order to make the information be identified and analyzed by different systems. STEP, EDI and XML are typical ones [6-8]. These technologies unify message interchange format, that is, has clarified fields constructing information and the data format of these fields. But it does not explicit the semantic information of fields. Therefore, it still needs to write special programs to analyze the interchange message. Because of this, the study of semantic information shows up. It indicates that semantic expression of information is cleared and standard by semantic annotation, ontology modeling and ontology-mapping technology [9-11], so that information can be recognized. There are not yet widely accepted standards now, so that semantic integration lacks effective applications.

Expression of information format and information semantics, as well as the construction of the systems, needs the support of certain models. So information modeling research is very important. Existing information modeling research mainly two aspects: a) traditional information modeling; b) Semantic Information Modeling. Traditional information modeling research mainly focuses on how to build information system

models and support design and development of information systems with these models.

UML and IDEF are typical ones [12, 13]. Semantic information modeling research mainly focus on how to build a semantic model of information, usually using ontology modeling technology, formal modeling technology and so on [14, 15].

Currently, traditional information modeling and semantic information modeling are studied separately in most cases, and there are few researches based on the combination of them. Furthermore, because the systems are built separately, information models are constructed respectively in the case of traditional information modeling, without considering information systems integration, namely these models do not contain any integration elements.

In summary, the middleware-based integration technology and semantic information integration technology are currently research hotspots, and they can solve the problems of information integration of heterogeneous systems. But due to the high technical requirements, there is a lack of effective implementation.

The researches on information integration generally consist of the following three aspects:

(1) Integration in the database layer

The information of enterprise information systems is mainly stored in databases, so information integration can be regarded as the integration of databases.

The Integration technologies of databases mainly include data federation and data warehouse [1]. In the case of data warehouse, data warehouse is created and used for storing the data, which are extracted from one or more original databases, and processed according to the global schema. In the case of data federation, a virtual view is used to collect, but not really store, integration information from original databases.

The main disadvantage of data federation and data warehouse is that they are tightly coupled to the original databases. As a result, the flexibility is insufficient. In addition, they put more emphasis on data integration, regardless of interoperation among databases.

(2) Integration in the application layer

In this aspect, interoperation among systems is mainly focused on.

In the early time, interoperation was realized through p2p-based secondary development according to the integration requirements. E.g. Ref [2] applied the interoperation technology to the integration between PDM and ERP. The advantage is that this technology is easy to realize, so it has been widely used. However, because the p2p-based secondary development is necessary, the integration efforts increase rapidly as the systems involved increase. For example, if the amount of systems involved is n , secondary development must be conducted c_n^2 times. The flexibility and extendibility is poor.

Therefore, the author's research team proposes a new integration technology, namely semantic gateway [16, 17] technology on the basis of the above study. It is intended

to provide an effective solution to heterogeneous system integration with more convenience and greater flexibility.

II. HETEROGENEOUS SYSTEMS INFORMATION INTEGRATION BASED ON SEMANTIC GATEWAY

A. Framework for Heterogeneous Systems Information Integration Based on Semantic Gateway

The heterogeneous systems information integration framework based on semantic gateway is shown in figure 1. The integration framework consists of information systems, databases, database triggers, database adapters, semantic gateway server, and semantic gateway modeling tools. The database triggers designed as special database triggers for semantic gateway construction, are used to capture data changes; the database adapters are employed for adaption to different types of database system, such as Oracle, SQL Server, MySQL, etc.; the semantic gateway server, with the semantic gateway engine as the key part, is the core of the whole framework; the semantic gateway engine is constructed on basis of message-oriented middleware. Different from conventional one, this message-oriented middleware can conduct two-level semantic parsing and mapping, thus equipped with the ability of semantic understanding; the semantic gateway modeling tools are used to build integrated models, including semantic description model, semantic mapping model, semantic publishing modes and semantic subscription model, etc.

B. The Integration Process and Relevant Technologies

As enterprise information systems, in most cases, are based on databases, the information will be eventually stored in the databases. And various operations in information systems will finally mapping for four types of operations (insert, delete, update, select), and data changes in database follow. Therefore, the integration of information systems can be regarded as data synchronization among databases. From this perspective, the semantic gateway can also be seen as a data synchronization tool of databases. At the same time, different information systems and corresponding databases are mostly heterogeneous. Therefore, semantic gateway is also a tool to eliminate heterogeneity and help heterogeneous systems understand each other.

(1) Publish/Subscribe Process

The publish/subscribe techniques are able to realize decoupling between sender and acceptor. In this way, the sender is only responsible for sending messages to the semantic gateway, without caring about who will receive the message.

The detailed process of publish/subscribe is as follows.

Firstly, according to the integration requirements, the sender establishes database triggers, and database triggers will capture data alteration and send them through database adaptor to semantic gateway for further processing. Then, the acceptor as integration demands initiates subscription to semantic gateway. According to subscription items, semantic gateway engine distributes messages as required, and eventually write into the database of acceptor through database adaptor.

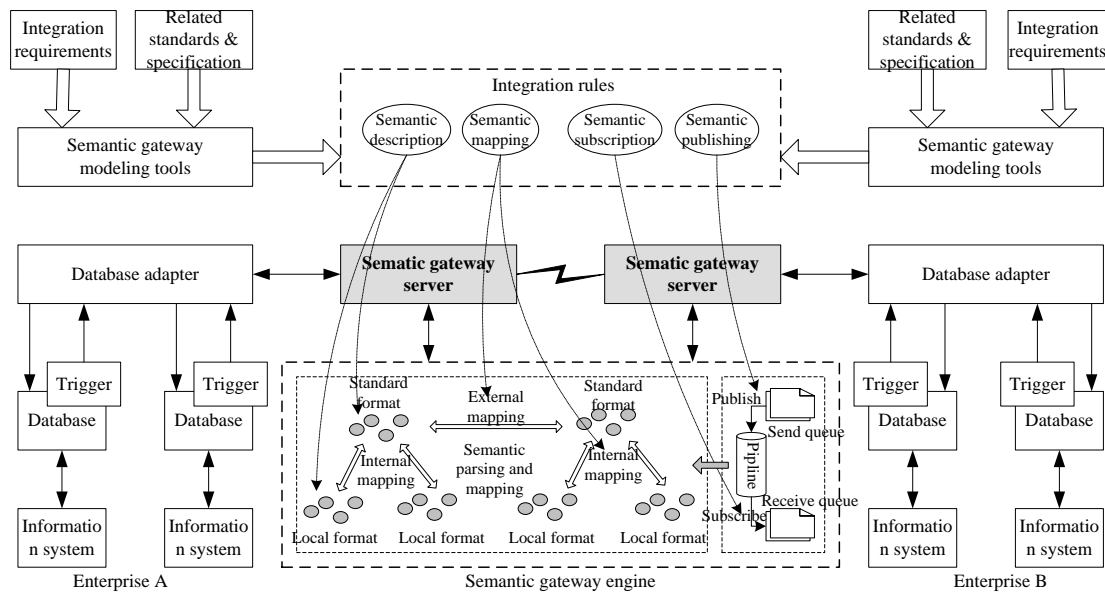


Figure 1. Framework for heterogeneous systems information integration based on semantic gateway

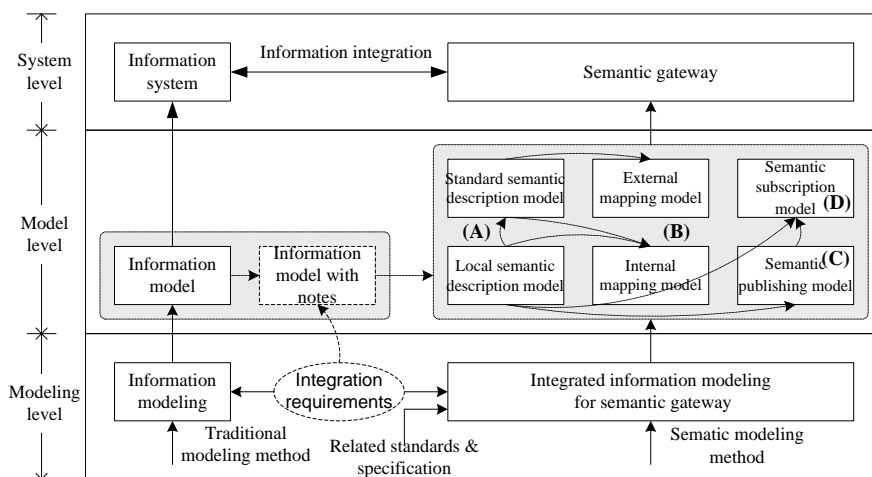


Figure 2. The modeling framework

(2) Two-Level Semantic Parsing and Mapping Process

Firstly, information in each information system has its own internal format called local semantic description. Secondly, every enterprise may set up a set of consolidated message interchange format called standard semantic description. And this is two-level semantic description. The mapping between local semantic description and standard semantic description is called internal mapping, and the mapping among multiple standards is called external mapping.

Internal information integration in enterprise should go through internal mapping twice. Firstly, the local semantic of system A maps the standard semantic of enterprise, and then standard semantic of enterprise maps the local semantic of system B.

Information integration among enterprises should go through internal mapping twice and external mapping once. Firstly, the local semantic of system A maps the standard semantic of enterprise A. Secondly, the standard semantic of enterprise A maps the standard semantic of

enterprise B. Finally, the standard semantic of enterprise B maps the local semantic of system B.

III. MODELING METHOD FOR HETEROGENEOUS SYSTEMS INFORMATION INTEGRATION BASE ON SEMANTIC GATEWAY

A. Modeling Framework for Heterogeneous Systems Information Integration Based on Semantic Gateway

The framework of the heterogeneous system integration modeling method based on semantic gateway is shown in figure 2.

The main content of this framework can be summarized in three levels and two main lines. Three levels refer to modeling method layer, model layer and system layer. The model is produced by modeling method, and the model serves for system.

Two main lines refer to information modeling line and semantic gateway information integration modeling line. Information modeling line means the conventional information system modeling process, using conventional

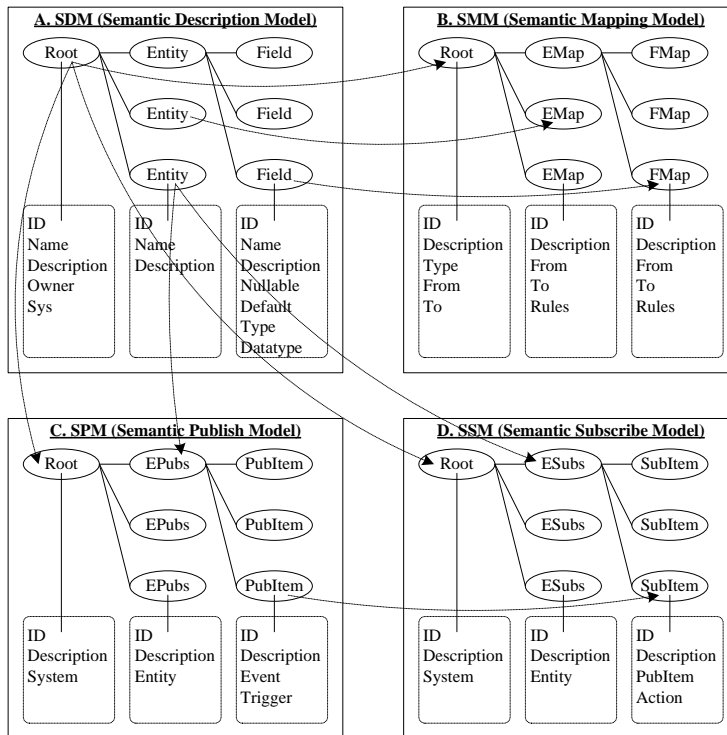


Figure 3. The information integration models

modeling method, such as UML, IDEF1x, etc., to construct information models of information systems. It will serves for the construction and development of information systems. Semantic gateway information integration modeling line is used to express the process of the heterogeneous systems integrated modeling method based on semantic gateway, using semantic modeling method and leading in relevant standards, to construct semantic gateway information integration models. It eventually supports the configuration and operation of semantic gateway. There is an obvious difference between semantic modeling method and conventional modeling method. The generated models of semantic modeling method should be analyzed by semantic gateway, so every model element and symbol needs specific semantic. Semantic gateway information integration model includes local semantic description model, standard semantic description model, internal semantic mapping model, external semantic mapping model, semantic publishing model and semantic subscription model.

Integration requirements are the link between information modeling and semantic gateway information integration modeling. Conventional information modeling has no regard for integration requirement. In order to support semantic gateway integration, the conventional information modeling method should be extended to add integration requirements description. Just adding the mark “☑” and corresponding annotations to model elements required to participate in integration, the prototype of the semantic gateway information integration models will appear without interruptions to the normal information modeling process.

B. The Models and Modeling Process

Semantic gateway information integration model includes local semantic description model, standard semantic description model, the internal semantic mapping model, the external semantic mapping model, semantic publishing model and semantic subscription model. And these six models can be divided into semantic description model, semantic mapping model, semantic distribution model and semantic subscription model. As shown in Figure 3.

(1) Semantic Description Model (SMM)

Semantic Description Model is defined as follows.

$$SMM = \{ID, Name, Description, Type, Owner, (Entity, Entity, Entity...)\} \tag{1}$$

ID is the unique number, *Name* is the designation, and *Description* is the additional information of the model. *Type* is the type of the model, valued as {Local|Standard}. *Local* stands for local semantic description model and *Standard* stands for standard semantic description model. *Owner* is used to represent which system or enterprise does the model belongs to.

Entity is similar to the table model of database. Each *SMM* model contains more than one *Entity* model.

$$Entity = \{ID, Name, Description, (Field, Field, Field...)\} \tag{2}$$

ID is the unique number, *Name* is the designation and *Description* is the additional information of *Entity* model.

TABLE I. FIELD TYPE

Type Name	Sample	Additional Semantic Description
Primary	Customer.ID	{Rules}, Rules represents the encoding rules of primary key
Foreign	Order.Customer.ID	{Entity, Field}, Entity represents the associated entity, Field represents the primary key of the associated entity
Calc	Total-price = Unit-price * Amount	{Fields, Formula}, Fields represent one or more fields which are reliant, Formula represents computational formula.
Enum	Sex	{(Item, Item, Item ...) Item= {Value, Description}}, Value is the value of enumeration term, Description is the semantic description of enumeration term.
Normal	Customer.Name	

TABLE II. DATA TYPE OF FIELD

Type Name	Sample	Additional Semantic Description
Decimal	16.7	{Length, Precision}, Length represents data bits, Precision represents decimal digits.
Currency	\$99.0	{Type}, Type represents kind of currencies.
Date	8:23:15	{Local, Format}, Local represents time zone, Format represents date format.
String	"hello"	

Field is similar to the field model of database. Each Entity contains more than one Field model.

$$Field = \{ID, Name, Description, Nullable, Default, Type, Datatype\} \quad (3)$$

ID is the unique number, Name is the designation and Description is the additional information of the model. Nullable represents whether the field value can be empty or not. Its value is {Yes|No}. Default is the default value of field. Type is the type of the model, as shown in Table 1, and Datatype is the data type of field, as shown in Table 2.

Through the analysis of the semantic description model and comparing it with the information model, it can be concluded that most model elements of semantic description model are similar, or even identical to the model elements of information model, such as Entity.Name, Field.Name and Field.Nullable, ect. So these model elements can be got directly from information model, but there are also some model elements not existing in information model, such as ID, Field.Type, which should be added and perfected.

The steps of establishing semantic description model are as follows:

Step 1: The main model elements of local semantic description models are formed through information models with integration notes.

Step 2: The complete local semantic description models are formed through supplements to the above models.

Step 3: The standard semantic description models are formed through two ways. One is extracting model elements from local semantic description models and conducting further integration (bottom-up molding); the other is building the models from scratch according to the relevant standards (top-down modeling).

(2) Semantic Mapping Model (SMM)

Semantic mapping model is defined as follows.

$$SMM = \{ID, Description, Type, From, To, (EMap, EMap, EMap...)\} \quad (4)$$

ID is the unique number and Description is the additional information of the model. Type is the type of the model, valued as "internal" or "external" which respectively indicates the internal semantic mapping and the external semantic mapping. From and To represent both sides of the semantic mapping.

EMap is entity mapping model. Each SMM contains more than one EMap.

$$EMap = \{ID, Description, From, To, (FMap, FMap, FMap...)\} \quad (5)$$

ID is the unique number and Description is the additional information of the model. From and To represent both sides of semantic mapping.

FMap is field mapping model. One EMap contains more than one FMap.

$$FMap = \{ID, Description, From, To\} \quad (6)$$

ID is the unique number and Description is the additional information of the model. From and To represent both sides of semantic mapping.

Semantic mapping model and semantic description model share analogous hierarchical structure. And every semantic mapping model depends on two semantic description models.

The steps to construct semantic mapping model are as follows:

Step 1: According to the integration requirements, the relationships between information system and semantic gateway and the relationships between semantic gateway and other semantic gateways should be definite. Thus a top-level model of semantic mapping model is formed, namely the Root node.

Step 2: The mid-level model of semantic mapping model, namely the EMap nodes are formed through identifying the relationships between entities from different semantic description models.

Step 3: The bottom-level model of semantic mapping model, namely the FMap nodes are formed through

identifying the relationships between fields from different semantic description models.

(3) Semantic Publishing Model (SPM)

Semantic publishing model is defined as follows.

$$SPM = \{ID, Description, System, (EPubs, EPubs, EPubs...)\} \tag{7}$$

ID is the unique number and *Description* is the additional information of the model. *System* indicates which information *System* represents which system does conduct information publishing.

EPubs represents the publishing rules based on entity. One *SPM* contains more than one *EPubs*.

$$EPubs = \{ID, Description, Entity, (PubItem, PubItem, PubItem...)\} \tag{8}$$

ID is the unique number and *Description* is the additional information of the model. *Entity* describes on the basis of which entity the *EPubs* are established.

PubItem is one of the publishing rules. One *EPubs* contains more than one *PubItem*.

$$PubItem = \{ID, Description, Event, Trigger\} \tag{9}$$

ID is the unique number and *Description* is the additional information of the model. *Event* represents the causes of data changes, including INSERT, UPDATE and DELETE. *Trigger* represents the events triggering condition.

For example, when synchronizing the “Order” information between ERP and MES, it only needs to construct the *PubItem* model for the UPDATE event in ERP and the triggering condition which is the order status changes from "releasing" to "released".

The steps of establishing semantic publishing model are as follows:

Step 1: Identify the system with information to be integrated and its local semantic description model, and form top-level model of semantic publishing model, namely the Root node.

Step 2: Analyze which entity can provide information needed by integration and further find out the business operations associated with these entities. Then the operations conforming to integration rules are located, along with the corresponding database events and triggering conditions, to form *EPubs* models and *PubItem* models. And complete semantic publishing model will eventually be constructed

(4) Semantic Subscription Model (SSM)

Semantic Subscription Model is defined as follows.

$$SSM = \{ID, Description, System, (ESubs, ESubs, ESubs...)\} \tag{10}$$

ID is the unique number and *Description* is the additional information of the model. *System* indicates which information system is targeted to conduct information subscription.

ESubs represents the subscription rules based on entity. One *SSM* contains more than one *ESubs*.

$$ESubs = \{ID, Description, Entity, (SubItem, SubItem, SubItem...)\} \tag{11}$$

ID is the unique number and *Description* is the additional information of the model. *Entity* describes on the basis of which entity the *ESubs* are established.

SubItem is one of the subscription rules. One *ESubs* contains more than one *SubItem*.

$$SubItem = \{ID, Description, PubItem, Action\} \tag{12}$$

ID is the unique number and *Description* is the additional information of the model. *PubItem* represents that on the basis of which publishing rule the subscription information is distributed. *Action* is the operation after receiving subscription.

The steps to establish semantic subscription model are as follows:

Step1: Identify the information systems requiring information integrated from other ones and locate the corresponding local semantic description models to form the top-level model of semantic subscription model, namely Root node.

Step2: Identify the entity data in need of synchronization and integration with others and further analyze the integration items. Establish the relationships between subscription items and distribution to eventually obtain complete semantic subscription model.

IV. APPLICATION VERIFICATION

The research results of this paper have been verified in one manufacture enterprise which produces the core components of marine diesel engine. The core information systems in this enterprise are ERP and MES. And they need to interchange the “Worksheet” information, as follows: after ERP system finishes the worksheets assignment, then these assigned worksheets will appear in MES system; in turn, any updated worksheet information in MES system should be synchronized with ERP system in real time. As shown in figure 4(a).

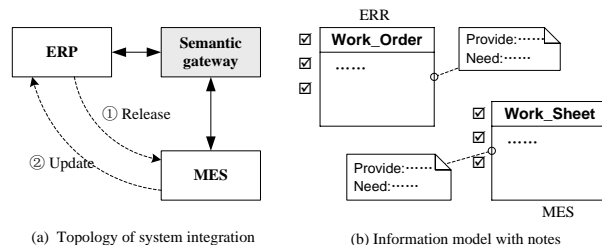


Figure 4. The worksheet’s information integration of ERP/MES

The process of establishing the information integration model based on semantic gateway:

(1) Mark the ERP/MES information models separately and then form information models with the integration

notes according to integration requirements. As shown in figure 4(b).

(2) Establish the worksheet's local semantic description model of ERP and MES respectively based on the information models with integration notes.

(3) Construct the worksheet's standard semantic description models combined with the worksheet's local semantic models of ERP/MES according to relevant standards.

(4) Establish the mapping from the worksheet's local semantic description models of ERP/MES to the worksheet's standard semantic description models separately.

(5) Establish the worksheet's semantic publishing models of ERP system according to the integration requirements.

(6) Establish the worksheet's semantic publishing models and the worksheet's semantic subscription models of MES according to the integration requirements.

Through the above steps, semantic gateway information integration models are established. The models can transform into semantic gateway setting, with the ability to support the rapid implementation of semantic gateway.

V. CONCLUSIONS

In this paper, a framework for heterogeneous systems information integration based on semantic gateway has been proposed. With the support of the framework, information integration of heterogeneous systems can be realized with more convenience and greater flexibility.

A modeling method for heterogeneous systems information integration based on semantic gateway has been put forward. The models and Modeling process has been discussed in detail. This method is able to rapidly generate the information integration models, which provide supports to the implementation and operation of semantic gateway

The application verification indicates that the research results of this paper have very good feasibility and effectiveness.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No. 51075414

REFERENCES

- [1] Xian-Di Yang, Zhi-Yong Peng, Jun-Qiang Liu, Xu-Hui L. I. An Overview of Information Integration. *Computer Science*. 2006, 33(7) pp. 55-59, 80.
- [2] Yu Jun-He, Qi Guo-Ning, Wu Zhao-Tong, Gu Xin-Jian The College Of, Zhejiang University, Hangzhou, et al. Research on Method and Application of Integrating PDM and ERP. *Computer Integrated Manufacturing Systems*. 2001(06) pp. 49-53.
- [3] Li Zhi-lin. The Data Integration in Banking Statement Management System Based on Messaging Middleware. *South China University of Technology*, 2012.
- [4] Valilai Omid Fatahi, Houshmand Mahmoud. A collaborative and integrated platform to support distributed manufacturing system using a service-oriented approach based on cloud computing paradigm. *Robotics and Computer-Integrated Manufacturing*. 2013, 29(1) pp. 110-127.
- [5] He Rongmao, Qin Futong, Hu Ran, Yu Xin. Research and Design on Heterogeneous Data Integration Middleware Based on SOA. *Ship Electronic Engineering*. 2012, 32(1) pp. 77-78, 124
- [6] Zhouyang Li, Xitian Tian, Guoding Chen. Study on CAD/CAPP/CNC Integrated System Based on STEP-NC. *China Mechanical Engineering*. 2006, 17(21) pp. 2243-2248.
- [7] Wang Ying-Lin, Qi Ke-Tao, Zhang Shen-Sheng. XML/EDI Integration Framework for Agile Supply Chain. *China Mechanical Engineering*. 2003, 14(22) pp. 1922-1925.
- [8] Li Qing, Mak Hon Chung, Zhao Jianmin, Zhu Xinzhong. OXML: an Object XML Database Supporting Rich Media Indexing and Retrieval. *Journal of Multimedia*. 2011, 6(2) pp. 115-121.
- [9] Yahia Esmat, Lezoche Mario, Aubry Alexis, Panetto Herv é Semantics enactment for interoperability assessment in enterprise information systems. *Annual Reviews in Control*. 2012, 36(1) pp. 101-117.
- [10] Yong Feng, Ming-Yu Wang. Lightweight data integration method based on semantics. *Computer Engineering and Design*. 2012, 33(1) pp. 402-406.
- [11] Buitelaar Paul, Cimiano Philipp, Frank Anette, Hartung Matthias, Racioppa Stefania. Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies*. 2008, 66(11) pp. 759-788.
- [12] Yang Tian-Jian, Wang Bo, Guo Wei, Zhao Yan-Bin. Integrated Modeloing Method of CAPP System with IDEFX and UML. *Journal of Tianjin University*. 2004(06) pp. 553-558.
- [13] Kim Cheol-Han, Weston R. H., Hodgson A., Lee Kyung-Huy. The complementary use of IDEF and UML modelling approaches. *Computers in Industry*. 2003, 50(1) pp. 35-56.
- [14] Guo Gang, Tang Hua-Mao, Luo Yu. Semantic -based Product Functional Formal Modeling. *Computer Integrated Manufacturing Systems*. 2011(6) pp. 1171-1177
- [15] Yuan Qiing-Ni, Xie Qing-Sheng, Xu Ming-Heng, Li Shao-Bo. Research on manufacturing resources ontology model based on semantic. *Journal of wuhan university of technology*. 2009(10) pp. 121-125
- [16] Song Yu-Chuan, Lei Qi, Liu Fei. A semantic gateway architecture for system integration in networked manufacturing. Laubisrutistr. 24, Stafa-Zuerich, CH-8712, Switzerland: *Trans Tech Publications Ltd*, 2010 pp. 419-420, 453-456.
- [17] Qi Lei, Yu-Chuan Song, Xian-Wang L. I. Heterogeneous system integration framework and key technologies with the support of semantic gateway. *Journal of Chongqing University (Natural Science Edition)*. 2010, 33(11) pp. 27-32.

Satellite Formation based on SDDF Method

Wang Yu, Wu Zhi-qiang, and Zhu Xin-hua

School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing, China

Email: wangyu.njust@gmail.com, wuzhiqiang@njust.edu.cn, zhuxinhua@njust.edu.cn

Abstract—The technology of satellite formation flying has been a research focus in flight application. The relative position and velocity between satellites are basic parameters to achieve the control of formation flight during the satellite formation flying mission. In order to improve the navigation accuracy, a new filter different from Extended Kalman Filter (EKF) should be adopted to estimate the errors of relative position and velocity, which is based on the nonlinearity of the kinetic model for the satellite formation flying. A nonlinear Divided Difference Filter (DDF) based on Stirling interpolation formula was proposed in this paper. According to the linearity of the measurement equation for the filter, a simplified differential filter was designed by means of expanding the polynomial of the nonlinear system equation and linear approximating of the finite differential interpolation. Digital simulation experiment for the relative positioning of satellite formation flying was carried out. The result demonstrates that the filter proposed in this paper has a higher filtering accuracy, faster convergence speed and better stability. Compared with the EKF, the estimation accuracy of the relative position and velocity has improved by 77.1% and 47% respectively in the method of simplified DDF, which indicates the significance for practical applications.

Index Terms—Divided Difference Filter; Extended Kalman Filter; Satellite Formation; Relative Navigation

I. INTRODUCTION

Satellite formation flying is a new concept in satellite applications which represents the future of space system development. Generally, cluster of formation flying satellites consists of a reference satellite and several companion satellites. All of these the satellites cooperate with each other to realize the function of a single large satellite. The system of satellite formation has the advantage of powerful function, low cost, high reliability and strong adaptability [1]. At present, formation flying has been widely used in long baseline interferometry, stereo imaging, and synthetic aperture and so on. It is important to make clear the operation state of all satellites during the formation flying. According to the needs of the mission and circumstances change, it is necessary to recombine the formation in order to provide what the user need quickly. And during the invalidation of certain satellite, compensation will be carried out immediately. For a single satellite, it needs to have the ability to enter the queue and operate autonomously based on the related information of the formation, which is supported by the autonomic location technology and station-keeping technology. The satellite also needs to realize the position

and orbit determination independently according to the nature environment feature without the external assistant information, and it is supported by the technology of attitude and position determination based on geomagnetic field for the space or ground target, spaceborne GPS orbit determination, relay satellite orbit determination, the accurate orbit determination for whole formation, attitude measurement and control for whole formation satellites, and the method to improve the orbit determination accuracy by correcting the ionosphere refraction for satellite beacon [2] [3].

In order to meet the requirement of many tasks especially for the interferometric measurement, the accurate information of relative position and attitude should be provided by the relative navigation filter. Thus nonlinear kinetic equations are used to be the filtering state equations [4]. Moreover, to design the relative navigation filter of satellite formation flying, a variety of relative measurement sensors which can provide the measurement value of the relative position, velocity and angle of sight are introduced. As a result, the nonlinearity of the measurement values and system state model will increase the nonlinearity of the filtering process. Consequently, a lot of research work has been done [5]. Garrion gave the detailed designed method by utilizing the EKF in relative navigation for satellite formation. Chen further studied the UKF and applied it to satellite formation. Heyne proposed the Sigma Point Kalman. Nocosia improved the traditional filter and proposed the idea of designing the distributed filter. Azizi studied the actuator fault estimation and designed a new distributed filter [6].

In the field of engineering application, the Extended Kalman Filter (EKF) is usually utilized for the relative navigation. The basic idea of the EKF is the linear approximation of nonlinear expressions, which realized by Taylor expansion for nonlinear state transfer equation and observation equation in the field of state space and then the first-order terms will be obtained. The above process implies the hypothesis that there exist the derivations of the state and observation equations. However, the EKF will result in the following two problems [7]: First, the linear errors induced by the first-order linear approximation of the system state equation will influence the relative navigation accuracy. For example, the nonlinearity of formation will increase as the inter-satellite baseline increased. Thus, there is a significant difference between system model based on the first order linear approximation and the actual moving

process which leads to a lower filtering accuracy; Second, the complex system state equation and the calculation of Jacobian matrix will increase the complexity of the filtering algorithm. Nonlinear distributed filters based on sampling methods to solve the above problems have appeared recently. Literature [8] and [9] presented a kind of Divided Difference Filter (DDF) using Stirling interpolation (an equidistant nodes difference interpolation method). Expanding the polynomial of the nonlinear system equation by center interpolation and approximating the finite items. Keeping the first-order difference terms as an approximation and carrying out the first-order difference filter, then we will get a higher precision than the first-order Taylor expansion [10]; Reserving the second-order difference items as an approximation, the result of the second-order difference filter will be obtained, which has a higher precision than the method of second-order Taylor expansion. Both of them are able to achieve better performance than the EKF. In addition, the DDF method doesn't have to calculate the Jacobian matrix during the filtering processing. Therefore, it has the advantage of low computational complexity and easy application.

Satellite formation flying is a typical nonlinear problem. Large error of parameter measurement in practical application makes the study of nonlinear filtering algorithms become the key to improve the accuracy of relative navigation. In the standard DDF algorithm, the process noise and observation noise are expanded as a new state vector which will dramatically increase the complexity of the algorithm as the dimension of the state vector increased. However, in most applications, especially for satellite formation flying relative positioning, the noise is additive. We can simplify the DDF algorithm to reduce the computational complexity on condition that there's no influence to the filter's performance.

For practical application, this paper proposed a new filter to estimate the relative position and velocity according to the nonlinearity of the kinetic model for satellite formation flying. Firstly, the algorithm of DDF is explained and the relative navigation kinetic equation for satellite formation flying is illustrated. Secondly, since the observed equation of the filter for satellite formation flying is linear, on the basis of the literature [8,9], this paper developed a Simplified Divided Difference Filter (SDDF) algorithm aiming at the discrete nonlinear model of linear measurement output. Finally, the simulation was carried out and the result was analyzed which proves the validity and efficient of the proposed algorithm.

II. STANDARD DIVIDED DIFFERENCE FILTER

A nonlinear system can be defined as follows:

$$x_{k+1} = f(x_k, v_k) \tag{1}$$

$$y_k = g(x_k, w_k) \tag{2}$$

where (1) is the system state equation and (2) is the observation equation. x_k and v_k are n_x -dimensional vectors, y_k and w_k are n_y -dimensional vectors. v_k

and w_k are the zero-mean white Gaussian and the variance of them are Q_k and R_k , which are mutually independent of each other and irrelevant from the state variables before time K. The augmented state variables can be got by combing the state vectors with the process noise and observation noise respectively. According to the transfer rule of random variables in nonlinear function and the principle of nonlinear filter [11], the differential filter algorithm can be obtained.

At time K, the state prediction and the filter result are \hat{x}_k and \hat{y}_k with the variances \hat{P}_k and \hat{P}_k . The Cholesky decomposition is carried out as follows:

$$\hat{P}_x \triangleq \hat{S}_x \hat{S}_x^T, \hat{P}_y \triangleq \hat{S}_y \hat{S}_y^T \tag{3}$$

$\hat{S}_{x,p}$ and $\hat{S}_{y,p}$ denote the p-th row of \hat{S}_x , \hat{S}_y , respectively. Four differential matrices are given by:

$$S_{x\hat{x}}^{(1)}(k) = \left\{ S_{x\hat{x}}^{(1)}(k)_{(i,j)} \right\} = \left\{ \frac{[f_i(\hat{x}_k + h\hat{S}_{x,j}) - f_i(\hat{x}_k - h\hat{S}_{x,j})]}{2h} \right\} \tag{4}$$

$$S_{y\hat{x}}^{(1)}(k) = \left\{ S_{y\hat{x}}^{(1)}(k)_{(i,j)} \right\} = \left\{ \frac{[g_i(\hat{x}_k + h\hat{S}_{x,j}) - g_i(\hat{x}_k - h\hat{S}_{x,j})]}{2h} \right\} \tag{5}$$

$$S_{x\hat{x}}^{(2)}(k) = \left\{ S_{x\hat{x}}^{(2)}(k)_{(i,j)} \right\} = \left\{ \frac{\sqrt{h^2-1}}{2h} [f_i(\hat{x}_k + h\hat{S}_{x,j}) + f_i(\hat{x}_k - h\hat{S}_{x,j}) - 2f_i(\hat{x}_k)] \right\} \tag{6}$$

$$S_{y\hat{x}}^{(2)}(k) = \left\{ S_{y\hat{x}}^{(2)}(k)_{(i,j)} \right\} = \left\{ \frac{\sqrt{h^2-1}}{2h} [g_i(\hat{x}_k + h\hat{S}_{x,j}) + g_i(\hat{x}_k - h\hat{S}_{x,j}) - 2g_i(\hat{x}_k)] \right\} \tag{7}$$

And then the standard DDF algorithm can be described as follows [12, 13]:

Initialization

$$\begin{aligned} \hat{x}_0 &= E[x_0] \\ \hat{P}_x(0) &= E[(x_0 - \hat{x}_0)(x_0 - \hat{x}_0)^T] \end{aligned} \tag{8}$$

State Prediction

$$\hat{x}_k = \frac{h^2 - n}{h^2} f(\hat{x}_{k-1}) + \tag{9}$$

$$\frac{1}{2h^2} \sum_{p=1}^{n_x} [f(\hat{x}_{k-1} + h\hat{S}_{x,p}) + f(\hat{x}_{k-1} - h\hat{S}_{x,p})]$$

$$\begin{aligned} \hat{P}_x(k) &= [S_{x\hat{x}}^{(1)}(k-1) \quad S_{x\hat{x}}^{(2)}(k-1)] \times \\ & \quad [S_{x\hat{x}}^{(1)}(k-1) \quad S_{x\hat{x}}^{(2)}(k-1)]^T + Q_{k-1} \\ & \triangleq \hat{S}_x(k) \hat{S}_x(k)^T \end{aligned} \tag{10}$$

Observation Prediction

$$\hat{y}_k = \frac{h^2 - n}{h^2} g(\hat{x}_k) + \frac{1}{2h^2} \sum_{p=1}^{n_x} [g(\hat{x}_k + h\hat{s}_{x,p}) + g(\hat{x}_k - h\hat{s}_{x,p})] \quad (11)$$

$$P_y(k) = \begin{bmatrix} S_{yx}^{(1)}(k) & S_{yx}^{(2)}(k) \end{bmatrix} \times \begin{bmatrix} S_{yx}^{(1)}(k) & S_{yx}^{(2)}(k) \end{bmatrix}^T + R_k \quad (12)$$

$$P_{xy}(k) = \hat{S}_x(k) [S_{yx}^{(1)}(k)]^T \quad (13)$$

State Filter Renewal

$$K_k = P_{xy}(k) [P_y(k)]^{-1} \quad (14)$$

$$\hat{x}_k = \hat{x}_k + K_k (y_k - \hat{y}_k) \quad (15)$$

$$\hat{P}_x(k) = \begin{bmatrix} \hat{S}_x(k) - K_k S_{yx}^{(1)}(k) & K_k S_{yx}^{(2)}(k) \end{bmatrix} \times \begin{bmatrix} \hat{S}_x(k) - K_k S_{yx}^{(1)}(k) & K_k S_{yx}^{(2)}(k) \end{bmatrix}^T + K_k R_k K_k \hat{S}_x(k) \hat{S}_x(k)^T \quad (16)$$

III. THE RELATIVE NAVIGATION OF FORMATION FLYING SATELLITES

A. Relative Kinetic Equation

There are two common reference frames in satellite formation flying [14]. One is Earth Center Inertial Frame (ECIF), the other is Local Vertical Local Horizontal Frame (LVLH), which are shown in Fig1. In the frame of ECIF, O is geocentric, OX points to equinox, OZ points to the North Pole and the direction of OY meets the right-hand rule with OX and OZ . In the frame of LVLH, o means the centroid of the satellite, ox is the radius vector of satellite, oz is the normal direction of orbit plane, and then the direction of oy meets the right-hand rule with ox and oz .

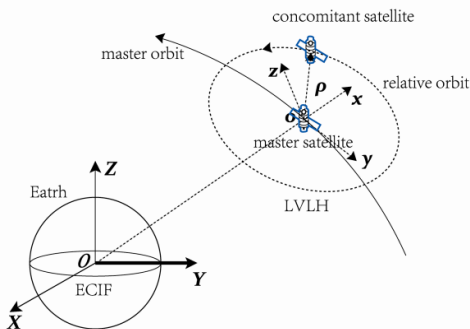


Figure 1. The local coordinate and the model of satellites formation.

The coordinate origin of formation flying satellite relative motion is located at the centroid of master satellite. x, y, z represent for the radial, tangential and normal directions respectively. The relative position

vector of formation flying satellites is defined as $\rho = [x \ y \ z]^T$ while the relative velocity vector is $\dot{\rho} = [\dot{x} \ \dot{y} \ \dot{z}]^T$.

The kinetic equation for the master satellite and concomitant satellite is [15]:

$$\begin{cases} \frac{d^2 r_c}{dt^2} + \frac{\mu}{r_c^3} r_c = f_c \\ \frac{d^2 r_d}{dt^2} + \frac{\mu}{r_d^3} r_d = f_d \end{cases} \quad (17)$$

f_c and f_d are the acceleration sum of control and perturbation for the master satellite and concomitant satellite. For the relative position of them is $\rho = r_c - r_d$, then we can get the following expression:

$$\frac{d^2 \rho}{dt^2} + \frac{\mu}{r_d^3} r_d - \frac{\mu}{r_c^3} r_c = f \quad (18)$$

Rewritten (18), we can get:

$$\begin{aligned} \frac{d^2 \rho}{dt^2} + \frac{\mu}{r_d^3} r_d - \frac{\mu}{r_c^3} r_c &= f \ddot{\rho} + 2\omega_c \times \dot{\rho} + \omega_c \times \omega_c \times \rho + \\ \dot{\omega}_c \times \rho + \mu \left(\frac{r_c + \rho}{|r_c + \rho|^3} - \frac{r_c}{|r_c|^3} \right) &= f \end{aligned} \quad (19)$$

where ω_c and $\dot{\omega}_c$ are the rotational angular velocity and acceleration for the relative motion frame relative to the inertial frame. Equation (19) is the relative motion kinetic equation for satellite formation with active control force by nonlinear relative motion and perturbation around any orbit or formation satellites. In the relative frame $o-xyz$, (19) can be written as follows according to Cartesian coordinate [16]:

$$\begin{aligned} \ddot{x} &= 2\dot{\theta}\dot{y} + \ddot{\theta}y + \dot{\theta}^2 x + \frac{\mu}{r^2} \\ &\quad - \frac{\mu(r+x)}{[(r+x)^2 + y^2 + z^2]^{3/2}} + w_x \\ \ddot{y} &= -2\dot{\theta}\dot{x} - \ddot{\theta}x + \dot{\theta}^2 y \\ &\quad - \frac{\mu y}{[(r+x)^2 + y^2 + z^2]^{3/2}} + w_y \\ \ddot{z} &= -\frac{\mu z}{[(r+x)^2 + y^2 + z^2]^{3/2}} + w_z \end{aligned} \quad (20)$$

r is the geocentric distance of the master satellite orbit, μ is the constant of the earth gravitation, and θ is the true anomaly of the master satellite. w_x, w_y and w_z are the disturbing accelerations in the direction of x, y and z , which are approximated by the zero-mean white Gaussian noise. The accelerations for the true anomaly of the master satellite and the geocentric distance of the orbit are:

$$\begin{aligned} \ddot{r} &= r\dot{\theta}^2 - \frac{\mu}{r^2} \\ \ddot{\theta} &= -2\frac{\dot{r}}{r}\dot{\theta} \end{aligned} \quad (21)$$

B. GPS Measurement Model

The carrier difference measurement equation is [17]:

$$\Delta\Phi_{ij}^k + lN_{ij}^k = D_{ij}^k + c\tau_i - c\tau_j + \eta_{ij}^k \quad (22)$$

$\Delta\Phi_{ij}^k$ is the measurement value between the master and concomitant satellites according to the k-th visible GPS satellite. l is the carrier wavelength of the GPS navigation satellite. N_{ij}^k is the single difference integer ambiguity. D_{ij}^k is the baseline vector of the formation flying satellite projected onto the line of sight vector for the k-th visible GPS satellite, which contains the relative position between formation flying satellites. $\Delta\tau = c\tau_i - c\tau_j$ is the clock error between the master and concomitant satellite, which can be modeled by the second-order Markov process:

$$\begin{aligned} \Delta\dot{\tau} &= \Delta f + w_1 & w_1 &\in N(0, S_{\Delta\tau}) \\ \Delta\dot{f} &= w_2 & w_2 &\in N(0, S_{\Delta f}) \end{aligned} \quad (23)$$

Δf is the frequency drift. η_{ij}^k is the measurement noise.

C. Orbit Model

The orbit model in this paper takes the following acceleration perturbation into consideration [18]: the nonspherical perturbation of the earth a_{NS} , the sunlight pressure perturbation a_{SRP} , the atmosphere drag perturbation a_D and the three-body perturbation a_{3B} . During the filtering process, the actual relative motion state of formation flying satellites can be obtained by solving the disturbed kinetic model of the master and concomitant satellites and then getting the difference of them.

IV. SDDF ALGORITHM FOR SATELLITE FORMATION FLYING

In many practical applications, the system model is not so complex as (1) and (2). The process noise and observation noise are additive. In this case, the proper simplification for the algorithm will not affect the filtering performance and it can reduce the computation. Take the following discrete nonlinear into consideration:

$$\begin{aligned} x_k &= f(x_{k-1}) + v_{k-1} \\ y_k &= H_k x_k + w_k \end{aligned} \quad (24)$$

Since the measurement equation is linear, the DDF process can be corresponding simplified [19]. The priori estimation of the system state is shown in (9). The priori matrix of square root for the system state error covariance matrix is given by:

$$\hat{S}_x(k+1) = HT\left(\begin{bmatrix} S_{xx}^{(1)}(k) & S_{xx}^{(2)}(k) & S_w(k) \end{bmatrix}\right) \quad (25)$$

where $S_w(k)$ stands for the square root of $Q(k)$. It satisfies $Q(k) = S_w(k)S_w(k)^T$. HT means Householder transform which transforms S into a triangular matrix. It satisfies $HT(S)HT(S)^T = SS^T$. The gain matrix $K(k+1)$ and the posterior matrix of square root for the system error covariance $\hat{S}_x(k+1)$ are determined by (26) and (27).

$$K(k+1) = \begin{bmatrix} \hat{S}_x(k+1)\hat{S}_x(k+1)^T H_k^T \\ (H_k \hat{S}_x(k+1)\hat{S}_x(k+1)^T H_k^T + R(k+1))^{-1} \end{bmatrix} \quad (26)$$

$$\begin{aligned} \hat{S}_x(k+1) &= HT\left(\left[(I - K(k+1)H_k)\hat{S}_x(k+1) \right. \right. \\ &\quad \left. \left. K(k+1)S_v(k+1)\right]\right) \end{aligned} \quad (27)$$

where $S_v(k)$ stands for the square root of $R(k)$. It satisfies $R(k) = S_v(k)S_v(k)^T$.

The innovation is defined by: $\gamma(k+1) = y_{k+1} - H_k \bar{x}_{k+1}$, and then the posteriori estimation of the system state can be obtained:

$$\hat{x}_{k+1} = \bar{x}_{k+1} + K(k+1)\gamma(k+1) \quad (28)$$

From the aforementioned filtering process, the difference between the simplified DDF and the EKF is as follows: 1) The state vector is obtained by approximating the finite terms from the center difference interpolation in the method of simplified DDF, while in the method of EKF it is calculated by the iteration of the first order approximation for the Taylor expansion of the system state equation; 2) In the prediction of the system state error covariance, the square root of the matrix is utilized to update in the simplified DDF, while the traditional Kalman filter algorithm is utilized in the EKF; 3) The symmetry and positive definition of the state error covariance matrix can be guaranteed in the simplified DDF by utilizing the square root of the matrix to update.

V. SIMULATION

In order to verify the validity and efficiency of the proposed algorithm, simulation for satellite relative navigation is carried out based on the background of Bi-Satellite formation flying. The EKF and simplified DDF are utilized in the estimation of relative position and velocity. The RMSE (Root Mean Square Error) is usually used to evaluate the performance of the filter [20]. And it is defined as follows:

$$RMSE_k = \sqrt{\frac{1}{N} \sum_{i=1}^N [(x_k^i - \hat{x}_k^i)^2 + (y_k^i - \hat{y}_k^i)^2 + (z_k^i - \hat{z}_k^i)^2]} \quad (29)$$

The initial orbital parameters of the formation flying satellite used in the simulation were: $e_c = [7178\text{km } 1 \times 10^{-5}]$

$51.7 \ 0 \ 0 \ 0]^T$, $e_{df} = [7178\text{km} \ 6.966 \times 10^{-5} \ 51.7 \ 0.010171 \ 0 \ 0.634 \ 9]^T$.

In the simulation process, the twelve dimensional state vector of the system was $x = [\rho \ \bar{\rho} \ r \ \bar{r} \ \theta \ \bar{\theta} \ \Delta\tau \ \Delta f]^T$. In order to update the clock error and frequency drift of the receiver in time, the clock error and frequency drift terms were incorporated into the navigation filter. The spectral density for the process noise in (12) was $\sqrt{5} \times 10^{-11} \text{m/s}^{3/2}$. The spectral density for the difference of the clock error and the frequency drift were $S_{\Delta\tau} = 10^{-4}$ and $S_{\Delta f} = 10^{-11}$ respectively. The measurement noise error was $0.01\text{m} (1\sigma)$. The covariance matrix of the measurement noise was $R = 1 \times 10^{-4} I \text{ m}^2$.

The elements for the variance matrix of the initial filter state error were set as follows. The relative position error was $4I_{3 \times 3} \text{m}^2$. The speed error was $0.01I_{3 \times 3} (\text{m/s})^2$. The geocentric distance and the change rate of the master satellite were 50m^2 and $0.01(\text{m/s})^2$. The true anomaly error and the change rate of the master satellite were $1 \times 10^{-4} \text{rad}^2$ and $1 \times 10^{-4} (\text{rad/s})^2$. The difference of the clock error and the frequency drift were 100m^2 and $1(\text{m/s})^2$.

The digital simulation was realized by means of MATLAB using the method of Monte Carlo and it was repeated by 20 times. The length of the simulation period was 3000s. The sampling period of GPS measurement was 1s. Different Gaussian random noises of the same distribution were adopted in the simulation. The filtering error was the average RMSE of the 20 estimation errors. The actual relative motion state of the master and concomitant satellite was obtained by the fourth-order Runge-Kutta integration of the disturbed two-body kinetic equation. The actual relative motion state of the formation flying satellites was obtained by the difference of the absolute motion state between the master and concomitant satellites. The period for the relative orbital motions of formation flying satellites is about five orbital periods in the situation with or without space perturbation. The results are shown in Fig. 2 and Fig. 3. It is obvious that the relative motion trajectory diverges significantly when there is the orbit perturbation.

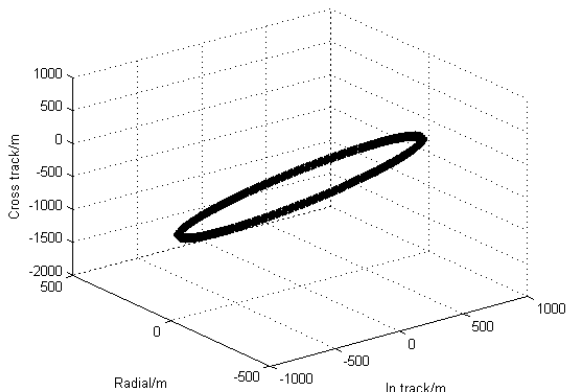


Figure 2. The motion orbit without orbit perturbation

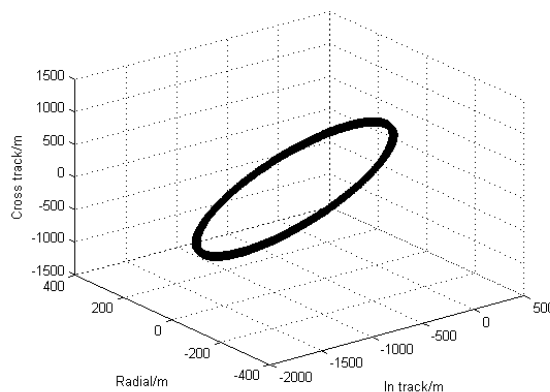


Figure 3. The motion orbit with orbit perturbation

The results of uniaxial relative state error are shown in Fig. 4 to Fig. 7 by the filter of EKF and simplified DDF. The simulation results indicates that it needs much time for convergence to estimate the errors of relative position and velocity when using the method of the EKF, while in the method of DDF the convergence time to estimate the relative position error is accelerated but the convergence time to estimate the relative velocity error is the same as the EKF. Furthermore, after the convergence, the error fluctuations for the EKF are wildly while the error fluctuations are quite small. According to the statistical result, in the method of the EKF, the relative position error was $\leq 0.0094\text{m}(1\sigma)$ and the relative velocity error was $\leq 0.0051\text{m/s}(1\sigma)$, while in the method of simplified DDF the relative position error was $\leq 0.00215\text{m}(1\sigma)$ and the relative velocity error was $\leq 0.0027\text{m/s}(1\sigma)$. Therefore, compared with the EKF, the estimation accuracy of relative position and velocity was improved by 77.1% and 47% in the method of simplified DDF.

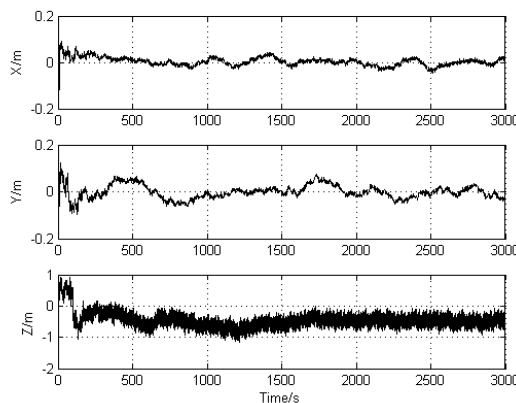


Figure 4. The relative position error of the EKF

In order to compare the performance of the two algorithms better, the 2-norm of the relative errors for the position and velocity were calculated which were illustrated in Fig.8 and Fig.9. For the reason that the system state equation of formation flying satellites is nonlinear, the EKF only carried out a first-order linear approximation during the filtering process, while the simplified DDF utilized a difference interpolation to approach the nonlinear model of the system. Thus, the

latter can achieve the accuracy with faster convergence speed and better stability when compared the EKF.

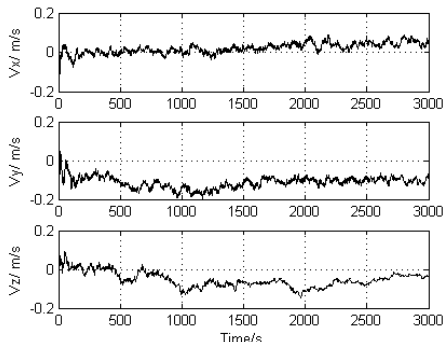


Figure 5. The relative speed error of the EKF

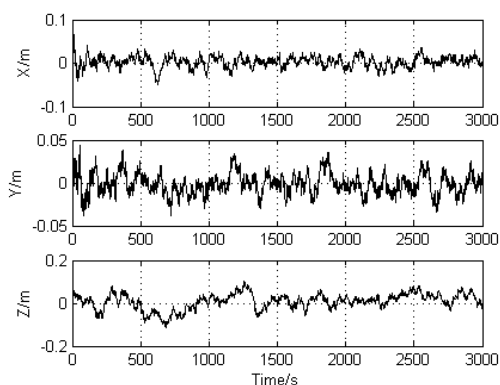


Figure 6. The relative position error of the simplified DDF

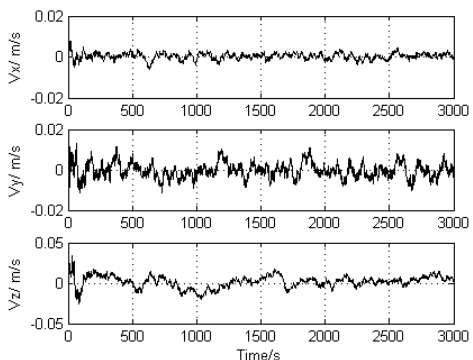


Figure 7. The relative speed error of the simplified DDF

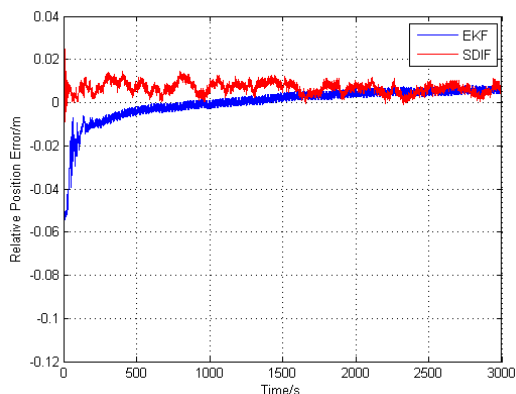


Figure 8. The relative position error of the EKF and the SDDF

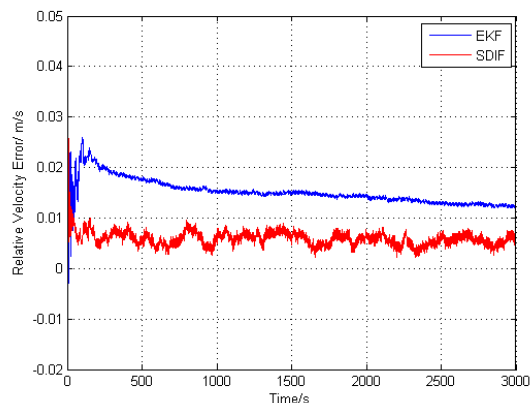


Figure 9. The relative speed error of the EKF and the SDDF

VI. CONCLUSIONS

As we know, the accuracy for error estimation will decrease with slow convergence rate when applying the EKF to solving the nonlinearity for the system error model of the relative navigation in satellite formation flying. According to the linear order of the error model for the nonlinear system and the realization of real-time filtering, a nonlinear Divided Difference Filter based on Stirling interpolation formula was proposed for the relative navigation of formation flying satellites. According to the approximation level of the DDF in the error model for the nonlinear system and the characteristic of the linearity of the measurement equation, the state equation for the filter was established by the polynomial expansion of nonlinear system equation and the approximation of finite differential interpolation. The relative kinetic equation with actual orbit perturbation was considered as the filtering model for the performance comparison between the simplified DDF with the EKF in the relative navigation when using the GPS carrier differential measurement method. The simulation for the performance verification of the relative navigation filter is carried on in the environment of Bi-Satellite formation flying. And the simulation result demonstrates that the proposed simplified DDF can achieve higher precision, faster convergence speed and better stability than the EKF during the flying process of the satellite formation. Thus it is feasible to use the simplified DDF in the design of the relative navigation filter for formation satellites which is significant for actual application.

ACKNOWLEDGMENT

This work was supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2012BAJ23B02).

REFERENCES

[1] DONG Xiao-Guang, CAO Xi-Bin and ZHANG Jin-Xiu, "A Robust Adaptive Control Law for Satellite Formation Flying," *ACTA AUTOMATICA SINICA*, vol. 39, no. 2, pp. 132-141, 2013.

- [2] HUANG Hai-bin, MA Guang-fu, and ZHUANG Yu-fei, "Real-Time Re-Planning for Satellite Formation Reconfiguration in Deep Space," *Journal of Astronautics*, vol. 33, no. 3, pp. 325-333, 2012.
- [3] LU Jian-ting, CAO Xin-bin and GAO Dai, "Relative attitude control of satellite formation flying," *Journal of Harbin Institute of Technology*, vol. 42, no. 1, pp. 9-12, 2010.
- [4] Kim, S. G., Crassidis, J. L., Yang, C. and Fosbury, A. M., "Kalman filtering for relative spacecraft attitude and position estimation," *Journal of Guidance, Control, and Dynamics*, vol. 30, no. 1, pp. 133-143, 2007.
- [5] Tao Jiuliang, *Research on relative measurement for formation flying satellite*. Shanghai Jiao Tong University, China, 2012.
- [6] Azizi, S. M. and Khorasani, K. A., "Distributed Kalman Filter for Actuator Fault Estimation of Deep Space Formation Flying Satellites," *Systems Conference*, pp. 354-359, March, 2009.
- [7] Chanier F, Checchin P and Blanc C, et al. Comparison of EKF and PEKF in a SLAM context. *International IEEE Conference on Intelligent Transportation Systems*, pp. 1078-1083, October, 2008.
- [8] Norgaard, M., Poulsen, N. G. and Ravn, O., "New developments in state estimation for nonlinear systems automatic," *Automatica*, vol. 36, no. 11, pp. 1627-1638, 2000.
- [9] Norgaard, M., Poulsen, N. G. and Ravn, O., "Advances in Derivative-Free State Estimation for Nonlinear Systems," *Technical report, IMM-REP-1998-15*, Technical University of Denmark, 2800 Lyngby, Denmark, April, 2000.
- [10] Mu, Jing, Cai, Yuan-Li, "Iterated divided difference filter and its applications," *Control and Decision*, vol. 26, no. 9, pp. 1425-1428, 2011.
- [11] Bar-Shalom, Y., Li, X. and Kirubarajan, T, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2001.
- [12] Tang, Xiao-Jun, Wei, Jian-Li and Chen, Kai, "Square-root divided difference Rauch-Tung-Striebel smoother," *Journal of Harbin Institute of Technology (New Series)*, vol. 19, no. 5, pp. 36-40, 2012.
- [13] Karlgaard, Christopher D., and Shen, "Haijun. Robust state estimation using desensitized Divided Difference Filter," *ISA Transactions*, vol. 52, no. 5, pp. 629-637, 2013.
- [14] Yoo, S. M., Song, Y. J., Park, S. Y. and Choi, K. H., "Spacecraft formation flying for Earth-crossing object deflections using a power limited laser ablating," *Advances in Space Research*, vol. 43, no. 12, pp. 1873-1889, 2009.
- [15] Mei Jie and Ma Guang-Fu, "Robust adaptive control of relativeorbit for nearby spacecraft," *Journal of Astronautics*, vol. 31, no. 10, pp. 2276-2282, 2010.
- [16] Schaub, H. and Junkins, J. L., *Analytical mechanics of space systems*. Aiaa, 2003.
- [17] Lane, C. M, *Formation design and relative navigation in high earth orbits*, Department of Aerospace Engineering Sciences, University of Colorado, Colorado USA, 2007.
- [18] Tewari, A. *Atmospheric and Space Flight Dynamics: Modeling and Simulation with MATLAB and Simulink*. Birkhauser Verlag, Viadukt-strasse 42, CH-4051 Basel, Switzerland, 2007.
- [19] Subrahmanya, N. and Shin, Y. C., "Adaptive divided difference filtering for simultaneous state and parameter estimation," *automatic*, vol. 45, no. 7, pp. 1686-1693, 2009.
- [20] WU Shun-hua, XIN Qin and WAN Jian-wei, "Simplified DDF Algorithm and the Application for Satellite to Satellite Passive Orbit Determination and Tracking," *Journal of Astronautics*, vol. 30, no. 4, pp. 1557-1563, 2009.

Heterogeneous Web Data Extraction Algorithm Based On Modified Hidden Conditional Random Fields

Cheng Cui

Department of Information Engineering Jilin Police College, Changchun, China

Email: ch_cui@163.com

Abstract—As it is of great importance to extract useful information from heterogeneous Web data, in this paper, we propose a novel heterogeneous Web data extraction algorithm using a modified hidden conditional random fields model. Considering the traditional linear chain based conditional random fields can not effectively solve the problem of complex and heterogeneous Web data extraction, we modify the standard hidden conditional random fields in three aspects, which are 1) Using the hidden Markov model to calculate the hidden variables, 2) Modifying the standard hidden conditional random fields through two stages. In the first stage, each training data sequence is learned using hidden Markov model, and then implicit variables can be visible. In the second stage, parameters can be learned for a given sequence. (3) The objective functions of hidden conditional random fields are revised, and the heterogeneous Web data are extracted by maximizing the posterior probability of the modified hidden conditional random fields. Finally, experiments are conducted to make performance evaluation on two standard datasets—"EData dataset and "Research Papers dataset". Compared with the existing Web data extraction methods, it can be seen that the proposed algorithm can extract useful information from heterogeneous Web data effectively and efficiently.

Index Terms—Hidden Conditional Random Fields; Web Data Extraction; Undirected Graph; Potential Energy Function; Hidden Variable

I. INTRODUCTION

With the popularization of the World Wide Web, a great number of data from different domains have become available. Hence, the popularity of the World Wide Web provides opportunities for users to benefit from the useful Web data. In the traditional modes, users search Web data by browsing Web page and searching by keyword, which are intuitive forms of seeking data from the Web. Unfortunately, the above searching methods have several limitations. The browsing behavior is not suitable to locate particular contents of the Web data, the reasons lie in that following links does not make sense and it is easy to get lost. On the other hand, although searching by keyword is much efficient than browsing behavior, it usually returns massive data, which is beyond users' processing ability. Therefore, in spite of being publicly and readily available, it is hard to extract useful information from the Web data [1].

To extract useful information from Web data more effectively, many existing methods resorted to the ideas from the research domain of database area. As is well known that the data in database are belonged to structured data, hence, traditional database techniques can not used in Web data extraction. There it is very important to extract useful information from the unstructured and semi-structured Web data, and then to populate databases for further data process [6-10].


Structured data in Web pages usually include important information. Such web data are often obtained from underlying databases and displayed in Web pages using fixed templates. Particularly, the structured data objects are denoted as data records. Extracting data records enables one to integrate data from multiple Web sites and pages to provide value-added services [2]. As is shown in Fig.1 (a), an E-commerce website is used as an example, and information of three kinds of iPhone products are illustrated. In Fig.1 (b), a Web page segment containing a data table is given, of which each table row is a data record. However, seeking and organizing Web data is not easy, because traditional database techniques can not be utilized directly in Web data extraction [3].

In the research field of heterogeneous Web data extraction, hidden conditional random fields are powerful technologies, which could model the state sequence as being conditionally Markov given the observation sequence. Hidden conditional random fields are designed based on Conditional random fields which can be used in many complex computing tasks, and have been widely exploited in the field of intelligent information processing [4].


However, there are some limitations in Conditional random fields, for example, Conditional random fields cannot capture intermediate structures using hidden-state variables. Therefore, we use Hidden conditional random fields in heterogeneous Web data extraction. Hidden conditional random fields utilize the intermediate hidden variables to construct the latent structure of the input data [5].

The main innovations of this paper lie in the following aspects:

(1) The proposed algorithm uses the hidden Markov model to calculate the hidden variables more accurately and modified the standard hidden conditional random



Apple iPhone 5 16GB (Black) - Unlocked by Apple


~~750.00~~ **\$709.80** 

Order in the next **9 hours** and get it by Wednesday, Jul 10.

More Buying Choices

\$569.99 new (125 offers)

\$489.99 used (65 offers)



Apple iPhone 3G 8GB (Black) - AT&T by Apple


~~549.00~~ **\$142.88**

In Stock


More Buying Choices

\$142.88 new (14 offers)

\$70.99 used (77 offers)



Apple iPhone 4 32GB (Black) - Verizon by Apple

~~600.00~~ **\$314.50** 

Order in the next **9 hours** and get it by Wednesday, Jul 10.

More Buying Choices

\$300.00 new (15 offers)

\$163.99 used (84 offers)

(a) A list of product in Amazon

Stock #	Model	Description	Odometer	Price↑
57605	Dodge SX 2.0	Loaded/Keyless	28000	14495
58205	Dodge SX 2.0	Loaded/Keyless	19500	15495
57805	Chrysler Sebring Touring	Keyless/Trac Cont	31500	15995
58465	Chrysler Sebring Touring	Keyless/Trac Cont	32500	15995
58455	Chrysler Sebring Touring	Keyless/Trac Cont	34000	16695
58495	Chrysler Sebring Touring	Keyless/Trac Cont	22500	16695
58375	Chrysler PT Cruiser	Cruise/KeylessD	29500	17795
58475	Dodge Grand Caravan	Quads/Rear AC	52000	19895
58285	Dodge Grand Caravan	Sto&Go/Keyless	43500	21695
57965	Chrysler PT Cruiser Convertible	Touring/Loaded	7000	22195

(b) Data records in datatable

Figure 1. An example of a Web page with structured data

fields through two stages. In the first stage, each training data sequence is learned using hidden Markov model, and then implicit variables can be visible. In the second stage, for a given sequence, parameters can be learned.

(2) The objective function of hidden conditional random fields is modified in our algorithm, and the heterogeneous Web data are extracted by maximizing the posterior probability of the modified hidden conditional random fields.

The rest of the paper is organized as the following sections. Section 2 introduces the related works. Section 3 illustrates the proposed scheme for heterogeneous Web data extracting by modified hidden conditional random fields. In section 4, a series of experiments are designed and conducted to make performance evaluation. Finally, we conclude the whole paper in section 5.

II. RELATED WORKS

In this section, we will survey related works of this paper in two aspects, which include 1) Web data extraction related works and 2) applications of Hidden conditional random fields.

Su et al. presented a data extraction system based on the browser and the Web services resource framework. This framework is designed based on the interaction of DE system, which regulates the access and the operation state of the Web service specification. The coupling

interaction DE system can provide users with packaging status, and the status of the Web service (named Web service resources). The interactive DE system can accept parameters from the application in the extraction process by notification WSRF design pattern as well [11].

Kayed et al. regarded the data extraction problem as the page generated decoding process based on structured data and tree templates. Furthermore, the authors propose an unsupervised, page-level data extraction method. In this method, each individual site is derived from deep architecture and a template is proposed to contain a single or multiple data records in a page. The FiVaTech system applied tree matching, tree alignment, and data mining to implement an effective computing process [12].

Lin et al. proposed a novel method to acquire collocations from the Web. Three classical lexical association measures (co-occurrence frequency, mutual information and t-test) are used to automatically extract collocation. Based on the experimental results, the benchmarks indicate that superior performance of this new Web-based approach in both high precision and recall [13].

Cafarella et al. proposed three typical extraction systems which can be implement on most of the websites. Particularly, the TEXTRUNNER system is designed based on natural language texts, and the WEBTABLES system focuses on HTML-embedded tables. Moreover, in

this paper, several unique data applications are discussed which are enabled by aggregating extracted Web data [14].

Zhu et al. illustrated a paradigm of hierarchical model which is used to achieve an integrated Web data extraction. Furthermore, this model model is named dynamic hierarchical Markov random field (DHMRFs). Considering the structural uncertainty, in DHMRFs, the joint distribution, model structure and class label are defined. Particularly, joint distribution refers to exponential family distribution. As a powerful model, DHMRFs relax the indicator model through the independence assumption. Hence the exact inference is intractable, a variational approach is proposed to construct a parameter learning model [15].

Afterwards, the applications of hidden conditional random fields are given as follows.

Indio et al. developed a system named TPpred, which is a novel predictor of organelle-targeting peptides based on Grammatical-Restrained Hidden Conditional Random Fields. The proposed system is trained on a non-redundant dataset of proteins where the presence of a target peptide was experimentally validated [16].

Wang et al. proposed a hidden conditional random field model with independent component analysis mixture feature functions for event classification for videos. Hidden conditional random field model refers to a discriminative model without conditional independence assumption of observations, and can be used in the application of video analysis as well. [17].

Qian et al. utilized hidden conditional random field to model the observations of mid-level semantics of an event clip for event detection. Comparisons are made with the dynamic Bayesian networks, hidden Markov model, enhanced HMM and conditional random field-based event detection approaches [18]. Other research works about hidden conditional random fields' application please refer to [19] and [20]

III. THE PROPOSED SCHEME

Conditional random fields belong to statistical modeling methods, which are widely utilized in pattern recognition and information retrieval. On the other hand, conditional random fields is a discriminative undirected probabilistic graphical model. Particularly, Conditional random fields can be used to find applications in shallow parsing, named entity recognition and gene finding and so on. As is well known that Conditional random fields is an alternative model to hidden Markov models. The framework of conditional random fields is shown in Fig.2.

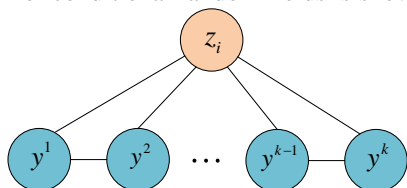


Figure 2. Framework of conditional random fields

For the undirected graph $G=(V,E)$, V and E represent the nodes and edges of the graph G

respectively. z refers to the data which are required to be observed. y^i denotes a random variable, and the value of y^i is chosen from a labeled set, Moreover, $y=[y^1,y^2,\dots,y^k]^T$ is satisfied. If $y^i(i \in [1,k])$ meets the Markov properties, (y,z) is called conditional random fields. In the theory of conditional random fields, for a given z which is observed, the distribution of label sequence y meets the following equation.

$$p(y|z) = \frac{1}{Z} \exp\{\sum_j \phi_s(y^j, z) + \sum_{l \in N_j} \phi_r(y^j, y^l | z)\} \quad (1)$$

where the parameter Z should be normalized, and N_j refers to the neighbor node of G . Next, $\phi_s(y^j, z)$ denotes the probability of the node j which is labeled by y^j according to the observing information, and $\phi_r(y^j, y^l | z)$ refers to the dependency degree between node j and its neighbors.

However, the traditional linear chain based conditional random fields can not effectively model complex and heterogeneous Web data. Hence, in this paper, we proposed modified hidden conditional random fields and efficient Web data extraction algorithm to implement a heterogeneous Web data extraction process with high accuracy. The framework of the standard hidden conditional random fields is shown in Fig. 3

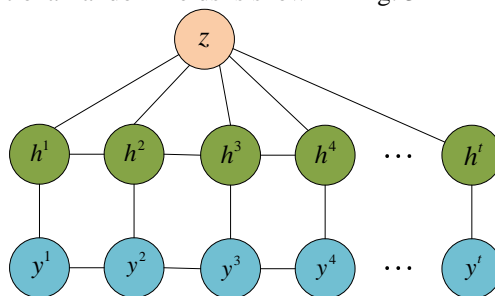


Figure 3. Framework of hidden conditional random fields

Hidden conditional random fields make a judgement according to the input sequence y belonged to the category of z . y refers to local observation vector, and $y=\{y^1,y^2,\dots,y^t\}$ is satisfied. Particularly, y^i denotes a feature vector. For a given sequence y , there may existing a related vector $h=\{h^1,h^2,\dots,h^t\}$, $h^i \in H$, where H means all the set which includes all possible hidden states. The modeling process for observed data using hidden conditional random fields is represented by the following equation.

$$\begin{aligned} p(z|y, \theta, \omega) &= \sum_h p(z, h|y, \theta, \omega) \\ &= \frac{\sum_h e^{\phi(z, h, y; \theta, \omega)}}{\sum_{z \in Z, h \in H^m} e^{\phi(z, h, y; \theta, \omega)}} \end{aligned} \quad (2)$$

where the potential energy function $\phi()$ with parameter θ and ω is used to represent the influence degree between hidden state structure, observing data and data category, and parameter ω refers to the range of the window. $\phi()$ is calculated by the following equation.

$$\begin{aligned} \phi(z, h, y : \theta, \omega) = & \sum_{(j,k) \in \gamma} \theta_e[z, h_j, h_k] + \sum_{j=1}^m \theta_z[z, h_j] \\ & + \sum_{j=1}^m \varphi(y, j, \omega) \cdot \theta_h[h_j] \end{aligned} \quad (3)$$

In Eq. 3, γ refers to a chain structure based graph, of which each node corresponds to a hidden state of time t , and $\varphi(y, j, \omega)$ represents the arbitrary feature of the observing window. The key step of the hidden conditional random fields' modeling process is the model training process and parameter learning. The parameter θ is equal to $[\theta_e, \theta_z, \theta_h]$. θ_e is used to measuring the compatibility degree between the continuous state j, k , and the variable z . θ_z refers to the compatibility degree between state j and the variable z and θ_h is the parameter of the state h_j ($h_j \in H$). The training process of the above model utilizes the following objective function.

$$\Gamma(\theta) = \sum_{i=1}^n \log p(z_i | y_i, \theta, \omega) - \frac{\|\theta\|^2}{2\sigma^2} \quad (4)$$

where parameter n is the total number of sequences in the training set. Parameters of hidden conditional random fields can be estimated by the following formula.

$$\Gamma(\theta) = \sum_{i=1}^n \log p(z_i | y_i, \theta, \omega) - \frac{\|\theta\|^2}{2\sigma^2} \quad (5)$$

$$L = \arg \max_L p(L | X, \theta) \quad (6)$$

where L represents the label of the observing data, and it can be deduced by Eq.7 as follows.

$$p(L | H, X, \theta) = \frac{1}{Z(L, X, \theta)} \exp\{\phi(L, H, X, \theta)\} \quad (7)$$

However, the difficulties of extracting the heterogeneous Web data directly using hidden conditional random field lie in that the proposed model depends on the initial choice of parameters severely. Therefore, in order to improve the accuracy of Web data extraction, we propose a modified hidden conditional random fields model. The main idea is to use the hidden Markov model to calculate the hidden variables more accurately.

The main modifications of the hidden conditional random fields are implemented through two stages. In the first stage, each training data sequence is learned using hidden Markov model, and then implicit variables can be

visible. A particular hidden Markov class can be solved by the following equation.

$$\begin{aligned} \theta^*(c) = & \arg \max_{\Theta} \sum_{i=1}^{s(c)} p(Y_i | z = c; \Theta) \\ = & \arg \max_{\Theta} \log \sum_{i=1}^{s(c)} \sum_H p(Y_i, H | z = c; \Theta) \end{aligned} \quad (8)$$

After the implicit variables becoming visible, in the second stage, for a given sequence y , the parameter θ^* can be learned from the above process. Next, the category of which the test sequence y is belonged to can be obtained by the following equation.

$$p(z | Y; \Theta) = \frac{p(z, Y; \Theta)}{p(Y; \Theta)} = \frac{e^{\phi(z, H(Y); \Theta)}}{\sum_y e^{\phi(z, H(Y); \Theta)}} \quad (9)$$

Afterwards, the objective function of hidden conditional random fields is modified as Eq.10.

$$L(\Theta) = -\sum_{i=1}^s \log p(z_i | y_i, \Theta) + \frac{\|\Theta\|^2}{2\sigma^2} \quad (10)$$

Based on the above analysis, the heterogeneous Web data (denoted as z) can be extracted from the websites by maximizing the posterior probability of the modified hidden conditional random fields.

$$z^* = \arg \max_{z \in Z} p(z | Y; \Theta^*) \quad (11)$$

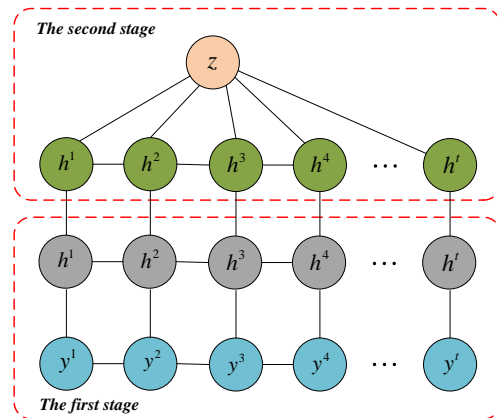


Figure 4. Framework of modified hidden conditional random fields

IV. EXPERIMENTS

A. Dataset and Performance Evaluation Metric

In this paper, two standard datasets are used, which are 1) EData dataset [6] and 2) Research Papers dataset [23].

We use the EData dataset [6] to testify the performance of the Web data extraction. The EData dataset is created to test the precision of Web data extraction on online advertising. As is illustrated in paper [6], EData is made up of 18,000 online advertisements, which were extracted from Craigslist, Ebay, and KSL. The advertisements in EData are uniformly distributed among the eight chosen domains, which are Cars-for-Sale, Computer Science

Jobs, Food Coupons, Furniture, Houses-for-Sale, Jewelry, Motorcycles-for-Sale, and Musical Instruments, and there are 750 advertisements in each of the eight advertisements domains extracted from each of the three advertisements websites. The advertisements domains in EData vary in terms of their 1) diversity, which include advertisements in jobs, transportation, food, housing, and entertainment that are essential to our daily lives, 2) advertisements sizes, from arbitrary long advertisements to relatively short ones, and 3) word distribution, different word usage in closely related advertisements which are similar in contents and nature. Moreover, advertisements in EData were extracted from various online data with different data structures, which can be utilized to test the performance of Web data extraction approach.

Research Papers dataset [23] is made up of the headers of research papers. The header of a research paper is defined to be all of the words from the beginning of the paper, usually the introduction, or to the end of the first page, whichever occurs first. We randomly select 500 for training and the other ones for testing. This sub-dataset is denoted as RP.

To evaluate the performance of Web data extraction, we use Recall, Precision and F1. For Web data extraction dataset i (denoted as d_i), the number of information we extract in d_i is $N_f(d_i)$ and the number of ground truth information in d_i is $N_g(d_i)$. The precision and recall of d_i is defined as follows.

$$P(d_i) = \frac{|N_f(d_i) \cap N_g(d_i)|}{|N_f(d_i)|} \tag{12}$$

$$R(d_i) = \frac{|N_f(d_i) \cap N_g(d_i)|}{|N_g(d_i)|} \tag{13}$$

To obtain the weighted harmonic mean of precision and recall, the F1 metric is used as well.

$$F1 = \frac{2 \times P(d_i) \times R(d_i)}{P(d_i) + R(d_i)} \tag{14}$$

Recall measures the accuracy of returning ground-truth information which are extracted from Web data, while Precision assesses the ability of excluding false positives. F1 metric can calculate the fitness of ground-truth and detect information extracted from Web data by jointly considering recall and precision.

B. Experimental Results and Analyses

Firstly, we test the performance of the proposed algorithm on EData dataset, and eight kinds of online advertisements are utilized, including: “Cars”, “Food”, “Furniture”, “Houses”, “Jewelry”, “CS Jobs”, “Motorcycles”, and “Musics”. Particularly, the average of all the above kinds of online advertisements is given as well. As is shown in Fig.5, we can see that the average precision, recall and F1 values are all larger than 0.8.

Therefore, the proposed algorithm can effectively extract useful online advertisements information from EData dataset.

Afterwards, to make performance evaluation more objective, some existing information extraction methods are compared with the proposed method, which are: 1) ADEx [6], 2) WDE-SS [21], 3) DHMRFs [22], 4) CRFs [23], 5) 2D-CRFs [24]. Then, the precision, recall and F1 are utilized to make performance evaluation, and the experimental results are shown in Fig.6, from which we find that the proposed algorithm performs better than other methods.

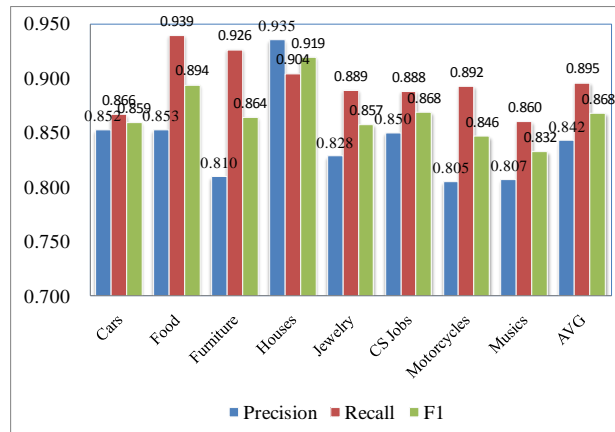


Figure 5. Performance evaluating for different kinds of online advertisements utilizing the proposed algorithm

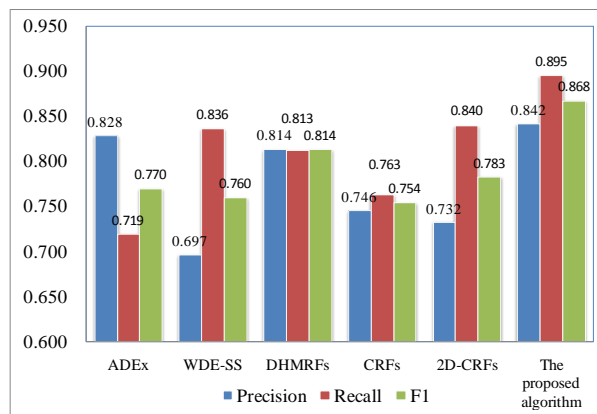


Figure 6. Performance evaluating for different methods

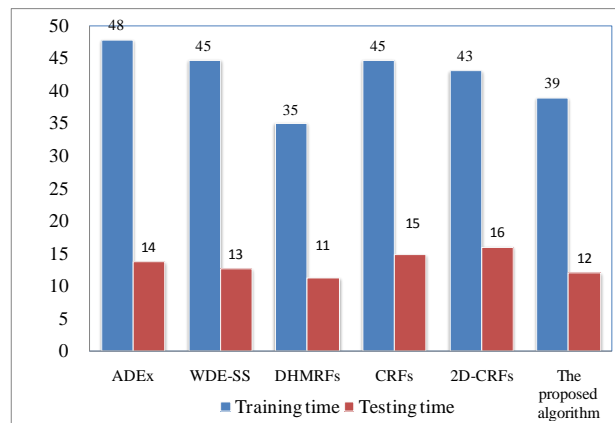


Figure 7. Comparison of time cost for different method

TABLE I. DATA EXTRACTION RESULTS USING RP DATASET

Category	ADEx			WDE-SS			DHMRFs		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Title	0.820	0.776	0.797	0.914	0.676	0.777	0.908	0.806	0.854
Author	0.806	0.929	0.863	0.638	0.701	0.668	0.802	0.863	0.831
Affiliation	0.626	0.757	0.685	0.616	0.804	0.697	0.721	0.788	0.753
Address	0.804	0.753	0.778	0.692	0.599	0.642	0.804	0.828	0.816
Note	0.708	0.787	0.745	0.649	0.921	0.761	0.865	0.850	0.857
Email	0.684	0.812	0.742	0.515	0.802	0.627	0.707	0.876	0.783
Date	0.756	0.750	0.753	0.714	0.879	0.788	0.844	0.824	0.834
Abstract	0.795	0.682	0.734	0.710	0.653	0.680	0.790	0.799	0.794
Phone	0.755	0.755	0.755	0.643	0.662	0.652	0.790	0.890	0.837
Keyword	0.582	0.631	0.605	0.735	0.779	0.756	0.703	0.728	0.715
Web	0.878	0.669	0.759	0.759	0.699	0.728	0.915	0.809	0.859
Degree	0.845	0.749	0.794	0.777	0.821	0.798	0.846	0.907	0.876
Pubnum	0.629	0.773	0.694	0.767	0.633	0.694	0.727	0.858	0.787
Category	CRFs			2D-CRFs			The proposed algorithm		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Title	0.541	0.514	0.527	0.869	0.797	0.831	0.938	0.880	0.908
Author	0.556	0.551	0.554	0.792	0.804	0.798	0.847	0.936	0.889
Affiliation	0.502	0.749	0.601	0.690	0.796	0.740	0.768	0.833	0.799
Address	0.696	0.627	0.660	0.735	0.801	0.767	0.862	0.874	0.868
Note	0.697	0.876	0.776	0.782	0.844	0.812	0.901	0.939	0.920
Email	0.709	0.698	0.704	0.643	0.842	0.729	0.767	0.878	0.819
Date	0.560	0.638	0.596	0.849	0.795	0.821	0.915	0.896	0.905
Abstract	0.750	0.807	0.778	0.759	0.697	0.727	0.807	0.822	0.815
Phone	0.541	0.514	0.527	0.869	0.797	0.831	0.938	0.880	0.908
Keyword	0.556	0.551	0.554	0.792	0.804	0.798	0.847	0.936	0.889
Web	0.502	0.749	0.601	0.690	0.796	0.740	0.768	0.833	0.799
Degree	0.696	0.627	0.660	0.735	0.801	0.767	0.862	0.874	0.868
Pubnum	0.697	0.876	0.776	0.782	0.844	0.812	0.901	0.939	0.920

In the following parts, we will compare the training time and testing time for the above methods under EData dataset utilizing the same PC. All the experiments are conducted on the PC with Intel Core i7 processor, the main frequency of this CPU we used is 2.9GHz. The capacity of the internal memory we chosen is the 4GB, and the hard disk we utilized is 1.0TB. Based on the above hardware settings, the algorithm running time are shown in Fig 7.

As is shown in Table.1, it can be seen that under the dataset RP, the proposed algorithm performs significantly better than other methods.

Based on the above experimental results, it can be seen that the proposed heterogeneous Web data extraction algorithm is effective both in the system performance and in the system efficiency. The reasons lie in the following aspects:

(1) The proposed algorithm utilizes the hidden Markov model to calculate the hidden variables more accurately and modified the standard hidden conditional random fields through two stages. Particularly, the objective function of hidden conditional random fields is modified, and the heterogeneous Web data are extracted by maximizing the posterior probability of the modified hidden conditional random fields. Therefore, the proposed algorithm can effectively extract useful information from the heterogeneous Web data with high accuracy.

(2) ADEx is currently designed to extract data from advertisements that include a single product/service in an ad. Particularly, ADEx should be enhanced to solve any online advertisements that include multiple products,

such as in video games advertisements. Furthermore, even though ADEx currently handles advertisements from various domains, it is less accurate in distinguishing ads in closely related domains, such as advertisements on different means of transportation, or advertisements that advertise professional jobs that cross multiple disciplines, such as Bioinformatics. Therefore, ADEx should be extended to discern advertisements with closely related domains.

(3) The approach of WDE-SS only considers the structure of Web documents. How to combine semantic information in documents to further improve the proposed framework is to be deeply studied. As WDE-SS requires to parse documents into trees in memory, it is not efficient on very large data sets.

(4) CRFs denotes investigates the issues of regularization, feature spaces, and efficient use of unsupported features in CRFs, with an application to information extraction from research papers. Conditional random fields (CRFs) are undirected graphical models trained to maximize a conditional probability. However, fundamental advances in regularization for CRFs should be studied in future research works.

(5) 2D-CRFs refers to a two-dimensional Conditional Random Field model, which provides a new way of incorporating two-dimensional neighborhood dependencies to improve the performance of Web information extraction. By marginalizing variables progressively along the diagonals, efficient parameter learning and labeling can be performed. When the proposed model is applied to product information extraction, significant

improvements are achieved compared with linear-chain CRF models.

V. CONCLUSIONS

In this paper, we present a heterogeneous Web data extraction algorithm based on a modified hidden conditional random fields model. The standard hidden conditional random fields are modified by three aspects. Firstly, we utilize the hidden Markov model to calculate the hidden variables. Secondly we modify the standard hidden conditional random fields through two stages. In the first stage, each training data sequence is learned using hidden Markov model, and then implicit variables can be visible. In the second stage, parameters can be learned for a given sequence. Thirdly, the objective function of hidden conditional random fields is revised, and the heterogeneous Web data can be obtained by maximizing the posterior probability of the modified hidden conditional random fields.

REFERENCES

- [1] Laender AHF, Ribeiro-Neto BA, da Silva AS, "A brief survey of Web data extraction tools", *SIGMOD Record*, 2002, 31(2) pp. 84-93.
- [2] Zhai Yanhong, Liu Bing, "Structured data extraction from the web based on partial tree alignment", *IEEE Transactions on Knowledge and Data Engineering*, 2006, 18(12) pp. 1614-1628.
- [3] Lage JP, da Silva AS, Golgher PB, "Automatic generation of agents for collecting hidden Web pages for data extraction", *Data & Knowledge Engineering*, 2004, 49(2) pp. 177-196.
- [4] Chen Shi-Wen, Wu Jiang-Xing, Ye Xiao-Long, Guo Tong, "Distributed denial of service attacks detection method based on conditional random fields", *Journal of Networks*, 2013, 8(4) pp. 858-865.
- [5] Qattoni Ariadna, Wang Sybor, Morency Louis-Philippe, "Hidden conditional random fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(10) pp. 1848-1853.
- [6] Pera Maria S., Qumsiyeh Rani, Ng Yiu-Kai, "Web-based closed-domain data extraction on online advertisements", *Information Systems*, 2013, 38(2) pp. 183-197.
- [7] Furche Tim, Gottlob Georg, Grasso Giovanni, "OXPATH: A language for scalable data extraction, automation, and crawling on the deep web", *VLDB Journal*, 2013, 22(1) pp. 47-72.
- [8] Hong Jer Lang, "Data Extraction for Deep Web Using WordNet", *IEEE Transactions on Systems Man and Cybernetics Part C-applications and Reviews*, 2011, 41(6) pp. 854-868.
- [9] Chen Kerui, Zuo Wanli, He Fengling, "Data Extraction and Annotation Based on Domain-specific Ontology Evolution for Deep Web", *Computer Science and Information Systems*, 2011, 8(3) pp. 673-692.
- [10] Liu Wei, Meng Xiaofeng, Meng Weiyi, "ViDE: A Vision-Based Approach for Deep Web Data Extraction", *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(3) pp. 447-460.
- [11] Su Jui-Yuan, Chen Lung-Pin, Wu I-Chen, "A Loosely Coupled Interactive Web Data Extraction System", *Journal of Internet Technology*, 2010, 11(2) pp. 237-249.
- [12] Kayed Mohammed, Chang Chia-Hui, "FiVaTech: Page-Level Web Data Extraction from Template Pages", *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(2) pp. 249-263.
- [13] Lin Jianfang, Li Sheng, Cai Yuhan, "Collocation Extraction Using Web Feedback Data", *Chinese Journal Of Electronics*, 2009, 18(2) pp. 312-316.
- [14] Cafarella Michael J., Madhavan Jayant, Halevy Alon, "Web-Scale Extraction of Structured Data", *Sigmod Record*, 2008, 37(4) pp. 55-61.
- [15] Zhu Jun, Nie Zaiqing, Zhang Bo, "Dynamic Hierarchical Markov Random Fields for integrated web data extraction", *Journal of Machine Learning Research*, 2008, 9 pp. 1583-1614.
- [16] Indio Valentina, Martelli Pier Luigi, Savojardo Castrense, "The prediction of organelle-targeting peptides in eukaryotic proteins with Grammatical-Restrained Hidden Conditional Random Fields", *Bioinformatics*, 2013, 29(8) pp. 981-988.
- [17] Wang Xiaofeng, Zhang Xiao-Ping, "An ICA Mixture Hidden Conditional Random Field Model for Video Event Classification", *IEEE Transactions on Circuits and Systems for Video Technology*, 2013, 23(1) pp. 46-59.
- [18] Qian X., Hou X., Tang, Y. Y., "Hidden conditional random field-based soccer video events detection", *IET Image Processing*, 2012, 6(9) pp. 1338-1347.
- [19] Bousmalis Konstantinos, Zafeiriou Stefanos, Morency Louis-Philippe, "Infinite Hidden Conditional Random Fields for Human Behavior Analysis", *IEEE Transactions on Neural Networks and Learning Systems*, 2013, 24(1) pp. 170-177.
- [20] Zhang Jianguo, Gong Shaogang, "Action categorization with modified hidden conditional random field", *Pattern Recognition*, 2010, 43(1) pp. 197-203.
- [21] Li Zhao, Ng Wee Keong, Sun Aixin, "Web data extraction based on structural similarity", *Knowledge and Information Systems*, 2005, 8(4) pp. 438-461.
- [22] Zhu Jun, Nie Zaiqing, Zhang Bo, "Dynamic Hierarchical Markov Random Fields for integrated web data extraction", *Journal of Machine Learning Research*, 2008, 9 pp. 1583-1614.
- [23] Peng Fuchun, McCallum Andrew, "Information extraction from research papers using conditional random fields", *Information Processing & Management*, 2006, 42(4) pp. 963-979.
- [24] Zhu Jun, Nie Zaiqing, Wen Ji-Rong, Zhang Bo, Ma Wei-Ying, "2D conditional random fields for web information extraction", *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 1044-1051.

Nearly Optimal Solution for Restricted Euclidean Bottleneck Steiner Tree Problem

Zimao Li* and Wenying Xiao

College of Computer Science, South-Central University for Nationalities, Wuhan City, Hubei Province, P. R. China

*Corresponding author, Email: lizm@mail.scuec.edu.cn, 1047464075@qq.com

Abstract—A variation of the traditional Steiner tree problem, the bottleneck Steiner tree problem is considered in this paper, which asks to find a Steiner tree for n terminals with at most k Steiner points such that the length of the longest edge in the tree is minimized. The problem has applications in the design of WDM optical networks, design of wireless communication networks and reconstruction of phylogenetic tree in biology. We study a restricted version of the bottleneck Steiner tree problem in the Euclidean plane which requires that only degree-2 Steiner points are possibly adjacent in the optimal solution. The problem is known to be MAX-SNP hard and cannot be approximated within $\sqrt{2}$ unless P=NP, we propose a nearly optimal randomized polynomial time approximation algorithm with performance ratio $\sqrt{2}+\epsilon$, where ϵ is a positive number.

Index Terms—Bottleneck Steiner Tree; Approximation Algorithm; Performance Ratio; Wireless Networks

I. INTRODUCTION

Given a weighted graph $G=(V,E;W)$ and a subset $S\subset V$ of required vertices (also called terminals), the traditional Steiner tree problem [1] asks a shortest acyclic network connecting S . In fact, the acyclic network is a tree and it may use additional points (also called Steiner points) in $V-S$. We call such a tree a Steiner tree. In the past 20 years, the traditional Steiner tree problem attracts considerable attention and interests from both theoretical point of view and its applicability and once occupied a central place in the emerging theory of approximation algorithms.

The problem is MAX-SNP hard even when the edge weights are only 1 or 2 [2]. For the Steiner tree problem in Euclidean plane, it is still NP-hard and there is a polynomial-time approximation scheme (PTAS) [3] for Euclidean Steiner trees, i.e., a near-optimal solution can be found in polynomial time [4].

New applications of Steiner tree problem in VLSI routing [5], wireless communications [6] and phylogenetic tree reconstruction in biology [7] have been found and studied deeply. These applications generally need to do some modification to the traditional Steiner tree problem. Therefore, the study of variations of traditional Steiner tree problem become a hot issue.

For example, recent advances in affordable and efficient electronics have had a dramatic impact on the availability and performance of radio-frequency wireless communication equipment. A number of defense and

civil applications involve deployment of computing devices or sensors able to communicate digital information through wireless connections. In most cases the sensors are battery powered and therefore operate for a limited time before they consume all power and stop working. In order to prolong the network lifetime in general, it is desirable to minimize the distance between nodes [8].

Another example, in the design of wavelength division multiplexing (WDM) optical network, suppose we need to connect the n nodes located at p_1, p_2, \dots, p_n by WDM optical network, due to transmit power limit, signal can only transmit a limited distance to ensure correct transmission. If the distance between some nodes in the connection tree is large, signal amplifiers are required to place at proper positions to shorten the connection distance.

Two examples leads us to consider minimizing the maximum edge length problem and minimizing the number of Steiner points problem, implying the two variants of the classic Steiner tree problem: the bottleneck Steiner tree problem [9] and the Steiner tree problem with minimum number of Steiner points and bounded edge-length [8, 10, 11, 13].

In this paper, we consider one related variation of the traditional Steiner tree problem, the bottleneck Steiner tree problem, which is defined as follows: given a set $P=\{p_1, p_2, \dots, p_n\}$ of n terminals and a positive integer k , we want to find a Steiner tree with at most k Steiner points such that the length of the longest edges in the tree is minimized.

The problem can be applied to extend the lifetime of a wireless network when n nodes have fixed locations and a number of up to k additional nodes can be placed at arbitrary positions. The objective is to build a spanning tree that connects the n fixed points and up to k additional nodes in the Euclidean plane, so that the length of the longest tree edge is minimized. Hence, the power required to transmit on the longest link is minimized also, and the network lifetime, in terms of connectivity, is extended.

Other applications such as design of multifacility location, VLSI routing, network routing, optical switching networks and phylogenetic tree reconstruction indicates the broad applicability of the bottleneck Steiner tree problem.

The problem is showed to be NP-hard. In [9], D.-Z Du and L. Wang proved that unless $P=NP$, the problem cannot be approximated in polynomial time within performance ratios 2 and $\sqrt{2}$ in the rectilinear plane and the Euclidean plane, respectively. Moreover, they gave an approximation algorithm with performance ratio 2 for both the rectilinear plane and the Euclidean plane. For the rectilinear plane, the performance ratio is best possible, that is, the performance ratio is tight. For the Euclidean plane, however, the gap between the lower bound $\sqrt{2}$ and upper bound 2 is still big. Based on the existence of a 3-restricted Steiner tree, we presented a randomized polynomial approximation algorithm with performance ratio $1.866 + \epsilon$, for any positive number ϵ for the Euclidean plane [12]. Later I. Cardei, M. Cardei, L. Wang, B. Xu, and D.-Z Du improved the performance ratio to $\sqrt{3} + \epsilon$, for any positive number ϵ [8, 13]. This is so far the best results possible.

In 2004, a restricted version of the problem in the Euclidean plane which requires that no edge connects any two Steiner points in the optimal solution was considered. We proved that the problem is NP-hard and cannot be approximated in polynomial time within performance ratio $\sqrt{2}$ and proposed a randomized polynomial approximation algorithm with performance ratio $\sqrt{2} + \epsilon$, for any positive number ϵ [14]. S. Bae, C. Lee, and S. Choi studied the Euclidean bottleneck Steiner tree problem when k is restricted to 1 or 2, they gave exact solutions to this problem [15]. M. Li, B. Ma and L. Wang studied the bottleneck Steiner tree problem in String space when $k = 1$ (also called the closest string problem). They proved the problem to be NP-hard and present a PTAS for it, and hence solved it perfectly in theory [16].

In this paper, we study the bottleneck Steiner tree problem in the Euclidean plane by allowing only degree-2 Steiner points are possibly adjacent in the optimal bottleneck Steiner tree. The case we consider is more general than the restricted version in [14]. We denote the problem *restricted-BST* for short. We have shown that the problem is MAX-SNP hard and cannot be approximated within performance ratio $\sqrt{2}$ and provide an $O(n \log n + k \log n)$, approximation algorithm with performance ratio $\sqrt{3}$ [17]. But there still exist a gap between the lower bound $\sqrt{2}$ and upper bound $\sqrt{3}$. In this paper, by introducing the notion of 3-restricted Steiner tree, we prove the existence of ratio $\sqrt{2}$ and propose a randomized polynomial time approximation algorithm with performance ratio $\sqrt{2} + \epsilon$, for any positive number ϵ , which is nearly optimal and almost close the problem.

In Section II, we show that the existence of 3-restricted Steiner tree with the length of the longest edge not exceeding $\sqrt{2}$ of the optimal solution. Section III provide a method to construct a weighted 3-hypergraph and a polynomial time randomized approximation algorithm with performance ratio $\sqrt{2} + \epsilon$. By introducing the binary search strategy, we also give a fast implementation. The concluding remarks appear in Section IV.

II. THE EXISTENCE OF PERFORMANCE RATIO

In this section, by introducing the notion of 3-restricted Steiner tree, we will show the existence of performance ratio $\sqrt{2}$ for the restricted-BST problem. First, the following theorem in [17] shows the hardness of the problem.

Theorem 1: Unless $P=NP$, the restricted-BST problem in the Euclidean plane cannot be approximated within performance ratio $\sqrt{2}$ in polynomial time.

For describing convenience, we need some related notions. Usually, every leaf in a Steiner tree is a terminal. However, a terminal may not be a leaf. A Steiner tree is *full* if all terminals are leaves. Thus, if a Steiner tree is not full, there must exist a terminal which is not a leaf, we can decompose the tree at this terminal into several small trees, and these small trees share a common terminal. In this way we can always decompose any Steiner tree into the union of several small trees, in each of them a vertex is a leaf if and only if it is a terminal. These small trees are called *full Steiner components*, or formally,

Definition 1: A full Steiner component of a Steiner tree is a subtree in which each terminal is a leaf and each internal node is a Steiner point.

Consequently, We can define the notion of *k-restricted Steiner tree* and in this paper we focus on the case when $k = 3$.

Definition 2: A Steiner tree for n terminals is a k -restricted Steiner tree if each of its full component spans at most k terminals.

Below notation is adopted in the proof of our main theorem-Theorem 2. Let a and b be two points in the plane, we denote ab an edge and $|ab|$ the length of ab . Without loss of generality, we assume the length of the longest edges in the optimal Steiner tree is 1.

Theorem 2: Given a set of n terminals P in the Euclidean plane, let T be an optimal bottleneck Steiner tree for the restricted-BST problem. Then, there exists a 3-restricted Steiner tree T' for P with the same number of Steiner points as T such that the length of the longest edges in T' is at most $\sqrt{2}$.

Proof: Because only degree-2 Steiner points are possibly adjacent in the optimal solution, any optimal bottleneck Steiner tree T can be decomposed into the union of its full components, each of which is either a star with a Steiner point as center (see Figure 1) or just a line-segment path connecting two terminals with $l \geq 0$ intermediate degree-2 Steiner points (See Figure 2).

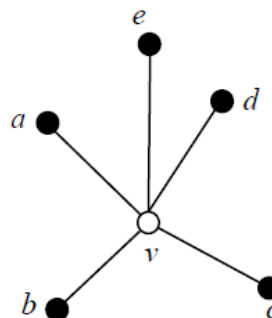


Figure 1. A star with a Steiner point

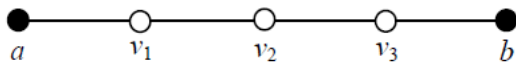


Figure 2. A line-segment path with 3 degree-2 Steiner points

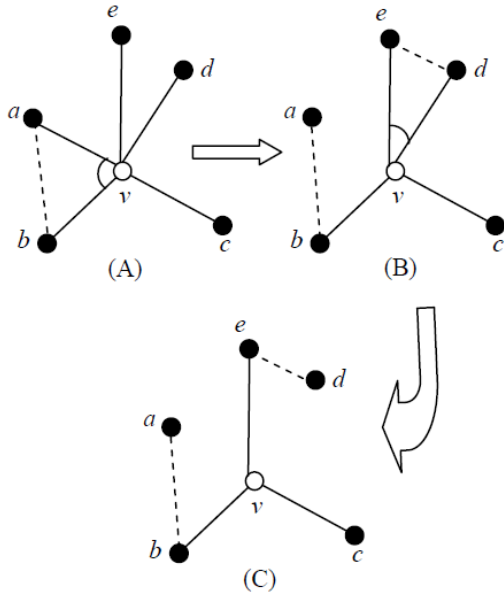


Figure 3. Transformation of a star to 3-restricted Steiner subtree

For a star T_s with at least 4 terminals, we can always decrease the degree of the Steiner point step by step to 3 and guarantee the length of the longest edges in the modified tree is at most $\sqrt{2}$. The procedure is as below:

Suppose the Steiner point is labeled as v , there must exist two terminals a and b satisfying $\angle avb \leq 90^\circ$, by directly connecting a and b and removing the longer edge of va and vb , the degree of v is decreased by 1, and it is easily seen that

$|ab| = \sqrt{|va|^2 + |vb|^2 - 2|va| \cdot |vb|\cos\angle avb} \leq \sqrt{2}$ (remember the assumption that the length of the longest edges in the optimal Steiner tree is 1). Repeat the procedure until the degree of v becomes 3. Figure 3 gives an example to illustrate the procedure.

Thus we transform the star T_s into a Steiner subtree in which the length of the longest edges is at most $\sqrt{2}$ and the number of Steiner points in the Steiner subtree does not increase. For the line-segment path like full Steiner component, no transformation work is needed because the length of its edges is at most 1. Finally we union all the Steiner subtrees to form a steinerized spanning tree T' with the same number of Steiner points as the optimal bottleneck Steiner tree T , apparently T' is a 3-restricted Steiner tree and the length of the longest edges in T' is at most $\sqrt{2}$.

A hypergraph $H=(V, F)$ is a generalization of a graph where the edge set F is an arbitrary family of subsets of vertex set V . A 3-hypergraph $H_3=(V, F)$ is a hypergraph, each of whose edges has cardinality at most 3. A weighted 3-hypergraph $H_3=(V, F; W)$ is a 3-hypergraph with each edge associated with a weight. A minimum spanning tree for a weighted 3-hypergraph $H_3=(V, F; W)$

is a subgraph T of H_3 that is a tree containing every node in V with the least weight.

The following theorem proves the existence of a randomized algorithm for computing a minimum spanning tree for a weighted 3-hypergraph [18].

Theorem 2: There exists a randomized algorithm for the minimum spanning tree problem for weighted 3-hypergraphs, with probability at least 0.5, running in $\text{poly}(n, w_{\max})$ time, where n is the number of nodes in the hypergraph and w_{\max} is the largest weight of edges in the hypergraph.

III. THE APPROXIMATION ALGORITHM

In this section, we transform the computation of an optimal 3-restricted Steiner tree into the minimum spanning tree problem for weighted 3-hypergraphs.

To construct a weighted 3-hypergraph, we need to know B , the length of the longest edges in an optimal solution. It is hard to find the exact value of B in an efficient way because of the hardness of the restricted-BST problem. However, we can guess the length of the longest edges in an optimal solution. The following procedure finds a value B' that is at most $(1+\epsilon)B$ for any $\epsilon > 0$:

Run the polynomial time approximation algorithm with performance ratio $\sqrt{3}$ in [17] to get an upper bound X of B .

Try to use one of $\frac{X}{\sqrt{3}}, \frac{X}{\sqrt{3}}(1 + \epsilon), \frac{X}{\sqrt{3}}(1 + 2\epsilon), \dots, X$ as B' , where ϵ is a positive number.

Thus, we can assume that $B'=(1+\epsilon)B$ is the approximation of the longest edges in an optimal solution. Now we can construct a weighted 3-hypergraph $H_3=(V, F; W)$ from the set P of terminals. The construction process is shown in Figure 4.

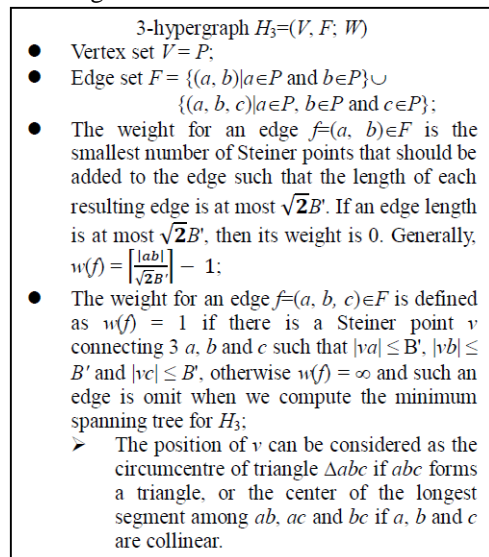


Figure 4. Construction process of a 3-hypergraph

It is easily seen that for a given bound B' , the above construction can be done in $O(n^3)$ time because the number of edges is $O(n^3)$ and the computation of the weight of an edge uses only constant time. After obtaining the weighted 3-hypergraph $H_3(V, F; W)$, we can

use the algorithm in [18] to computer a minimum spanning for H_3 . Now it is ready for us to present our approximation algorithm with performance ratio $\sqrt{2}+\varepsilon$, where ε is a positive number.

Algorithm restricted-BST(P, n, k, ε)

Input: A set P of n terminals in the Euclidean plane, an integer k and a positive number ε .

Output: A 3-restricted Steiner tree T for P with at most k Steiner points.

Call the $O(n\log n+k\log n)$ approximation algorithm with performance ratio $\sqrt{3}$ for restricted Euclidean bottleneck Steiner tree problem in [17] to get the length of the longest edge X .

For $B' \leftarrow \frac{X}{\sqrt{3}}, \frac{X}{\sqrt{3}}(1 + \varepsilon), \frac{X}{\sqrt{3}}(1 + 2\varepsilon), \dots, \frac{X}{\sqrt{3}}(1 + \lceil \frac{\sqrt{3}-1}{\varepsilon} \rceil \times \varepsilon)$ **do**

Construct a weighted 3-hypergraph $H_3(V, F; W)$ according to B' and Figure 4.

Call the polynomial randomized algorithm in [18] to compute a minimum spanning tree T' for $H_3(V, F; W)$.

if $w(T') \leq k$ **then** exit the for loop.

Replace every edge f of the minimum spanning tree T' on $H_3(V, F; W)$ with a Steiner subtree as below descriptions.

If $f = (a, b)$, replace f with a path connecting a and b by adding $w(f)$ intermediate Steiner points with a even partition of f .

If $f = (a, b, c)$, replace f with a star centered at the circumcenter triangle Δabc if abc forms a triangle, or at the center of the longest segment among ab, ac and bc if a, b and c are collinear.

Output the resulting 3-restricted Steiner tree.

Theorem 1 and Theorem 2 indicates the existence and performance of a randomized approximation algorithm for the restricted Euclidean bottleneck Steiner tree problem. Combined with algorithm restricted-BST, we have the following Theorem.

Theorem 3: For any given ε , there exists a randomized algorithm that computes a Steiner tree with n terminals and k Steiner points with probability at least 0.5 such that the longest edge in the tree is at most $\sqrt{2}+\varepsilon$ times of the optimum, and the algorithm's running is $\frac{1}{\varepsilon} \times poly(n, k)$.

In fact, by using a binary search strategy, we can decrease the number of loops in Step 2 from $\frac{1}{\varepsilon}$ to $\log(\frac{1}{\varepsilon})$ and hence improve Algorithm restricted-BST(P, n, k, ε). Algorithm faster-restricted-BSP(P, n, k, ε) is an improvement of Algorithm restricted-BST(P, n, k, ε).

Algorithm faster-restricted-BST(P, n, k, ε)

Input: A set P of n terminals in the Euclidean plane, an integer k and a positive number ε .

Output: A 3-restricted Steiner tree T for P with at most k Steiner points.

Call the $O(n\log n+k\log n)$ approximation algorithm with performance ratio $\sqrt{3}$ for restricted Euclidean bottleneck Steiner tree problem in [17] to get the length of the longest edge X .

Initialize $low \leftarrow 0$ and $high \leftarrow \lceil \frac{\sqrt{3}-1}{\varepsilon} \rceil$

while ($low < high$) **do**

$mid \leftarrow (low+high)/2$ and $B' \leftarrow \frac{X}{\sqrt{3}}(1 + mid \times \varepsilon)$

Construct a weighted 3-hypergraph $H_3(V, F; W)$ according to B' and Figure 4.

Call the polynomial randomized algorithm in [18] to compute a minimum spanning tree T for $H_3(V, F; W)$.

Consider the solution T , **if** $w(T) > k$, **then** $low \leftarrow mid+1$; **else** $high \leftarrow mid$.

Replace every edge f of the minimum spanning tree T on $H_3(V, F; W)$ with a Steiner subtree as below descriptions.

If $f = (a, b)$, replace f with a path connecting a and b by adding $w(f)$ intermediate Steiner points with a even partition of f .

If $f = (a, b, c)$, replace f with a star centered at the circumcenter triangle Δabc if abc forms a triangle, or at the center of the longest segment among ab, ac and bc if a, b and c are collinear.

Output the resulting 3-restricted Steiner tree.

IV. CONCLUSION

We mainly considered a restricted version of the bottleneck Steiner tree problem in the Euclidean plane. The problem is MSX-SNP hard and cannot be approximated with ratio $\sqrt{2}$ unless $P=NP$. In this paper we presented a polynomial time randomized approximation algorithm with performance ratio $\sqrt{2}+\varepsilon$. The algorithm is near optimal and almost close the gap between lower bound $\sqrt{2}$ and upper bound $\sqrt{2}+\varepsilon$.

Further study include the derandomization of the randomized algorithm efficiently.

As an application, the algorithm can be used to improve the lifetime of wireless networks by minimizing the length of the longest edge in the interconnecting tree by deploying additional relay nodes at specific locations.

ACKNOWLEDGMENT

The author wish to thank Rongbo Zhu for his double-check and suggestions on this paper. This work was fully supported by National Natural Science Foundation of China (Project 61103248 and 61379059).

REFERENCES

- [1] M. R. Garey, R. L. Graham and D. S. Johnson, "The Complexity of Computing Steiner Minimal Trees," *SIAM Journal on Applied Mathematics*, vol. 32, pp. 835-859, 1977.
- [2] M. Bern and P. Plassmann, "The Steiner Problem with Edge Lengths 1 and 2," *Information Processing Letters*, vol. 32, pp. 171-176, 1989.
- [3] M. R. Garey and D. S. Johnson, "Computers and Intractability, A Guide to the Theory of NP-Completeness," W. H. Freeman and Company, New York, 1979.
- [4] S. Arora, "Polynomial Time Approximation Scheme for Euclidean TSP and Other Geometric Problems," *Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, Burlington VT, pp. 2-11, Oct. 1996.
- [5] A. Kahng and G. Robins, "On Optimal Interconnections for VLSI," Kluwer Publishers, 1995.
- [6] A. Caldwell, A. Kahng, S. Mantik, I. Markov and A. Zelikovsky, "On Wirelength Estimations for Row-based Placement," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol 18, pp. 1265-1278, 1999.
- [7] F. K. Hwang, D. S. Richards and P. Winter, "The Steiner Tree Problem," North-Holland, 1992.
- [8] I. Cardei, M. Cardei, L. Wang, B. Xu, D. -Z. Du, "Optimal relay location for resource-limited energy-efficient wireless communication," *Journal of Global Optimization*, vol. 36, pp. 391-399, 2006.
- [9] L. Wang and D. -Z Du, "Approximations for a Bottleneck Steiner Tree Problem," *Algorithmica*, vol. 32, pp. 554-561, 2002.
- [10] M. Sarrafzadeh and C. Wong, "Bottleneck Steiner Trees in the Plane," *IEEE Transactions on Computers*, vol. 41, pp. 370-374, 1992.
- [11] G. Lin and G. Xue, "Steiner Tree Problem with Minimal Number of Steiner Points and Bounded Edge-length," *Information Processing Letters*, vol. 69, pp. 53-57, 1999.

- [12] L. Wang and Z. Li, "An approximation algorithm for a bottleneck k -Steiner tree problem in the Euclidean plane", *Information Processing Letters*, vol. 81, pp. 151-156, 2002.
- [13] D. -Z Du, L. Wang and B. Xu, "The Euclidean Bottleneck Steiner Tree and Steiner Tree with Minimum Number of Steiner Points," in *Proceedings of the 7th Annual International Conference on Computing and Combinatorics*, Guilin, China, LNCS vol. 2108, pp. 509-518, August 2001.
- [14] Z. Li, D. Zhu and S. Ma, "Approximation algorithm for bottleneck Steiner tree problem in the Euclidean plane," *Journal of Computer Science and Technology*, vol. 19, pp. 791-794, 2004.
- [15] S. Bae, C. Lee, and S. Choi, "On Exact Solutions to the Euclidean Bottleneck Steiner Tree Problem," in *Proceedings of the Third Annual Workshop on Algorithms and Computation*, Kolkata, India, LNCS vol. 5431, pp. 105-116, Feb. 2009.
- [16] M. Li, B. Ma and L. Wang, "On the Closest String and Substring Problems," *Journal of the ACM*, vol. 49, pp. 157-171, 2002.
- [17] Z. Li and W. Xiao, "Fast Approximation Algorithm for Restricted Euclidean Bottleneck Steiner Tree Problem," unpublished.
- [18] H. J. Prömel and A Steger, "A new approximation algorithm for the steiner tree problem with performance ratio $5/3$," *Journal of Algorithms*. vol. 36, pp. 89-101, 2000.



Zimao Li, Ph.D., associate professor and vice dean, born in Linqing, P.R, China, 1974, received his Bachelor's degree in Mathematics, Master's degree in Computer Science from Shandong University in 1996 and 1999, respectively, and his Ph.D. degree in Computer Science from City University of Hong Kong in 2002. His research

interests include computational complexity, approximation algorithms and design and analysis of algorithms.

From 2002 to 2005, he was a lecturer of College of Computer Science and Technology, Shandong University, P.R. China; from 2005, he is an associate professor of College of Computer Science, South-Central University for Nationalities, P.R. China; from 2012, he is the vice dean of College of Computer Science, South-Central University for Nationalities. He has been supported by 1 national fund and 3 provincial level funds and published more than 10 research papers in journals such as *SIAM Journal on Computing*, *IEEE Transactions on SMC*, *Journal of Computer Science and Technology*, *Information Processing Letters*, etc.



Wenying, Xiao, born in Anguo, P.R. China, 1988, received his Bachelor's degree in Computer Science and Technology from South-Central University for Nationalities in 2012, currently he is a Master degree candidate. His research interest is design and analysis of algorithms.

Computer Crime Forensics Based on Improved Decision Tree Algorithm

Ying Wang¹, Xinguang Peng¹, and Jing Bian^{1,2}

1. College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China

2. The Center of Information and Network, Shanxi Medical College of Continuing Education, Taiyuan, China

Email: ablewy@163.com, sxgrant@126.com, bianjing@126.com

Abstract—To find out the evidence of crime-related evidence and association rules among massive data, the classic decision tree algorithms such as ID3 for classification analysis have appeared in related prototype systems. So how to make it more suitable for computer forensics in variable environments becomes a hot issue. When selecting classification attributes, ID3 relies on computation of information entropy. Then the attributes owning more value are selected as classification nodes of the decision tree. Such classification is unrealistic under many cases. During the process of ID3 algorithm there are too many logarithms, so it is complicated to handle with the dataset which has various classification attributes. Therefore, contraposing the special demand for computer crime forensics, ID3 algorithm is improved and a novel classification attribute selection method based on Maclaurin-Priority Value First method is proposed. It adopts the foot changing formula and infinitesimal substitution to simplify the logarithms in ID3. For the errors generated in this process, an apposite constant is introduced to be multiplied by the simplified formulas for compensation. The idea of Priority Value First is introduced to solve the problems of value deviation. The performance of improved method is strictly proved in theory. Finally, the experiments verify that our scheme has advantage in computation time and classification accuracy, compared to ID3 and two existing algorithms.

Index Terms—ID3; Classification Attribute; Information Gain; Priority Value First; Decision Tree

I. INTRODUCTION

Recent years the experts and scholars at home and abroad are emphasizing on the problems of computer crime forensics. But the researches are limited in key word finding, pattern matching, analysis on file attributes etc [1]. The forensics process needs a lot of labor involved. It can not provide prediction for possible and potential computer crime, and it lacks the mining ability for data hidden information and crime mode [2, 3]. Therefore, people need to use other tools to get useful evidence from amounts of data or to acquire useful information for data analysis and acquisition of next time. Data mining is the most suitable tool and it can find out valuable knowledge and information from amounts of data, which has various types of patterns. At present, there are mainly two aspects in the field of computer crime forensics analysis [4, 5]: The first is using data mining technology to perform correlation analysis on

computer logs, firewall logs, intrusion detection logs and route access records. Then, the crime trace can be excavated and the crime process can also be restored. The other is excavating and analyzing amounts of computer crime data, to obtain disciplines and features from crime forensics. Based on the correlation between these features and different crime forensics, the computer crime forensics can be predicted. It can also provide clues and evidence for polices to resolve cases and to prevent crimes.

Association rule is a kind of mature method in data mining [6] and it can excavate mutually related methods between valuable data items from amounts of data. The association rule can be used to analyze users' forensics and to compare the incidents that may appear together. Users' forensics are studied to discover their inherent rules. There are many methods which can construct study models in classification analysis. Among them, the decision tree represented by ID3 algorithm is widely applied. For prototype system of computer forensics, ID3 is used generally. However, due to its generality and uniqueness for forensics data, it cannot be effective to excavate reasonable models. Many domestic and overseas researchers have optimized and extended this algorithm like C4.5 and CART, etc [7, 8]. Domestic scholars Chen Xiangtao, Qu Kaishe, etc [9] individually introduce approximate calculation and weighted value to optimize algorithm ID3 in different extent. However, approximate calculation can not make up the deviation by this optimization and the classification accuracy is not ideal. Although the latter solved the problem of value deviation partly, its calculation is so complicated and the generation time of decision tree is too long. [10] discovers that it needs equational scanning and ranking on data set many times, during the process to create decision tree. So it will lead to inefficiency of the algorithm. Meanwhile, ID3 algorithm takes the highest information gain as the standard of selection attribute. After analysis, this standard is inclined to select the attribute which has more attributes.

The computation of information entropy relies on the attribute owing many values. That is, the selection to judge attribute has deviation problem. During the data analysis on computer forensics, most data needs discretization such as the number of sending byte, IP address of sending end, etc. The attribute values after

discretization are related to discretization standard. So it decides that the decision attribute selection depends on discretization standard instead of data, as is not reasonable.

We will analyze and study the two improved schemes for ID3 algorithm. One of them is the project based on the idea of “approximate to simplify” referred in [11]. This improved method makes the computation of ID3 more convenient to some extent and it reduces the generation time of the decision tree. The other one introduces the prior knowledge to overcome value deviation in ID3 [12]. This scheme solves the problem of value deviation to some extent. However, this method adds multiply calculation to ID3 and it promotes the calculation of expected entropy to be more complex [13]. So it is slower than algorithm ID3 in classification speed. On the basis of the first improved algorithm, since it does not make up errors during equivalent substitution, it is not as good as ID3 on classification accuracy and it will affect the prediction accuracy to certain extent. We can conclude that this scheme can reduce the generation time of decision tree at the expense of classification accuracy. Although an empirical value is introduced in the second improved algorithm to overcome the defects of value deviation in ID3, when selecting classification attribute. So this method is still very complicated when calculating the expected entropy and it even needs more time of classification. Aiming at the features of computer crime forensics data, the current algorithms are improved from two aspects: weight value and twice information gain. The research mainly includes:

(1) This paper analyzes advantages and disadvantages of algorithm ID3 in detail. Through the analysis on ID3 and the features of these two improved schemes, we provide precise theoretical proof to its defects.

(2) A novel selection method of selected classification attribute is proposed, that is, the ID3 algorithm will be simplified with logarithm based on McLaughlin’s idea.

(3) For the problem of value deviation in ID3, this paper introduces the idea of “Priority Value First” [14] for attribute classification, which can process data set with many classification attribute values.

(4) Through the experiments, the improved algorithm is proved to have better performance in classification accuracy than ID3 and existing improved schemes. Meanwhile, it effectively overcomes the value deviation of the first improved method.

II. RELATED WORKS

A. Principle Algorithm

S is a set including s data samples. We assume the attribute of class label has m different value and define m different classes $C_i (i=1,2,\dots,m)$. Let s_i is the samples number in classes C_i . The expected information [15] for a given sample category is

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^m p_i \log_2(p) \quad (1)$$

$p_i = s_i / s$ is probability of any sample belonging to C_i . Since the information adopts binary coding the base of logarithmic function is 2.

Set attribute A has v different value. Then S can be divided into v subsets $\{S_1, S_2, \dots, S_v\}$ by A . Samples of S_j have the same value $a_j (j=1,2,\dots,v)$ in A . Set S_{ij} as the samples number of C_i in subset S_j . The expected information or entropy of the subsets divided by A is

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + s_{2j} + \dots + s_{mj}}{s} I(s_{1j}, s_{2j}, \dots, s_{mj}) \quad (2)$$

For give subset S_j , its expected information is

$$I(s_{1j}, s_{2j}, \dots, s_{mj}) = -\sum_{i=1}^m p_{ij} \log_2(p_{ij}) \quad (3)$$

$p_{ij} = s_{ij} / s_j$ is the probability of samples in S_j belonging to C_i . The information gain obtained by branches of A is

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A) \quad (4)$$

Decision tree algorithms compute the information gain of each attribute. They will select the biggest one as testing attribute of given set S and generate corresponding branch node. All the generated nodes are marked corresponding attribute. According to different value of this attribute the branches of decision tree are generated. Each branch represents a divided sample subset, to separate the samples.

B. Theoretical Analysis of the Defects

By formula 2 we can see that the expected is smaller if value of some attribute is more. Then the probability to be selected as classification attribute will be larger. So ID3 tends to select the attribute owning more values. We assume σ is an attribute in some training dataset and its value is $\sigma_1, \sigma_2, \dots, \sigma_n$. σ' is an extended attribute of σ . It means σ is divided into two value and the other factors are exactly consistent. So we can neglect other factors when comparing $gain(\sigma)$ and $gain(\sigma')$, which satisfies the uniqueness of experiment. The value of σ' is expressed by $\sigma'_1, \sigma'_2, \dots, \sigma'_n$. ID3 is used to get the information gain of σ and σ' . If $gain(\sigma') \geq gain(\sigma)$ is always tenable, then it demonstrate that there is defect in value deviation of ID3; else the defect does not exist. We assume the amount of classification attribute value is k . We get following formulas as formula 2:

$$E(\sigma) = \sum_{j=1}^n p(\sigma_j) I(\sigma_j) = -\sum_{j=1}^n p(\sigma_j) \sum_{i=1}^k [p(c_i / \sigma_j) \log_2(p(c_i / \sigma_j))] \quad (5)$$

$$E(\sigma') = \sum_{j=1}^{n+1} p(\sigma'_j) I(\sigma'_j) = -\sum_{j=1}^{n+1} p(\sigma'_j) \sum_{i=1}^k [p(c_i / \sigma'_j) \log_2(p(c_i / \sigma'_j))] \quad (6)$$

\therefore The attribute value of σ' is decomposed by σ

$$\begin{aligned} &\therefore gain(\sigma') - gain(\sigma) = E(\sigma') - E(\sigma) \\ &= p(\sigma'_n) \sum_{i=1}^k [p(c_i / \sigma'_n) \log_2(p(c_i / \sigma'_n))] \\ &+ p(\sigma'_{n+1}) \sum_{i=1}^k [p(c_i / \sigma'_{n+1}) \log_2(p(c_i / \sigma'_{n+1}))] \\ &- p(\sigma_n) \sum_{i=1}^k [p(c_i / \sigma_n) \log_2(p(c_i / \sigma_n))] \end{aligned} \tag{7}$$

Only the case of two decomposed value is discussed in this paper so $k = 2$. To simply formula 7 we use the following substitution:

$$\text{Let } \begin{cases} r = \frac{p(\sigma'_n)}{p(\sigma_n)} \\ x = p(c_1 / \sigma'_n) \\ p = p(c_1 / \sigma_n) \\ q = p(c_1 / \sigma_{n+1}) \end{cases}, \therefore \begin{cases} 1-r = \frac{p(\sigma'_{n+1})}{p(\sigma_n)} \\ 1-x = p(c_2 / \sigma_n) \\ x = rq + (1-r)q \end{cases} \text{ is}$$

tenable. Then

$$\begin{aligned} &[gain(\sigma') - gain(\sigma)] / p(\sigma_n) \\ &= r(p \log_2 p + (1-p)) + (1-r)(q \log_2 q + \\ &(1-q) \log_2(1-q)) - (x \log_2 x + (1-x) \log_2(1-x)) \end{aligned} \tag{8}$$

We use function derivation to prove whether $gain(\sigma') \geq gain(\sigma)$ is tenable. Let $f(x) = x \log_2 x - \log_2(1-x)$, then

$$\begin{aligned} &[gain(\sigma') - gain(\sigma)] / p(\sigma_n) \\ &= -f(x) + rf(p) + (1-r)f(q) \end{aligned} \tag{9}$$

$$\begin{aligned} &\therefore f'(x) = \log_2 x - \log_2(1-x) \\ &\therefore f(x) = f(rp + (1-r)q) \leq rf(p) + (1-r)f(q) \\ &\therefore f''(x) = 1/x + 1/(1-x) > 0, x \in (0,1) \\ &\therefore f(x) = f(rp + (1-r)q) \leq rf(p) + (1-r)f(q) \\ &\therefore [gain(\sigma') - gain(\sigma)] / p(\sigma_n) \geq 0 \\ &\therefore gain(\sigma') - gain(\sigma) \geq 0 \end{aligned}$$

That means when ID3 selects classification attributes the attribute owing more values tends to be selected.

III. IMPROVED METHOD

For convenient description we select a small amount of data samples shown in table 1. It includes part of the project content registered by the criminals. We use decision tree for knowledge mining. E_c =Extent of Crime, E_s =Economy Status, C_e =Extent of education, L_c = Legitimate Career and C_r =Criminal Record.

A. Improvment based on The Idea of Simplification

We introduce the formula $\ln(1+x) \approx x$ and simplify formula 2 as

$$\begin{aligned} E'(A) &= \sum_{i=1}^q \frac{p_i + n_i}{p+n} I(p_i, n_i) \\ &= \sum_{i=1}^q \frac{p_i + n_i}{p+n} \left(-\frac{p_i}{p_i + n_i} \log_2 \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} \log_2 \frac{n_i}{p_i + n_i} \right) \\ &= \frac{1}{(n+p) \ln 2} \sum_{i=1}^q \left(-p_i \ln \frac{p_i}{p_i + n_i} - n_i \ln \frac{n_i}{p_i + n_i} \right) \\ &= \frac{1}{(n+p) \ln 2} \sum_{i=1}^q \left[-p_i \left(1 - \frac{n_i}{p_i + n_i} \right) - n_i \ln \frac{p_i}{p_i + n_i} \right] \\ &\approx \frac{2}{(n+p) \ln 2} \sum_{i=1}^q \frac{n_i p_i}{p_i + n_i} \end{aligned}$$

For a given training set, the value of $\frac{2}{(n+p) \ln 2}$ is

fixed. So we only need to compute $\sum_{i=1}^q \frac{n_i p_i}{p_i + n_i}$. In this kind of improved method, the standard for selecting classification attributer is shown as

$$E'(A) = \sum_{i=1}^q \frac{n_i p_i}{p_i + n_i} \tag{10}$$

TABLE I. SAMPLES OF CRIME DATA

E_c	E_s	C_e	L_c	C_r
Low	Better	low	No	N
Low	Better	low	Yes	N
High	Better	low	No	P
Higher	Good	low	No	P
Higher	Bad	normal	No	P
Higher	Bad	normal	Yes	N
Low	Bad	normal	Yes	P
High	Good	Low	No	N
Low	Bad	Normal	No	P
Low	Good	normal	No	P
Higher	Good	Low	Yes	P
Low	Good	Low	Yes	N
High	Better	normal	No	No
Higher	Good	Low	Yes	Yes

The idea is similar to ID3 algorithm. For the attributes in the dataset, the attribute owing minimum $E'(A)$ is taken as classification attribute. We can analyze the improved performance in the time to generate decision tree and the accuracy of classification, setting the dataset in table 1 as an example. In formula 10 there are only multiplications and additions and there is not logarithm. So the computation process is relative easy and the time to generate decision tree is shorter than ID3. According to formula 10, we get the training dataset in table 1:

$$\begin{aligned} E'(E_c) &= \frac{2 \times 3}{2+3} + 0 + \frac{2 \times 3}{2+3} = 2.4 \\ E'(E_s) &= \frac{2 \times 2}{2+2} + \frac{2 \times 4}{2+4} + \frac{1 \times 3}{1+3} = 3.08 \\ E'(C_e) &= \frac{3 \times 3}{3+3} + \frac{2 \times 6}{2+6} = 3 \\ E'(L_c) &= \frac{3 \times 3}{3+3} + \frac{2 \times 6}{2+6} = 3 \end{aligned}$$

So $E'(E_c) < E'(C_e) = E'(L_c) < E'(E_s)$ is tenable. Therefore, according the standard for selecting

classification attributes we choose E_c as the root node of decision tree. Similarly $E'(A)$ of other attribute can be computed to generate the decision tree as figure 1. From this figure we can visually see that the leaf nodes of improved method are more than those generated by ID3.

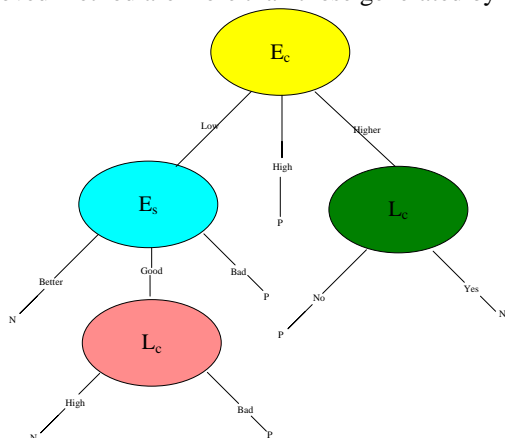


Figure 1. Decision tree generated in the 1st scheme

B. Interest Factor

This method is proposed for the problem in value deviation of ID3. We introduce the concept of “interest” [16] to rewrite the formula for expected entropy into following formula:

$$E(A) = \sum_{i=1}^q \left(\frac{p_i + n_i}{p + n} + \phi \right) I(p_i, n_i) \quad (11)$$

ϕ is attribute weight determined by experts in data mining and its calculated by above formula. Though this kind of classification needs more professional knowledge, it overcomes the value deviation of ID3 effectively. It has equivalent classification accuracy with ID3 and it is acknowledged by many professionals of the fields. But the defect in this scheme is: It makes ID3 algorithm more complicated for computation. If we set the interest of E_c as 0.33, we get

$$E(E_c) = \left(\frac{5}{14} + 0.33 \right) \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) + 0 + \left(\frac{5}{14} + 0.33 \right) \times \left(-\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \right) \approx 1.34$$

$$E(E_s) = \frac{4}{14} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) + \frac{6}{14} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) + \frac{4}{14} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) \approx 0.90$$

$$E(C_e) = \frac{7}{14} \left(-\frac{4}{7} \log_2 \frac{4}{7} - \frac{3}{7} \log_2 \frac{3}{7} \right) + \frac{7}{14} \left(-\frac{1}{7} \log_2 \frac{1}{7} - \frac{6}{7} \log_2 \frac{6}{7} \right) \approx 0.79$$

$$E(L_c) = \frac{8}{14} \left(-\frac{6}{8} \log_2 \frac{2}{7} - \frac{2}{8} \log_2 \frac{2}{8} \right) + \frac{6}{14} \left(-\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6} \right) \approx 0.87$$

So $E(C_e) < E(L_c) < E(E_s) < E(E_c)$ is tenable and we choose C_e as the root node of the decision tree. Similarly we can acquire the other classification nodes. The decision tree, time and accuracy for classification are shown in figure 2 and 3. From figure 2 it can be seen that the improved method overcome value deviation in ID3. Since multiplication is added this algorithm, its speed of classification is slower than ID3, by the results in figure 3.

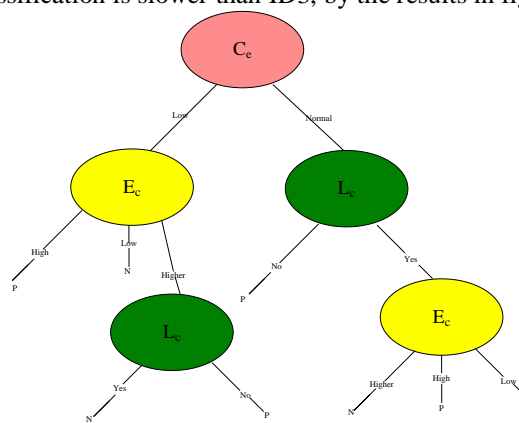


Figure 2. Decision tree generated by in the 2nd scheme

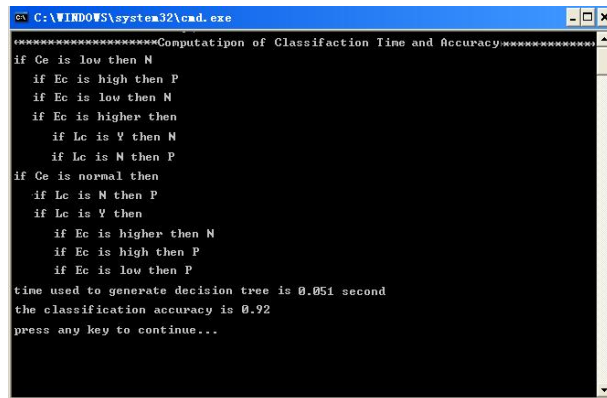


Figure 3. Classification time and accuracy

C. PVF-ID3 Algorithm

In the first optimized algorithm, the errors caused by equivalent substitution process are not compensated. So its classification accuracy is worse than ID3 and that will influence the prediction accuracy to some extent. The scheme shortens the time to generate decision at the price of reducing classification accuracy. But in many fields a little prediction error may cause huge data losses. In the second optimized algorithm, an experienced parameter is introduced. It overcomes the defects when ID3 selects the classification attribute. But the method to compute expected entropy is still complicated and it costs more classification time than ID3. Under most case, we do not concern about all the attributes in a large of dataset very much. So we expect to take the concerned attributes as

classification attributes. Then a constant $0 \leq \psi \leq 1$ can be set as the Priority Value referred in [19]. The formula to compute the expected entropy is changed as

$$E^*(A) = \sum_{i=1}^q \left(\frac{n_i + p_i}{P+N} + \psi \right) I(p_i, n_i) \quad (12)$$

$$\therefore \log_2 \lambda = \frac{\ln \lambda}{\ln 2}$$

$$\begin{aligned} \therefore E^*(A) &= \sum_{i=1}^q \frac{\ln 2}{(P+N)} \left(-p_i \ln \frac{p_i}{p_i + n_i} - n_i \ln \frac{n_i}{p_i + n_i} \right) \\ &+ \sum_{i=1}^q \psi \frac{1}{\ln 2} \left(-\frac{p_i}{p_i + n_i} \ln \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} \ln \frac{n_i}{p_i + n_i} \right) \end{aligned}$$

For the given training samples set, $(P+N)\ln 2$ is a fixed value. When computing the information entropy of each attribute it will be computed. So that can be neglected during the comparison. According to Taylor and Maclaurin formula we can simplify above formula as:

$$e(A) = \sum_{i=1}^q \left(1 + \frac{\psi}{n_i + p_i} \right) \frac{n_i p_i}{n_i + p_i} \quad (13)$$

The process of simplification will cause error, as we know from the principle. Thus, the above formula can not be used as a standard for selecting classification. We assume the amount of value for attribute A is θ . Then the computation for attribute selection proposed by us is acquired by the product of θ and $e(A)$.

$$e^*(A) = \theta e(A) \quad (14)$$

When s_i is null, ID3 algorithm will mark corresponding leaf nodes as the category which contains the maximum value of the samples [17]. To make the decision tree contain leaf nodes as few as possible, the improved algorithm will skip the intermediate procedure of ID3. It continues to find other sample subset which is not null as the next input sample set, generating corresponding branches. When the category can not be determined, to avoid incorrect classification, we prefer to feedback to deciders the un-classified information. So the deciders can perform prediction by other way.

ID3 can process the dataset which has only two kinds of attribute. PVF-ID3 algorithm can process the dataset that has multiple categories attribute values. If the dataset to be test has 3 categories attribute values: C_1, C_2, C_3 . The number of tuples is r_1, r_2, r_3 . Then the following formula can compute the expected information.

$$\begin{aligned} I(r_1, r_2, r_3) &= -\frac{r_1}{r_1 + r_2 + r_3} \log_2 \frac{r_1}{r_1 + r_2 + r_3} - \frac{r_2}{r_1 + r_2 + r_3} \log_2 \frac{r_2}{r_1 + r_2 + r_3} \\ &- \frac{r_3}{r_1 + r_2 + r_3} \log_2 \frac{r_3}{r_1 + r_2 + r_3} \end{aligned} \quad (15)$$

The improved decision tree algorithm is described as the pseudocodes below. When the attribute T of given dataset D has ζ different value, marked as $\{T_1, T_2, \dots, T_\zeta\}$. Then T can divide D into ζ subset as

$\{D_1, D_2, \dots, D_\zeta\}$. If D_{ij} is the samples number of D_i which has category attribute C_j , the needed information entropy for A to divide D can be computed as formula 15.

```

Algorithm Decision_Tree(samples,attribute_list)
Create node N;
If samples belong to the same class C
Then return N as leaf node and mark C ;
If attribute_list is null then
return N as leaf node and mark the it as the class
which has the maximum amount in samples
Mark node N as class_attribute
For known  $\alpha_i$  in each class_attribute
a branch satisfying condition class_attribute= $\alpha_i$  is
generated by N
Assume  $s_i$  is a sample set divided by
class_attribute= $\alpha_i$ 
If  $s_i$  =Null then continue to find the optimum
category attribute
Add a node returned by
Decision_Tree( $s_i$ ,samples,attribute_list)
}
    
```

$$E(A) = \sum_{i=1}^{\zeta} \left(\frac{D_{1j} + D_{2j} + \dots + D_{\zeta j}}{D} \right) I(D_{1j}, D_{2j}, \dots, D_{\zeta j}) \quad (15)$$

$$I(D_{1j}, D_{2j}, \dots, D_{\zeta j}) = -\sum_{i=1}^{\zeta} \frac{D_{ij}}{|D_i|} \log \left(\frac{D_{ij}}{|D_i|} \right)$$

TABLE II. TIME FOR CREATING DECISION TREE OF ID3 AND PVF-ID3

Record number	ID3	PVF-ID3	Saving time	Saving time rate
14	562.1	561.6	0.8	0.14
100	573	572	1	0.18
500	633	631	2	0.32
1000	689	687	2	0.29
2000	821	817	3	0.37
5000	1205	1201	4	0.33
8000	1589	1585	6	0.31
10000	1984	1970	14	0.66
15000	2783	2766	17	0.71

IV. THEORETICAL ANALYSIS AND EXPERIMENTAL VERIFICATION

A. Time Overhead Analysis

To prove that our method has higher construction efficiency, we use the dataset of different scales to test ID3 and PVF-ID3 in 12 times of computation. The average value of two algorithms is taken as the computing time to create the decision tree. Then through the experiments data we can analyze the time consumed by ID3 and PVF-ID3. The time saving rate is calculated by the ratio of the difference of average time of two algorithms and that of ID3 algorithm. From formula 13 we can see the optimized formula don not contain logarithm, only including multiplication, division and addition. In actual computation, this will be simpler the formula in ID3 and it takes less time to create the decision tree.

TABLE III. COMPARISON OF CLASSIFICATION RESULTS

Amount	Time for classification				Classification accuracy			
	ID3	After 1 st improvement	After 2 nd improvement	PVF-ID3	ID3	After 1 st improvement	After 2 nd improvement	PVF-ID3
351	0.0476	0.0412	0.0576	0.0431	0.80	0.76	0.79	0.87
625	0.0567	0.0498	0.0649	0.0512	0.84	0.79	0.81	0.89
768	0.0803	0.0714	0.0897	0.0743	0.88	0.81	0.87	0.90
1000	0.0989	0.0815	0.1076	0.0816	0.90	0.83	0.89	0.97

In table 2 we can see that in different scales of dataset, the time overhead of PVF-ID3 mainly comes from the computation of priori knowledge parameter. But with the increase of record data, that can be made up for with the rapid growth of decision tree. So it fully demonstrates that the improved method can create decision tree with higher efficiency.

B. Improvement on Value Deviation

We select the prior value of classification attribute of E_c as 0.35 and other prior value of classification attribute as 0. Due to formula 14 we can obtain the information entropy of each attribute in Table 1:

$$e^*(E_c) = [(1 + \frac{0.35}{5}) \times \frac{6}{5}] + (1 + \frac{0.35}{4}) \times \frac{0}{4} + (1 + \frac{0.35}{5}) \times \frac{6}{5} \times 3 = 7.704$$

$$e^*(E_s) = (\frac{4}{4} + \frac{8}{6} + \frac{3}{4}) \times 3 = 9.249$$

$$e^*(C_e) = (\frac{12}{7} + \frac{6}{7}) \times 2 = 5.142$$

$$e^*(L_c) = (\frac{12}{8} + \frac{9}{6}) \times 2 = 6$$

So $e^*(C_e) < e^*(L_c) < e^*(E_c) < e^*(E_s)$ is tenable. We select C_e as the root node of decision tree for classification. Similarly the others nodes can be acquired and the created decision tree is shown as figure 4. By the comparison of figure 1 and 4, we can visually find that PVF-ID3 has overcome the value deviation effectively.

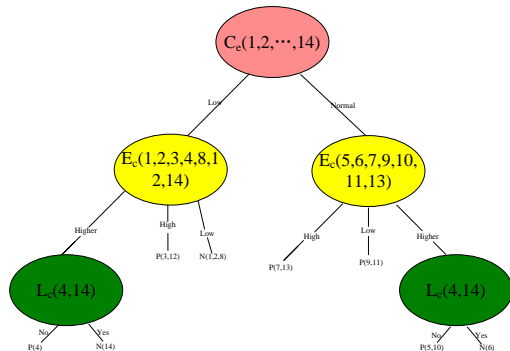


Figure 4. Decision tree created by PVF-ID3

C. Analysis on Accuracy and Multi-class Value Dataset

Though PVF-ID3 has simplified the computation in ID3, the simplification approximate substitution will cause errors, as is known from the mathematical theory. When computing the information entropy of some

attribute, we can make up these errors in the simplifying process, by introducing the multiplication of simplified formula and the amount of the values. Such improvement will increase the classification of ID3. As is described in above sectors, this advantage can be enlarged with the increasing scale of the dataset. The optimized information entropy computation formula is also applicable for the dataset owing multi-class values. We choose four kinds of different dataset in UCI [18]: ionosphere, balance-scale, diabetes and credit-g. They are tested in classification time and accuracy for PVF-ID3. The class attribute values of these dataset are more than one. The comparison results are shown as table 3 and figure 5, 6.

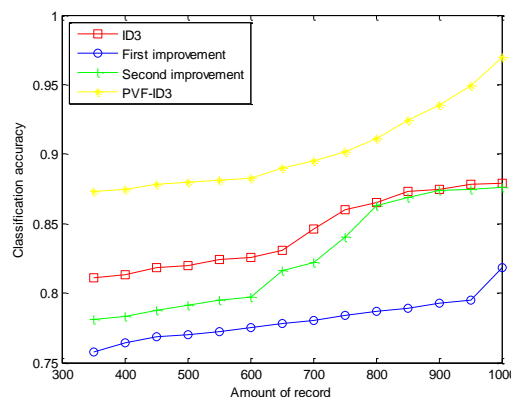


Figure 5. Comparison of classification accuracy

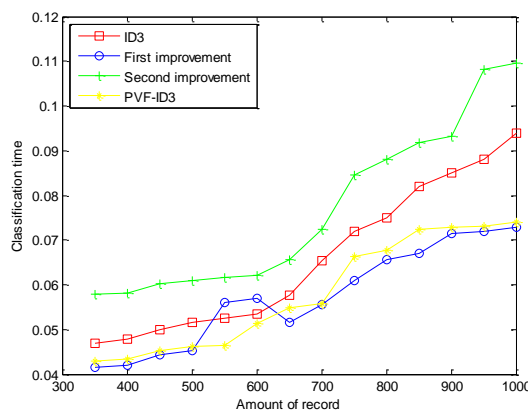


Figure 6. Comparison of classification time

By the theoretical and experimental analysis above, we can conclude that the classification time of PVF-ID3 is less than that of ID3 algorithm and the 2nd improved scheme. The classification accuracy is higher than all the methods mentioned above. It can also process the dataset which has multi-class values.

V. CONCLUSION

With the development of computer technology and popularity of Internet, the types of computer crime are constantly changing and the computer criminal characters are also different. This paper has introduced the disadvantages of current measures for crime data analysis. On this basis, combined with the advantages of data mining in data analysis, the idea of applying data mining technology to computer crime forensics is proposed. Contraposing the features like huge amount of analysis data, multiple attribute values and artificial participation, algorithm ID3 and its defects are analyzed and improved.

There are logarithms in ID3 and the computation is complicated, so we use bottom changing formula and Maclaurin formula to simplify the computation in ID3. Then the time to generate decision tree can be reduced. During the calculation of expected entropy of one attribute, by multiplying the simplified formula and the number of attribute value with this attribute, the error caused by simplification is effectively made up and the disadvantage in the second improved scheme is overcome. Meanwhile, this paper proposed the method to prevent the value deviation of ID3 algorithm, which belongs to the second improved program and it also provides a novel method PVF-ID3 to compute the expected entropy. PVF-ID3 has following features: (1) It effectively solves the problem of value deviation in ID3. (2) It effectively reduces the time for generating decision tree in ID3. (3) It effectively makes up the errors caused by the first improved method. (4) It can process data set with many classified attribute values.

There are also some limitations and shortcomings in our research: The selection of empirical value needs data mining experts who have rich working experience. In actual implementation, amounts of data will be changed along with time. The creation of decision tree will also change dynamically to provide correct prediction information. Therefore, the method in this paper is not very suitable for data set of dynamic scale, in classification and prediction. So it needs further improvement. Our future research focus will be about the parallel decision tree algorithm, accuracy of decision tree and discovering new decision tree algorithm. At present, the implementation of data mining technology is at the stage of theoretical research on computer crime analysis. It has not formed a practical product. Thus during the research on computer crime forensics analysis, it will become our future research on how to promote data mining technology to become real products.

ACKNOWLEDGMENT

This work was financially Supported by grant from the Shanxi Scholarship Council of China (No. 2009-28), the Natural Science Foundation of Shanxi Province (No. 2009011022-2) and the Natural Science Foundation for Young Scientists of Shanxi Province (No. 2012021011-3).

REFERENCES

- [1] Yasinsac Alec, Erbacher Robert F., "Marks Donald G., Computer forensics education", *IEEE Security and Privacy*, vol. 1, no. 4, pp. 15-23, 2003.
- [2] Oatley Giles, Ewart Brian, "Data mining and crime analysis", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 2, pp. 147-153, 2011.
- [3] Ding Li-Ping, Wang Yong-Ji, "Study on relevant law and technology issues about computer forensics", *Ruan Jian Xue Bao/Journal of Software*, vol. 16, no. 2, pp. 260-275, 2005.
- [4] Rudy Deca, Omar Cherkaoui, Yvon Savaria, "Rule-Based Network Service Provisioning", *Journal of Networks*, vol. 7, no. 10, pp. 1493-1505, 2012.
- [5] Tao Weidong, "Crime data mining based on extension classification", *Advances in Intelligent and Soft Computing*, vol. 115 no. 2, pp. 383-390, 2012.
- [6] ZHANG Bo, WU Lili, ZHOU Min, "The Analysis of User Forensics Based on Web Usage Mining", *Computer Science*, vol. 33, no. 8, pp. 213-215, 2006.
- [7] David McSherry, "Mixed-initiative problem solving with decision trees", *Artificial Intelligence Review*, vol. 23, no. 1, pp. 17-33, 2008.
- [8] Dai Zhen, Fei Hongxiao, Xie Wenbiao, "The Algorithm of users forensics associate rules mining Based on Specific Pattern Tree", *Computer Systems Applications*, vol. 13, no. 5, pp. 56-59, 2007.
- [9] Qu Kaise, CHEN Xiang-tao, "An Improved ID3 Algorithm of Decision Trees", *Computer Engineering & Science*, vol. 31, no. 6, pp. 109-111, 2009.
- [10] Ping Wang, Jing Han, Fuqiang Liu, "A Power Allocation Algorithm Based on Cooperative Game Theory in Multicell OFDM Systems", *Journal of Networks*, vol. 6, no. 11, pp. 1610-1618, 2011.
- [11] HUANG Ai-hui, CHEN Xiang-tao, "An Improved ID3 Algorithm of Decision Trees", *Computer Engineering & Science*, vol. 31, no. 6, pp. 109-111, 2009.
- [12] QU Kai-she, CHENG Wen-li, WANG Jun-hong, "Improved Algorithm Based on ID3", *Computer Engineering and Applications*, vol. 25, no. 39, pp. 104-107, 2003.
- [13] Zhong tao, Liu Hong, "ID3 optimization algorithm based on interestingness gain", *Proceedings of IEEE 3rd International Conference on Communication Software and Networks*, pp. 73-77, 2011.
- [14] Mahjoobi J, Etemad-Shahidi A, "An alternative approach for the prediction of significant wave heights based on classification and regression trees", *Applied Ocean Research*, vol. 30, no. 3, pp. 172-177, 2008.
- [15] GAO Yang, LIAO Jia-ping, WU Wei, "ID3 Algorithm and C4. 5 Algorithm Based on Decision Tree", *Journal of Hubei University of Technology*, vol. 26, no. 2, pp. 54-56, 2011.
- [16] Zhu Haodong, Zhong Yong, "Optimization of ID3 algorithm", *Huazhong Keji Daxue Xuebao*, vol. 38, no. 5, pp. 9-12, 2010.
- [17] Qian Zhang, "Reconstruction of Intermediate View based on Depth Map Enhancement", *Journal of Multimedia*, vol. 7, no. 6, pp. 415-420, 2012.
- [18] Zaman K. B. M. Q., Bridges J. E., Papamoschou, D., "Offset stream technology - Comparison of results from UCI and GRC Experiments", *Collection of Technical Papers*, vol. 8, no. 2, pp. 5243-5256, 2007.

Demand-oriented Traffic Measuring Method for Network Security Situation Assessment

Xu Zhenhua

Beijing Information Technology College, Beijing, China

Email: Shui0000@163.com

Abstract—In the information of security situation, traffic is necessary data to describe the network performance. It is also an important indicator to measure attacks such as worms or DoS. During the process of situational awareness, the traffic measuring methods are often needed without the privilege of routers. Our work emphasizes on the measuring algorithms based on packet pair in end-to-end measuring technologies. The simulations have shown that the congestion status of network and the length of packet have greater effect on the measurement of path capacity. So on the basis of improved packet pair measuring algorithms, a novel traffic measuring method DTM for demand is put forward. DTM aims to acquire the optimum compromise between measuring overhead and accuracy through the utilization of detecting packets. The relation between sending rate and available bandwidth are decided by one-way delay and changes of packets. Simultaneously, DTM assigns corresponding weight to the transition bandwidth, according to the disturbed degree of packet which is set as transition point, and its adjacent packets. Then it can better reflect the influence caused by background traffic. DTM also provides the rules to judge whether two packets are transition points and it discusses the method to determine the measuring range dynamically according to specific application bandwidth requirements. The experiments have studied the trend of relativity and the difference of packet pair, when the following factors are changing: size of detecting packet, network load, features of background traffic and bottleneck bandwidth. It is shown that DTM has characteristics in low overhead, high accuracy, good smoothness and sensitivity for network situation, compared to traditional algorithms.

Index Terms—Traffic Measuring; Packet Pair; Transition Point; Bandwidth; DTM; Congestion

I. INTRODUCTION

In order to find and defend sudden attack of cyberspace effectively as soon as possible, it is not enough to just rely on traditional safety protection technologies such as identity authentication, trusted computing, firewall, intrusion detection technology. By monitoring and identifying large-scaled intrusion intention and intrusion behavior in protected network, early warning technology based on safety situation perception can acquire more accurate safety threat behavior descriptions and more complete and timely network safety state estimations. Besides, this technology can predict attacking quantity and spatiotemporal characteristics before attack or its serious results. It can

take corresponding defending measures in advance to strengthen network safety. The study of large-scaled early warning technology in the networks is significant to improve emergency response capability in network system and ease the harm caused by attack [1]. It also helps to improve the systematic counterattack ability. The constant real-time measurement on network can enhance and guarantee connectivity and security in the whole network system. In addition, long-term monitoring data can be used to detect anomaly network traffic and illegal entry, which approaches the purpose of perceiving networked situation. Therefore, in safety state information, the traffic is not only the important data to describe network performance, but it is also an important index to measure some attacks including worm or DoS attack [2]. But under the condition that there is not authority to acquire flow data of network node, one of necessary problems is how to effectively measure network traffic in early warning system.

There exists certain proportional relation between link bandwidth and network traffic, while acquiring available bandwidth is the precondition and basis to measure network traffic. Many network bandwidth measuring technologies and tools are respectively proposed by academic circles and industrial communities: such as SNMP and RMP [3], the Netflow in CISCO company [4], packet pair measurement algorithm [5], measurement tool of link bandwidth Pathchar [6] and tool for network bottleneck Pathneek [7], etc. Dynamic characteristics of bandwidth present two important challenges: First, the measurement results reflect the network state during measuring execution. If network state has changed significantly when measurement is complete, the measurement results will be meaningless. This also demands that measuring time is not too long during execution of measurement tools; Secondly, if network is changing, especially available bandwidth is constantly changed along with time, the measurement results which can capture and reflect these dynamic characteristics are undoubtedly more valuable for application programs.

Packet pair is the kernel factor in many bandwidth measurement methods. Keshav [8] firstly proposes the packet pair idea which is also used to measure available bandwidth of fairness policy under network environment. Lai, etc [9] introduces the idea of packet pair in the network with FIFO strategy, such as link capacity measurement. Dovrolis, etc [10] furtherly analyzes the

multi-peak value of packet pair measurement from the perspective of capacity measurement. Martin [11] adopts the idea of packet discrete to propose a simple but effective capacity measurement method based on packet pair. Hul, etc [12] analyzes the available bandwidth measurement based on packet pair. Karame, etc [13] combines the idea of packet pair with PGM to propose Spruce method. Kang, etc [14] presents the analysis method of packet pair from theoretical analysis perspective.

This paper firstly studies the network bandwidth measurement method based on Packet Pair. We prove that this algorithm has some defects such as slow speed and low accuracy for measurement under heavy-loaded environment by simulation. Then according to conclusions from these results, we propose the improvement scheme based on tetrad packet group models and further presents a novel DTM (Demand-oriented Traffic Measuring) method. DTM makes full use of the information from detecting packet pair and combines packet pair with self-congestion measurement. By investigating one-way delay and the changing law of packet pair, the relationship between sending rate and available bandwidth can be acquired. The concept of transition point is proposed during measurement and the rules to judge whether packet pair is a conversion point or not are also presented. According to different interruption degrees of a packet which has become a conversion point on its adjacent packet, the weight with different bandwidth is assigned, so better measuring accuracy can be acquired at lower cost. Finally, based on specific methods to determine the measuring range of demand trends on bandwidth, the simulation experiments with NS-2 are used to verify the effectiveness of our scheme.

II. RELATED WORKS

A. Bandwidth Measurement based on Packet Pair

Packet pair technology is an important means of bandwidth measurement [15]: a packet pair is made up of two packets which are adjacent with similar sizes and sending rates. According to transmitting time, two packets are individually called the first packet and the second packet for packet pair. Time interval between two packets, that is, the difference value between the time after the second packet sending and the time after the first packet sending is called inner interval of packet pair. Correspondingly, the time interval of two adjacent packets between the second packet of previous one packet pair and the first packet of latter packet pair is called interval of packet pair. At this time, transmitting rate of packet pair can be defined as the ratio between its second packet size and inner interval of this packet during sending.

It is supposed that the path between source and destination is constituted of n hop links and two packets with L length are sent to destination end from source end, in the manner of back-to-back. These two back-to-back packets are called measuring packet pair. The manner of back-to-back indicates the initial interval of these two packets is as short as possible. The interval of packet pair

leaving the source end is defined as $T = L/R$ and R is the sending rate of source point on packet pair. Time interval that packet pair passes link α is T_α , time interval that packet pair reaches receiving terminal is T_n , R_b is the link bandwidth of link α and R_b is the measured path capacity.

With the consideration of the transmission process of the measurement packet on path: Before it reaches bottleneck node, if bandwidth of the i^{th} link is smaller than bandwidth of the $(i-1)^{th}$ link, time interval of packet pair after node i is $T_i = L/R_b$; else, $T_i = T_{i-1}$. After it passes the bottleneck node, time interval of measuring packet pair is $T_b = L/R_b$. Afterwards, time interval of measuring packet pair always keeps T_b till it reaches receiving terminal, which is shown as figure 1. In receiver, we can measure $T_n = T_b = L/R_b$ to calculate the bottleneck bandwidth of measured path with $R_b = L/T_n$.

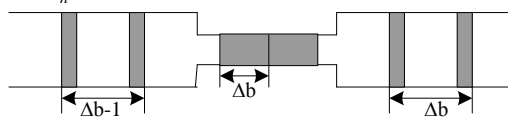


Figure 1. Interval of measuring packet pair passing through the bottleneck link

However, the simulation analyses of Packet Pair measurement algorithm indicate that measurement algorithm of Packet Pair cannot correctly measure network path capacity under heavy load. We will prove this through two experiments: A lot of researches on networked measurement have indicated that packet length transmitted by Internet mainly focuses on 3-4 numerical points. About 50% packet length is 40 bytes, 20% packet length is 552 bytes or 576 bytes and 15% packet length is 1500bytes [16]. Therefore, to simulate real network state, we impose networked load addition of 4 Pareto flow synthesis on each node path and simulate different network environment by controlling the loaded strength of traffic.

Figure 3 shows the distribution of each link utilization ratio u in the path when u is 10% and 60%. In this figure, when $u=10\%$, that is, when network is lightly loaded, among measured samples, the measured samples which are equal to those of true values will be predominant. After filtering, measured values of correct path capacity can be acquired. When $u=60\%$, that is, when network load is very high, among measured samples, the measured samples which are smaller than those of true values will be predominant. After filtering the path capacity measured values will be certainly smaller than those of practical path capacity. The reason is when network load is heavy, the probability of inserting other packets into two packets is largely increasing in packet pairs, which causes that increase of sample quantity, which is smaller than true values. Thus, when measured Packet Pair algorithm is in light load, the path capacity can be correctly measured. When network

loading is high, accurate results cannot be acquired. At this time, measured value is smaller than practical capacity value.

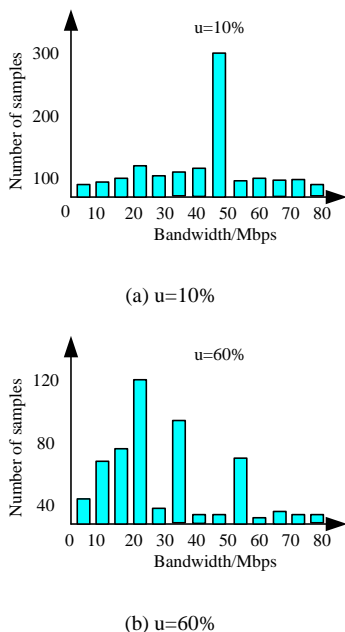


Figure 2. Samples distribution under different load

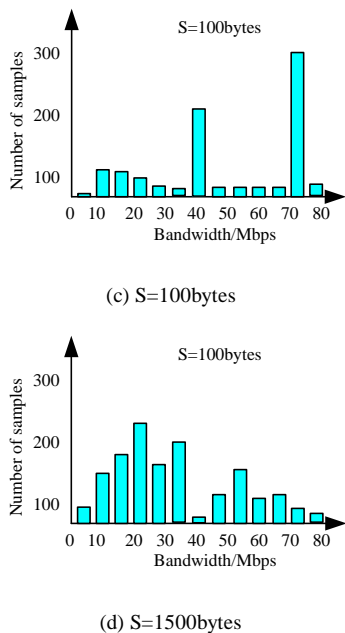


Figure 3. Samples distribution under different packet pair length

B. Determination of Transition Point

Packet pair series has the feature that inner interval is reducing continuously, which cause sending rate of packet pair increasingly enlarged for upper bound of measurement from lower terminal. Meanwhile, to keep the interval of packet pairs in enough size, which promote packet pair not to be interrupted. Measuring edge can be directly defined before measurement and it can also be defined during measurement based on network state. At the lowest expense, the measurement result can be

acquired which only needs to satisfy the packet pair sequence of above features. Figure 4 shows the detecting packet sequence formed by N packet pairs.

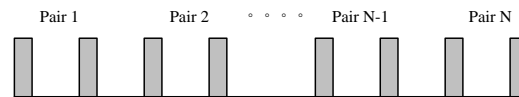


Figure 4. Detecting packet series composed N packets

By analyzing interruption of the packet pair, the relationship between the sender and available bandwidth on packet pair can be judged. For two packet pairs which are sent consequently but have opposite relation in available bandwidth, their background flow interruptions have significant difference. Those packet pairs whose sending rates are large will suffer more interruptions. According to definition of available bandwidth, existing network application can be interrupted. We will believe that this adjacent packet with significant interruption exists and it is reasonable to take slow sending rate as available bandwidth. In a group of packet pair series, when one packet is compared to its previous packet, if its background flow interruption is significantly increasing, this packet is called the transition point of measurement. The sending rate of previous packet pair is its transition bandwidth.

Abruptness of network traffic is an important factor to influence available bandwidth and measurement in short time. At busy stage of abruptness, that is, the time with background flow will reduce available bandwidth; while at free stage of abruptness, the time without background flow will increase available bandwidth. Abruptness may produce new transition point due to short time interval. Burst traffic cannot determine available bandwidth in certain time so the transition point caused by changing size of sending rate and available bandwidth; while the transition point caused by abruptness, must be distinguished. We will assign different weights with different transition points. The transition point weight caused by the change on the relationship between sending rate and available bandwidth is the maximum. It will make determination function for measuring results.

III. DEMAND-ORIENTED MEASURING STRATEGY

A. Measuring Process

Available bandwidth reflects the dynamic change of network. The measurement of available bandwidth should capture and respond the rapid change of network. It means two aspects: first, it can provide stable measuring results when network status keeps unchangeable; second, it has the sensitivity to reflect the change of network status rapidly. Therefore, our improved method analyzes the features of packet pair and makes full use of the implicated information. So it will control the overhead effectively with better stability and sensitivity, which can satisfy the demand for security assessment.

DTM method needs to measure the joint of sender and receiver. The sender is in charge of sending detecting packets and computation of available bandwidth; The receiver will return the timestamps to the sender when the

detecting packets arrive, then it can acquire the one-way delay of all the packets and the interval between packet pair of sender and receiver. When a group of time information of the detecting packet pair series is obtained, we use the following judging rules to acquire all the transition points and transition bandwidth. DTM introduces the weighted process disturbed by neighbour packets based on transition points. The measuring procedures in detail:

(a) Analyze the disturbance of detecting packets which are adjacent to the transition points. We define $d = n / N$ as the disturbed degree of packets. N denotes the total number of analyzed packets and n denotes the number of packets endured obvious disturbance. In DTM we only consider the packet pairs of the transition point series, or those prior and inferior to the transition point, so $N = 6$.

(b) Since the adjacent packets of the transition point in idle period will cause bigger sending rate than available width, it also suffers obvious disturbance of the background flow; On the contrary, the adjacent packets of the transition point in busy period causes smaller sending rate than available width, so it receives limited disturbance. Thus the value range of d is $(-\infty, 0.5)$ in idle period and $(0.5, 1)$ in busy period. The next step for weighting is to assign different weight value to corresponding transition points by weighted function. The weighted function takes $d \in (0, 1]$ of each transition point as input. When the input is 0.5 it obtains the maximum output value. It is increasing with the difference between d and 0.5 while the output weight will become smaller gradually.

(c) Set the transition width of all the transition points acquired by the process of a group of packet pair series as B_1, B_2, \dots, B_j . Simultaneously there corresponding weight are w_1, w_2, \dots, w_j . Then the measuring result of path available bandwidth can be expressed as

$$B = \frac{\sum_{i=1}^j BW_i \cdot w_i}{\sum_{i=1}^j w_i} \quad (3)$$

(d) DTM adopts multiple groups of packet pair series for measurement. The final measuring result is the mean value of measuring results of every series, as shown in formula 4. K denotes the number of packet pair series.

$$AW = \frac{\sum_{i=1}^n BW_i}{K} \quad (4)$$

B. Tetrad Packet Pair Measuring

Analysis on the simulation for Packet Pair algorithm has shown that: The measuring packets may insert packets of other flow into the pair at bottleneck node, which causes interval expansion; the packet pairs may also insert packets of other flow into the sequent node of bottleneck node, which causes interval compression. The

above will lead to smaller or bigger measuring value than real value. It needs more measurement and enough measuring samples to get correct measuring value. But it cause too much measuring time and network resource consumption. Then the measuring results are always smaller than real value when network reloading, leading to bad accuracy [17]. To overcome this defect, the measuring group in our model is a pair of sequent packet pair with different length, instead of single packet pair, to make up for the Packet Pair algorithm. The two pairs are called a packet tetrad in this paper.

From the conclusion in pervious sector, no matter how is the measuring environment, there are always equivalent samples with real value. Only the number is affected. Therefore, if we can judge if each sample is interrupted in time, the measurement will be finished when finding measuring samples without interruption. It has reduced the measuring time and raised the measuring accuracy. It is performed by a packet tetrad and the interval change of two sequent packets is shown in figure 5.

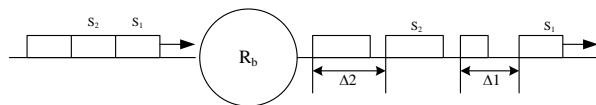


Figure 5. Interval of packet passed the bottleneck node

In this packet tetrad, the length of packet in the first packet pair is S_1 , the length of packet in the second packet pair is S_2 , $S_2 > S_1$. The interval when the first packet passes through the bottleneck node is Δ_1 , the interval when the first packet passes through the bottleneck node is Δ_2 . If no other interruption accepted, the capacity of measuring path is:

$$C_b = S_1 / \ddot{A}_1 = S_2 / \ddot{A}_2 \quad (5)$$

Then

$$S_2 / S_1 = \ddot{A}_2 / \ddot{A}_1 = \alpha \quad (6)$$

Even if the above formulas are feasible, the pack pairs still will suffer interruption from other packets. In the interference flow length ratio of two packet pairs is α . So we need take measures to differentiate these two cases. The packet length mainly concentrates in 3-4 points. About 50% packet length is 40 bytes, 20% is 552 bytes or 576 bytes, 15% is 1500 bytes. The ratio among these value may be 2.6, 2.72, 13.8, 14.4, 37.5. Thus, if the ration of two measuring pairs selected by us is away from these value, the interruption can be eliminated effectively. Table 1 shown the times needed to measure correct results for Tetrad Packet Pair algorithm, under different load for path $P = \{100, 70, 55, 40, 60, 80\}$ when $S_1 = 100$ bytes and $S_2 = 800$ bytes.

TABLE I. MEASURING TIMES NEEDED UNDER DIFFERENT LOAD FOR CORRECT RESULTS

Efficiency	10%	25%	40%	50%	70%	80%
Measuring times	3	7	19	36	95	152

From table 1 we can see, Tetrad Packet Pair method can measure the capacity of path rapidly and accurately. When network load is light, measuring speed is fast; On the contrary, though the speed is decreasing, we can still get accurate value. If the delay of RTT (Round Trip Time) from sender to receiver is 20ms, Tetrad Packet Pair algorithm only takes 60ms at its fastest speed and 3.04s at its lowest speed, to measure the bandwidth successfully.

C. Analysis on Transition Point and Weight

Related researches demonstrate that Internet traffic has properties of self-similar or long-range correlation [18]. So abruptness may cause the packet to be transition point. This transition only takes limited effect on available bandwidth. Since the relation between sending rate and available bandwidth has decisive function to transition points, the measurement of available bandwidth should handle with the effect including abruptness. Then the transition points are divided into following classes:

(1) If sending rate of the packet pair P is bigger than available bandwidth, and the sending rate of its adjacent previous packet pair is lower than available bandwidth. Compared to the previous packet pair at this moment, P will suffer obvious interruption. Then P is the defined transition point, which is the general case.

(2) If sending rate of P is lower than available bandwidth, but P encounters a burst or a transition period of large background streaming data, the interruption will be more serious compared to its previous packet pair. Because P meet the conditions of transition point, available bandwidth can get decreased by transition of P .

(3) If sending rate of the packet pair adjacent to P is larger than available bandwidth, and it just encounters an idle period without burst flows. At this point the packet pairs will only get limited interference. But the interruption of P is more during this period. Such packet pair also coincides with the condition of transition point.

To analyze different features of different kinds of transition points, we define $\rho = m/N$ to describe the interruption degree. m denotes the number of packets suffering obvious interruption, N denotes the total number of analyzed packets. The range includes transition points and its adjacent two packet pairs, six packets altogether, as is shown in figure 6, so $N = 6$.

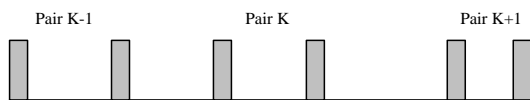


Figure 6. Detecting packet series composed N packets

For the first class, its adjacent previous packet pair has lower sending rate than available bandwidth so its interruption is limited; the adjacent previous packet pair suffers serious interruption. Then their corresponding ρ is close to 0.5. Similarly, the transition points of the second class should satisfy $\rho \leq 0.5$, the transition points of the third class satisfy $\rho > 0.5$. We can further use ρ to acquire the weight of transition point. The input of

weighted function is ρ and output is the weight of corresponding transition point. The weighted function should satisfy the following properties: its range is $(0,1]$; it gets the maximum value when input parameter is 0.5; along with the increasing or decreasing direction of input value, its output value will reduced gradually, which is close to 0 but is always greater than 0.

D. Termination for Threshold

The threshold has key function in judgement of transition point and corresponding weight. But it is a experienced value in fact. To test the performance of DTM method, we must observe the effects on measurement with different threshold. Figure 7 provides comparison between the measuring results and actual bandwidth with multiple thresholds. Chang of k decides corresponding threshold. When $k = 3$, the result is much closer to real value. If k is smaller, its threshold is big, which reduces the number of packet pairs or packets suffering serious interruption. It determines that the packets owing higher sending rate can be transition points and the packets owing low sending rate when encountering burst flow are excluded. According to definition, ρ is reducing simultaneously and it causes higher measuring results. Similar analysis is adaptive to the case with bigger k , which leads to lower measuring result. Therefore we adopt $k = 3$ in subsequent experiments.

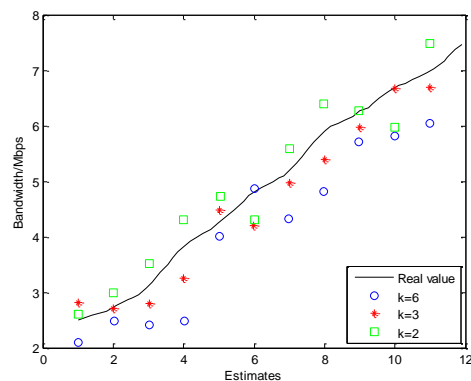


Figure 7. Measuring difference with different k value

IV. SIMULATION ANALYSIS

Figure 8 presents simulation topology structure based on NS2. The measuring path is made up of $n+1$ links between Sender and Sink. Background flow is n connections from P_i to Q_i ($i=1,2,\dots,n$). Previous $n-1$ connections and measurement paths only coincide with some link L_i . The last connection and measurement path between P_n and Q_n coincide with $n-1$ links. Each connection produces a group of Pareto flow ($\alpha = 1.6$). TCP and UDP flow respectively takes up 95% and 5%. Packet pair size contains 1500Bytes, 700Bytes and 40Bytes and it individually takes up 10%, 80% and 10% of the traffic. The shortest link of measuring path

capacity lies in the center of the path whose capacity is 10Mbps. All the other link capacity is 15Mbps and size of background flow in the path is all the same. By controlling the size of background flow, available bandwidths with different paths can be acquired. In this experiment, each packet pair series is made up of 25 packet pairs whose rates are increasingly enlarged. Its measurement scope is (400K,10M) and sending rate of adjacent packet pair has 400Kbps difference. When weighted function is input the value p to each transition point, weighted function will produce corresponding weight of this transition point.

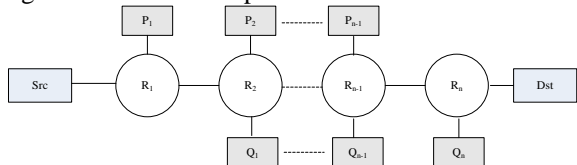


Figure 8. Topology of network simulation

A. Difference among Different Size of Packets

In order to compare packets with different packet sizes when they capture and reflect some differences, including available bandwidth, traffic abruptness, etc, when available bandwidth is 4.5Mbps. The sending rate is 1Mbps, 4.6Mbps and 8Mbps when multiple packet pairs are sent. During sending process, packet pair will keep enough interval to ensure that packet pairs will not have any interruptions. Figure 9 shows the inner interval difference SV distribution of packet pair on receiver and sender when packet pair size is 700Bytes. Figure 10 shows the distribution of each packet on the second one-way delay.

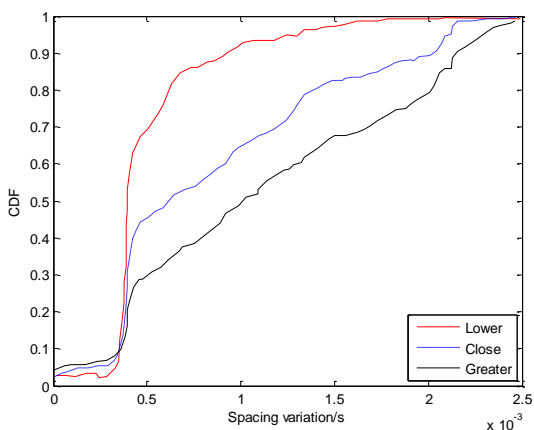


Figure 9. SV distribution of 700Bytes packet pair with different sending rate

When sending rate is slower than available bandwidth, inner interval 60% of packet on receiving end and sending end is approximating to 0 in figure 9. When they reach 80%, absolute value is less than 0.00055. From figure 10, the second one-way delay value of packet of more than 80% packet pair is less than 0.075 which means that most of packet pair is not or only influenced by limit background flow. However, when sending rate approximates to available bandwidth, inner interval

difference of packet pair will appear increasing tendency. Compared to short sending rate, one-way delay of more than 50% packet on the second packet will be more than 0.075. After further investigation, we can discover that one-way delay of the second packet has better stability when sending rate is slower than available bandwidth. When sending rate approximates to available bandwidth, one-way delay stability is significantly reduced. These tendencies will be more significant than those of approximating available bandwidth when sending rate is larger than available bandwidth. Therefore, packet with 700Bytes size cannot only correctly capture available bandwidth but it can reflect the abruptness more effectively.

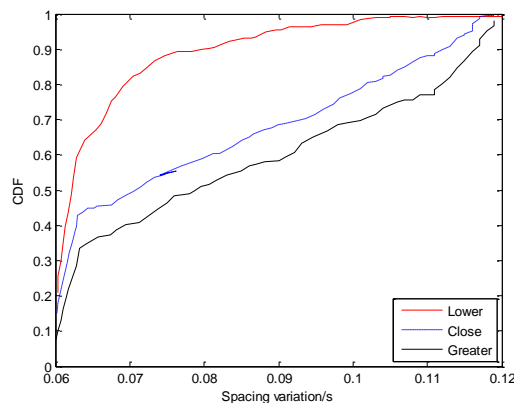


Figure 10. One-way delay distribution of the second packet pair with different sending rate

When packet size is 40Bytes or 1500Bytes, the previous one is found that it cannot capture available bandwidth. That is, when sending rate is larger than available bandwidth, packet pair does not show features influenced by significant interruption of background flow. The latter one is totally opposite. Even if the sending rate is slower than available bandwidth, there are certain packet pairs which are significantly interrupted. In summary, the detecting packet size of DTM is 700Bytes.

B. Performance under Congestion Conditions

The performance of DTM is tested under two background flow conditions as single congestion and multi-congestion. The size of background flow of other links and narrow links is the same. It focused on that narrow link capacity is the least one among all link capacities. Therefore, in the experiment, narrow-linked bandwidth is always the least value in the whole path. That is, at this time, narrow link and tight link are unified. Compared to single congestion, the much possibly interrupted link of detecting packet under multi-congestion condition contains tight link and two adjacent links of tight link

The performance of DTM is tested under two background flow conditions as single congestion and multi-congestion. For single congestion, all link background flow size is the same; for multi-congestion, the bandwidth utilization of narrow link and its two adjacent links is the same, too. The size of background flow of other links and narrow links is the same. It needs

attention that narrow link capacity is the least among all link capacities. Therefore, in the experiment, narrow linked available bandwidth is always the least value of the whole path. That is, at this time, narrow link and tight link are unified. Compared to single congestion, the much possibly interrupted link of detecting packet under multi-congestion condition contains tight link and two adjacent links of tight link.

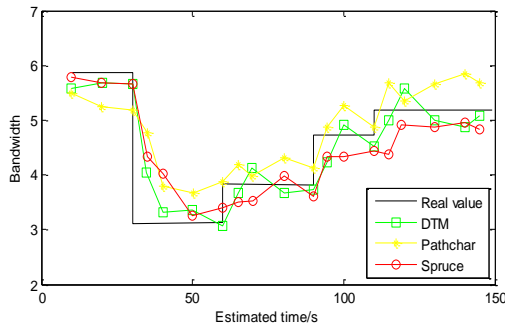


Figure 11. Comparison under single congestion

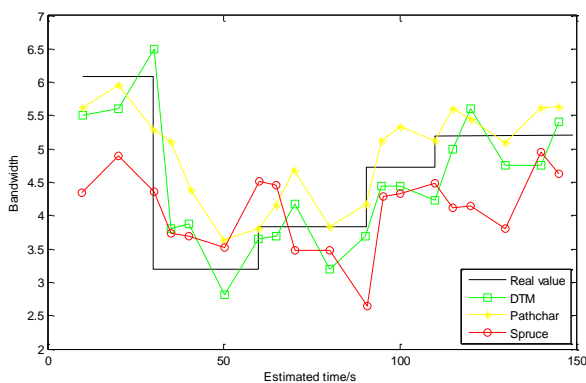


Figure 12. Comparison under multi-congestion

Through controlling the size of background flow, available bandwidth with two conditions will change at 15s, 45s, 80s and 110s. Meanwhile, it will keep stable between adjacent time points. Figure 11 and figure 12 respectively present the comparisons between DTM, Pathchar, Spruce and practical value, under single congestion and multi-congestion condition. The measurement result of Pathchar is the arithmetic mean between upper bound and lower bound on its recorded intervals.

Through controlling the size of background flow, available bandwidth with two conditions will change at 15s, 45s, 80s and 110s. Meanwhile, it will keep stable between adjacent time points. Figure 11 and figure 12 respectively present the comparison between DTM, Pathchar, Spruce and practical value, under single congestion and multi-congestion condition. The measurement result of Pathchar is the arithmetic mean between upper bound and lower bound on its recorded intervals.

C. Measuring Overhead

In this experiment, in more than 90% measurements, DTM only needs less than 4 packet pairs to get more

accurate measurement results. Therefore, measuring overload of DTM is not more than Spruce and is significantly lower than Pathchar. Previous dynamic process to determine upper bound and lower bound of measurement only needs limit detecting overhead. When networked state keeps stable, it will not need each measurement to perform this process [19]. The minimal detecting unit in these three methods, such as two packet pairs of Pathchar or between DTM and Spruce, always has large interval. Measuring process of each method will not result in long-term congestion in the network. Research shows the survival time of most Internet traffic is very short. To test the influence on these short-lived network flows, we define the longest congestion time of network as a minimum detecting unit time for the corresponding method to handle the bottleneck link. The shortest continuous congestion time is the needed time by the minimum detecting packet unit in measuring process.

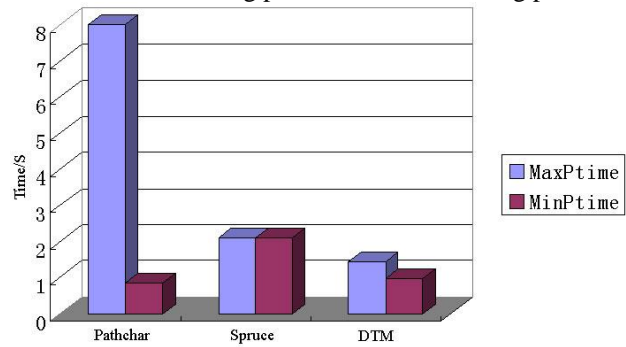


Figure 13. Comparison of the longest and shortest duration time

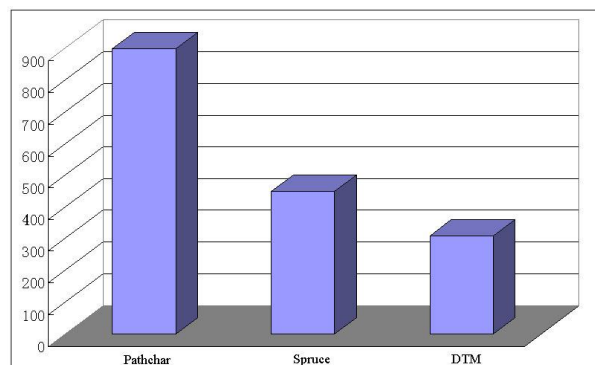


Figure 14. Comparison of measuring overhead

The shortest detecting unit of Pathchar is a packet series while the shortest detecting unit between Spruce and DTM is one packet pair. In addition, the shortest congestion time between Pathchar and DTM is the time to handle single packet. Since Spruce adopts back-to-back packet pair as measurement unit, its shortest congestion time is the time to deal with one packet in network. Figure 13 compares the longest and the shortest time between Pathchar, Spruce and DTM during measurement in simulations. From figure 14 we can see, the time of two congestions is the lowest for DTM. In different environments, each specific value of congestion time has difference. However, the size of corresponding values on different methods reflected by figure 14 is the same. In conclusion, DTM only needs little network

interruption to get more accurate measurement results, due to full use of packet pair information of DTM.

V. CONCLUSION

With the expansion of network scale and increasing complexity of network environment, it is too difficult to comprehensively inspect performance of network path capacity measurement algorithms in practical network. However, due to some advantages of controllability, repeatability, expansibility, etc on network simulation measurement, network simulation measurement has become a reasonable means to measure the performance of algorithms. Based on improved packet pair measuring method, we propose a novel available bandwidth traffic measuring strategy-DTM. It lies in making full use of existing demand information to get better measuring accuracy at lower cost. The basic idea of DTM is similar to Pathchar, which is by means of investigating the relationship between sending rate, acquired by one-way delay with its change and available bandwidth. According to different interruption degree of packet with transition point and its adjacent packet, DTM will assign different weights to corresponding transition bandwidth to reflect the influence of background abruptness on available bandwidth. We have provided two rules to judge whether the packet can become transition point. We also discuss specific methods for using demand trends of bandwidth to determine measurement range. The results prove that DTM has comprehensive performance advantages of low-cost, high-accuracy and better smoothness, compared to algorithms such as Pathload and Spruce.

REFERENCES

- [1] Q. He, C. Dovrolis and M. Amman, "On the Predictability of Lane Transfer TCP Throughput", *In Proceedings of ACM SIGCOMM*, pp. 131-139, Philadelphia PA, 2005.
- [2] Liu Xiao-Wu, Wang Hui-Qiang, Yu Ji-Guo, "Network security situation awareness model based on multi-source fusion", *Journal of PLA University of Science and Technology*, vol. 13, no. 4, pp. 403-407, 2012.
- [3] WANG Huan-ran, XU Ming-wei, "Survey on SNMP Network Management", *MINI-MICRO SYSTEMS*, vol. 25, no. 3, pp. 358-365, 2004.
- [4] Nevil Brownlee, "Network Management and Realtime Traffic Flow Measurement", *Journal of Network and Systems Management*, vol. 11, no. 2, pp. 223-228, 1998.
- [5] Jian Kang, Yuan-Zhang Song, Jun-Yao Zhang, "Accurate Detection of Peer-to-Peer Botnet using Multi-Stream Fused Scheme", *Journal of Networks*, vol. 6, no. 5, 2011.
- [6] Vinay Ribeiro, Rudolf Riedi, Richard Baraniuk, "pathChirp: Efficient Available Bandwidth Estimation for Network Paths", *Passive and Active Monitoring Workshop*, vol. 6, no. 4, pp. 1-10, 2003.
- [7] Park Kyung-Joon, Lim Hyuk, Choi Chong-Ho, "Stochastic analysis of packet-pair probing for network bandwidth estimation", *Computer Networks*, vol. 50, no. 12, pp. 1901-1915, 2006.
- [8] Keshav, L. Chen, L. Lao, "CapProbe: A Simple and Accurate Capacity Estimation Technique", *In Proceedings of SIGCOMM*, pp. 67-78, USA, 2004.
- [9] Yoo H., Jang J., "An improved packet pair method for filtering estimation noise and fast convergence in measuring bottleneck bandwidth", *Proceedings of the International Conference on Parallel and Distributed Systems*, pp. 763-767, Korea, 2001.
- [10] C. Dovrolis, P. Ramanathan, D. Moore, "Packet Dispersion Techniques and a Capacity Estimation Methodology", *IEEE/ACM Transactions on Networking*, vol. 12, no. 3, pp. 1000-1025, 2004.
- [11] Tunnicliffe Martin J., Winnett Maria, "An approximate stochastic analysis of the packet-pair probing technique for available bandwidth estimation", *International Journal of Communication Systems*, vol. 22, no. 6, pp. 651-669, 2009.
- [12] Ziqiang Wang, Xia Sun, "Manifold Adaptive Kernel Local Fisher Discriminant Analysis for Face Recognition", *Journal of Multimedia*, vol. 7, no. 6, pp. 387-393, 2012.
- [13] Karame Ghassan O., Danev Boris, Bannwart Cyrill, "On the security of end-to-end measurements based on packet-pair dispersions", *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 149-162, 2013.
- [14] Dey Bikash Kumar, Manjunath D., Chakraborty Supriyo, "Estimating network link characteristics using packet-pair dispersion: A discrete-time queueing theoretic analysis", *Computer Networks*, vol. 55, no. 5, pp. 1052-1068, 2011.
- [15] Huang Yu, Zhang Xiaoyi, Song Li, "An available bandwidth measurement method based on the Parity-Packet pair model", *Proceedings of International Colloquium on Computing, Communication, Control, and Management*, pp. 203-205, China, 2009.
- [16] LI Weiping, ZHANG Lei, "Measurement of Path Bandwidth in Wired/Wireless Ad Hoc Hybrid Networks", *Computer Communications*, vol. 21, no. 7, pp. 81-84, 2010.
- [17] Zhou ZiXuan, Lee Bu Sung, Fu Cheng Peng, "Packet triplet: A novel approach to estimate path capacity", *IEEE Communications Letters*, vol. 9, no. 12, pp. 1076-1078, 2005.
- [18] Guo-hong Gao, Wen-xian Xiao, Zhen Liu, Wen-long Wan, "An Application of the Modification of Slow Start Algorithm in Campus Network", *Journal of Networks*, vol. 6, no. 11, pp. 1549-1556, 2012.
- [19] Kevin Lai, Baker Mary, "Measuring link bandwidths using a deterministic model of packet delay", *Computer Communication Review*, vol. 30, no. 4, pp. 283-294, 2000.

Visual Simulation of Explosion Effects Based on Mathematical Model and Particle System

Gong Lin and Hu Dingjun

Zhenjiang Watercraft College, Zhenjiang, China

Email: glin402@126.com

Abstract—The paper firstly analyzed the factors related to the smoke diffusion and dissipation when explosion, and the mathematical model of smoke diffusion and dissipation are studied. Then the approach of particle system is used to management the movement of smoke particles. At last, the texture mapping techniques and billboard techniques are used to rendering the smoke. The results show that this method for visual simulation of exploding in virtual environment is efficiently and realistic.

Index Terms—Mathematical Models; Smoke Simulation; Particle Systems; Texture Mapping

I. INTRODUCTION

With the development of computer graphics technology, the scenes of smoke, flame, explosions, cloud, snow and other dynamic irregular fuzzy objects appear more and more in virtual battlefield environment. In virtual battlefield environment, the explosion effect is an important element, whether battlefield environment is realistic largely affects the trainer's training effect. But due to the dynamic irregular geometry and uncertainty internal of smoke, which makes the classical Euclidean geometry becoming powerless. Meanwhile, the movement of smoke in the combustion process is subjected to various internal and external factors, so the general modeling approach is difficult to describe it.

In recent decades, computer graphics researchers have made a lot of smoke simulation methods, some of which is effect, and the application fields also are larger. But for various reasons, the method which considered satisfactory generally is still relatively limited; it is difficult to say which method is more suitable for people's needs. Thus for computer workers, especially computer graphics researchers, the development of new and more convenient way to simulate the irregular fuzzy objects like smoke and so on is essential, and combining the existing methods organically to meet the needs of the times is still an important issue.

A. The Previous Work

Commonly, there are three methods for such dynamic irregular fuzzy object's modeling. There are mathematical and physics-based model, Particle system model and texture mapping method [1] [2] [3]. Particle system is relatively simple, it is easy to describe the performance of smoke's spread and disappeared details. Especially, the random changes in performance are

relatively easy, but because of the movement process is too random, we can not describe the changes of the smoke illumination accurately. The mathematical and physics-based methods use the fluid motion equations to describe the movement of the smoke. Its advantage is scientific and reasonable, so the effect of the visual simulation is real. But solving the hydrodynamic equations is an explicit method, only time step size is small enough to ensure operation stability, which also led to the operation speed is very slow, it will affect real-time. The method based on texture mapping technology has obvious advantages on speed though the application scope is limited. A simple texture image can replace a large number of particles, so it can resave a lot of computer resources and the speed is accelerated. The core of texture mapping technology is how to structure the application method, how to choose a texture image correctly, and how to handle the properties of a number of texture images. In addition, how to make it more in accordance with people's visual effects closely is needed to further research. The comparisons of three kind methods for smoke simulation are shown in table 1.

B. Our Work

Each method had their advantages and shortcomings, application scopes and result characteristics. Aiming to three methods above, this paper first studies the diffusion and dissipation mathematical model of smoke when explosion, then using the model of particle system to control the movement of smoke particle. The combining of particle systems and mathematical models can overcome the shortcomings of particle motion's random. When rendering particles, we use texture mapping technology and billboard technology to improve rendering speed. The result shows that this algorithm is simple and real-time.

II. MODELING OF SHELLS EXPLODING

Smoke diffusion of explosion can be seen as a rapid diffuse process from the point source to infinite space instantaneously, Second-order parabolic partial differential equations can be used to describe the smoke concentration variation. Smoke diffusion process and the smoke concentration variation which we observed is relate to light absorption, and also relates to the sensitivity of observation instruments or the naked eye. For example, with the naked eye that the smoke is gone,

TABLE I. THE COMPARISONS OF THREE KIND METHODS FOR SMOKE SIMULATION

	Particle systems model	Mathematical and Physics model	Texture mapping model
Computational	relating with the number of particles, the particle is more and the calculation amount is bigger	calculation process is complex, the total computation amount is very large	calculation process is relatively simple, depending on texture's controlling method
Speed	it is faster, but when particle is more, the time is slower	Calculation process is slow, tens of seconds to a few minutes per frame	Calculation is relatively small, depending on the rendering speed
Manifestations	the performance is realistic, but it has a great randomness	the performance is realistic, but flexibility is limited	the performance is realistic, but rendering presence is artificial
Memory Consume	memory requirements is affected by the data structure and particles number	memory consumes by Calculation is more, parallel processing is better	memory consume is less than the particle system, relating to the texture

and with the instrument of high sensitivity is still observed [4] [5].

Therefore, the entire modeling process should contain: the variation of the smoke concentration; light intensity variation through the smoke; the description to identify bright and dark of instrument sensitivity; changing process of opaque zone boundary. In order to simplify the modeling, First of all, we make the following assumptions:

(A) Under ideal conditions (without considering the impact of wind and earth), the explosion of shells as a point in the air releasing instantaneous smoke, smoke spread in the infinite space.

(B) The diffusion of smoke obeys the diffusion law, i.e. the area of the flow rate is proportional to its concentration gradient.

(C) The light strength passing through the smoke is reduced according to the absorption of the smoke, the reduction of light intensity on unit distance is proportional to the concentration of smoke; the light absorption in atmosphere effect is negligible.

(D) In the diffusion process of smoke, light intensity I_0 which does not pass through the smoke directly into the standard observation instrument remains unchanged. The light intensity I through the smoke into of the instrument or the naked eye, the observations result is

light and dark, only when $\frac{I}{I_0} > 1 - \mu$, the observations bright. μ is the sensitivity of the instrument or the naked eye, μ is smaller, the instruments or the naked eye is more sensitive, usually $\mu \ll 1$.

A. Change Law of Smoke Concentration

The explosion occurred time is counted as $t = 0$, the coordinate points for the origin. Smoke concentration on time t at any point (x, y, z) in infinite space is $C(x, y, z, t)$. The assumptions of the unit time through unit normal traffic to the area as following

$$q = -k \cdot \text{grad}C \tag{1}$$

K as Diffusion coefficient, grad as Gradient,, A negative sign indicates diffusion concentration by high concentration to low. Ω as spatial domain investigated, V as volume, Ω was surrounded by the surface, outer normal vector of S is n , the flowing through Ω in the $[t, t + \Delta t]$ as following

$$Q_1 = \int_t^{t+\Delta t} \iint_S q \cdot n d\sigma dt \tag{2}$$

The smoke increments within Ω as following

$$Q_2 = \iiint_V [C(x, y, z, t) - C(x, y, z, t + \Delta t)] dV \tag{3}$$

According to The law of conservation of mass

$$Q_1 = Q_2 \tag{4}$$

And The Albright formula based on surface integrals

$$\iint_S (q \cdot n) d\sigma = \iiint_V (\text{div}q) dV \tag{5}$$

where div is the divergence of the mark, by (1) - (5), and then using the integral mean value theorem, we get

$$\frac{\partial C}{\partial t} = k \text{div}(\text{grad}C) = k \left(\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2} \right) \tag{6}$$

This is parabolic partial differential equations. Based on the assumption (a), the initial conditions is point source function in the origin, can be counted as

$$C(x, y, z, 0) = Q \delta(x, y, z) \tag{7}$$

Q as the total cast smoke, $\delta(x, y, z)$ as the point source function. The solution of the equation (6) satisfies the conditions (7) is

$$C(x, y, z, t) = \frac{Q}{(4\pi kt)^{3/2}} e^{-\frac{x^2 + y^2 + z^2}{4kt}} \tag{8}$$

This result indicates that, for any moment, the smoke concentration surface C is the spherical as $x^2 + y^2 + z^2 = R^2$, and with the increase of spherical radius R , the value of C is reduced continuously.

B. Light Intensity Variation Through of Smoke

The light pass through the smoke in a certain direction, the length coordinate on the direction account as l , $C(l)$ is the smoke concentration, $I(l)$ is the light intensity, in accordance with the assumptions (c), we will get

$$\frac{dI}{dl} = -\alpha C(l)I(l) \tag{9}$$

α is light absorption coefficient, if the intensity of the light does not enter the smoke ($l = l_0$) meter for I_0

$$I(l) = I_0 \tag{10}$$

Solutions of Equation (9) under the conditions (10) is

$$I(l) = I_0 e^{-\alpha \int_0^l C(s) ds} \tag{11}$$

C. Instrument (Naked Eye) Sensitivity and the Opacity Area Boundary

From the above analysis we can get smoke concentration in the space is changing continuously, the light intensity through the smoke into the instrument is changing continuously too. then the reason will be observed to the spread of smoke opacity when the boundary of the area has a larger first, then eventually disappear completely because the instrument observations only light and dark of the points, while the bright and dark dividing line is determined by the sensitivity, based on the assumption that (d) only if the

$$\frac{I}{I_0} > 1 - \mu \tag{12}$$

Observations are bright, and then you can think that the smoke has completely disappeared. taken along the axis of light, set viewpoint in place $z = \infty$ and the light source in place $z = -\infty$, when smoke completely disappeared, the condition (12) by the formula (11) can be written

$$e^{-\alpha \int_{-\infty}^{+\infty} C(x, y, z, t) dz} > 1 - \mu \tag{13}$$

Because surface $C(x, y, z, t)$ is spherical, therefore, the area boundary of the spread of smoke projection on xy plane is the circumference, referred to as

$$x^2 + y^2 = r^2 \tag{14}$$

The variation $r(t)$ over time of circumference radius is determined by the condition (13).

D. Smoke Diffusion and Dissipation Variation

Using approximate relations $\ln(1+x) = x(x \ll 1)$, (13) can be turned into

$$\int_{-\infty}^{+\infty} C(x, y, z, t) dz < \frac{1}{\alpha} \ln \frac{1}{1-\mu} = \frac{\mu}{\alpha} \tag{15}$$

The formula (8) is integrated by substituting into (15), and using the formula $\int_{-\infty}^{+\infty} e^{-x^2/a} dx = \sqrt{\pi a}$, we will get

$$\frac{Q}{4\pi kt} e^{-\frac{x^2+y^2}{4kt}} = \frac{1}{\alpha} \tag{16}$$

Smoke spread radius of the area can be impressed by

$$r(t) = \sqrt{4kt \ln \frac{\alpha Q}{4\pi kt}} \tag{17}$$

Formula (17) can be drawn in Figure 1, when

$$t = t_1 = \frac{\alpha Q}{4\pi ke} \tag{18}$$

Smoke diffusion region radius r reaches its maximum r_m , when

$$t = t_2 = \frac{\alpha Q}{4\pi k} \tag{19}$$

$r = 0$, at the moment t_2 , smoke disappear completely.

(18), (19) shows that: t_1, t_2 is proportional to the light absorption coefficient α and Smoke who discharge Q ; inversely proportional to the diffusion coefficient k . from (18), (19), we can get $t_2 = t_1 \cdot e$, So when smoke diffusion region after the maximum moment t_1 , it can be predicted moments t_2 when smoke disappear completely.

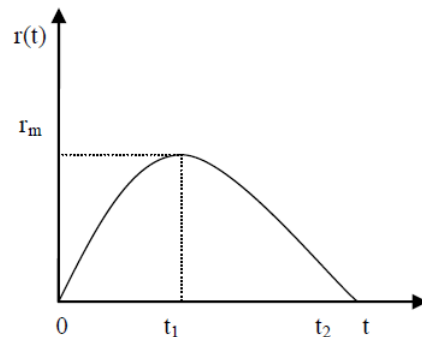


Figure 1. The radius $r(t)$ of smoke diffusion and dissipation area

We get the mathematical model of smoke diffusion and dissipation of shells' explosion; it has two advantages as following:

(1) The smoke diffusion and dissipation is obeyed to the physical rule. The results obtained only from the physical rule can not explain the process of diffusion and dissipation of smoke we observed well. we introduces an instrument indicator or sensitivity of human eye; it can explain the phenomenon above successfully and get the correct conclusions.

(2) The factors related to the smoke diffusion and dissipation has been analyzed, the light intensity variation has been researched, and the precise time of the smoke disappearing is forecast. The model has laid the foundation for the data visualization in the virtual reality platform.

E. Particle System Model

Particle system model is the most successful way to simulate dynamic irregular fuzzy objects. The basic idea is to use a large number of vitality particle elements to describe the irregular nature fuzzy scene. Particles are objects that have mass, position, and velocity, and respond to forces, but that have no spatial extent. Because they are simple, particles are by far the easiest objects to

simulate. Despite their simplicity, particles can be made to exhibit a wide range of interesting behavior. For example, a wide variety of non rigid structure scan be built by connecting particles with simple damped springs. In this portion of the course we cover the basics of particle dynamics, with an emphasis on the requirements of inter active simulation The properties like Position, velocity, color, size, and transparency of the particles changes follow certain principles. The movement of particles is controlled by combing particle system and mathematical model can overcome randomness shortcomings of the particles themselves.

(a) The generally simulating process of particle system model

(1) The basic steps to render smoke each frame according to the general procedure of particle system as following:

(2) A new smoke particle and field of force are generated;

Each one new smoke particle and field of force is given certain initial properties;

(3) The smoke particles which are over the lifetime is removed;

(4) Updating the properties of field of force, such as position and velocity according to the rules that is first set;

(5) Updating the color, size, transparency and other properties of particles according to the rule; update the position and speed properties of particles by the smoke force;

(6) Detecting the collision with outside world, and if collision occurs, the collision is eliminated;

(7) Rendering the particle to generate image.

We can get a dynamic loop process by going step (3) to (7) continuously.

(b) Our algorithm processes of Exploding

Since the explosion happened in an instant time, when the explosion occurred later no longer produce new particles, so the particle emission type can use explosive particle system model, that is, at the same time all particles launched from a region and outward spherical expansion. At beginning, the smoke expansion rapidly, then gradually slowed down until disappearing. The whole process can be simplified as following, the initial position of the smoke particles is randomly generated near the origin, the explosion smoke outward diffusion by sphere radius, and change with time to a maximum radius of the explosion, the particles initial speed is larger, and changes over time gradually until decrease; the particle size gradually increases over time. The phenomenon of fog rolling is occur when explosion, in order to simulate this phenomenon, the explosion is initially fast rotating particle texture, and then reduce the rotational speed varies with time. The same time the color of the particles need be set, and the rate of change in transparency, and gradually adjust the color and transparency of the particles in the particle movement process, when the particles extended to a maximum radius, the particle color gradually weakened, and fusion with the background, smoke disappear. The algorithm processes is shown in Figure 2.

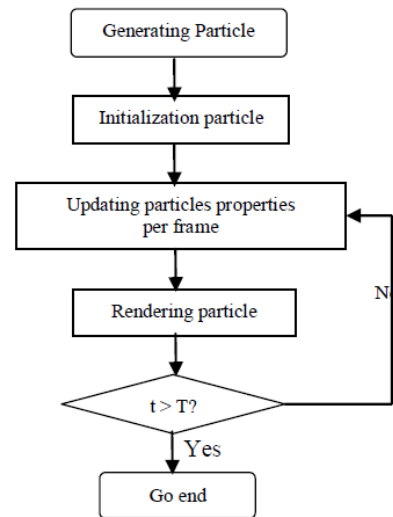


Figure 2. The algorithm processes of smoke diffusion and dissipation based on particle system

III. RENDERING OF SHELLS EXPLODING

The render of the smoke is a very important part. Particle-based simulation of smoke can be used OpenGL [6], and texture mapping can be used to improve realistic. When rendering the smoke particles, we can use the Billboard technology; it is a two-dimensional texture mapping technology [7].

A. Texture Mapping

In computer graphics, the application of a type of surface to a 3D image, a texture can be uniform, such as a brick wall, or irregular, such as wood grain or marble. The common method is to create a 2D bitmapped image of the texture, called a "texture map," which is then "wrapped around" the 3D object. An alternate method is to compute the texture entirely via mathematics instead of bitmaps. The latter method is not widely used, but can create more precise textures especially if there is great depth to the objects being textured.

A texture map is applied (mapped) to the surface of a shape or polygon. This process is akin to applying patterned paper to a plain white box. Every vertex in a polygon is assigned a texture coordinate (which in the 2d case is also known as a UV coordinate) either via explicit assignment or by procedural definition. Image sampling locations are then interpolated across the face of a polygon to produce a visual result that seems to have more richness than could otherwise be achieved with a limited number of polygons. Multi texturing is the use of more than one texture at a time on a polygon. For instance, a light map texture may be used to light a surface as an alternative to recalculating that lighting every time the surface is rendered. Another multi texture technique is bump mapping, which allows a texture to directly control the facing direction of a surface for the purposes of its lighting calculations; it can give a very good appearance of a complex surface, such as tree bark or rough concrete that takes on lighting detail in addition to the usual detailed coloring. Bump mapping has become popular in recent video games as graphics

hardware has become powerful enough to accommodate it in real-time.

B. The Representation of Particle Texture Image

Particle texture image can be generated by the algorithm or an existing image. The texture is described by using RGBA. And $R = G = B = A$, A is transparent, transparency is smaller when gray is bigger. Because the thickness of the particles descending from the center to the edge, it can be considered the gray values of texture also obey to the rule, while it is required a transition continuously. Generally this trend can be used Gaussian distribution formula to distribution, as following:

$$h(d) = \frac{\rho}{2\pi^{1/2}\sigma} \exp\left(-\frac{d^2}{2\sigma^2}\right) \quad (20)$$

where: d represents the length from the sphere center to edge; $h(d)$ represents the gray value of texture from the sphere center to distance d ; the variance σ is the Gaussian distribution variance, in order to achieve normalization (i.e. texture data is only defined in the $[-1.0, 1.0]$), σ will be taken as 3; ρ as the modulation values of central peak, the it can adjust the maximum gradation; P is in the range $[0, 2\pi^{1/2}\sigma]$, we found the peak is taken as 0.4, the simulation results would be more appropriate.

The Particle texture image produced by Gaussian distribution formula is shown in figure 3.

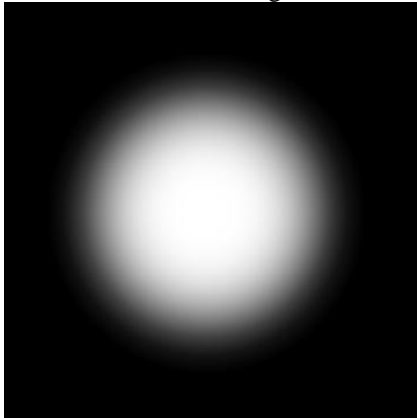


Figure 3. The Particle texture image produced by algorithm

C. Billboard Technology

The basic idea of the Billboard is using two-dimensional texture instead of the more complex three-dimensional geometric entities to improve the realism efficiency [8]. The essence of the technique of the Billboard is the plane rotating always facing the observer's viewpoint. Texture patterns are mapped by texture, affixed to the surface of the polygon on the calculated viewpoint and the relative position of the particles, and then the particles texture mapping plane is rotated, it is always perpendicular to the view vector; whether viewpoint how translational rotation, seen from the perspective of the observer, are real three-dimensional space objects, eliminating the distortion due to viewpoint transformation.

BillBoard realization lies in two places, one is the mode of texture mapping, for rendering irregularly fuzzy objects, usually with transparent textures mode, where we can use Alpha test function in OpenGL. Particle systems are generally used for the Alpha Blending function. The second key is how to make the a polygon face always toward the viewer, when rendering the particles, mainly for changing viewer cones, adjusting the normal vector of each particle in data field in dynamically, so that it is on the cone midline, thereby maintaining consistent rendering. Basic principles as follows:

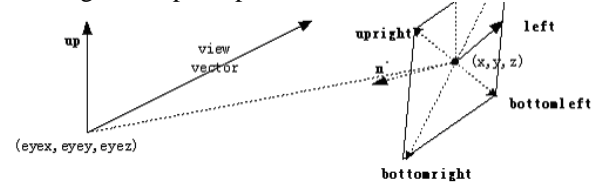


Figure 4. The normal vector of Billboard technology

Since our goal is to make the polygon maximize t to the final scene, so simplicity, we need to adjust the position of view vector coincides with the normal vector, as shown in Figure 4. Because the normal vector ultimately reflected by the coordinates of the vertices, so, as long as the four unit vectors (upleft, bottomleft, upright, bottomright) are calculated. Specific implementation steps are as follows.

$$(1) \vec{v}' = (x \quad y \quad z) - (eye_x \quad eye_y \quad eye_z);$$

$$(2) \textit{left} = \vec{v}' \times \textit{up};$$

(3) Because up is not exactly perpendicular to \vec{v}' , it needs to be corrected $up = \vec{v}' \times \textit{left}$

$$(4) \textit{upleft} = \textit{left} + \textit{up}$$

$$(5) \textit{bottomleft} = \textit{left} - \textit{up}$$

$$(6) \textit{upright} = \textit{up} - \textit{left}$$

$$(7) \textit{bottomright} = -\textit{upleft}$$

(8) Finally obtained four vertices coordinates of quadrilateral as follows:

$$(x + \textit{upleft}.x \quad y + \textit{upleft}.y \quad z + \textit{upleft}.z)$$

$$(x + \textit{bottomleft}.x \quad y + \textit{bottomleft}.y \quad z + \textit{bottomleft}.z)$$

$$(x + \textit{bottomright}.x \quad y + \textit{bottomright}.y \quad z + \textit{bottomright}.z)$$

$$(x + \textit{upright}.x \quad y + \textit{upright}.y \quad z + \textit{upright}.z)$$

Billboard is a low-level, hardware-independent technology. It can be used for solving any questions about the normal vector angle of elements changing with viewer. When rendering we have adopted following approaches: using the depth buffer, rendering quadrilateral and texture under depth buffer, and then use transparent, mixed effects and linear texture filtering can significantly improve the display effect of smoke particles.

Finally, it's worth noting that there are many different algorithms for blending colors in computer graphics. These are often referred to as "blend modes." By default, when we draw something on top of something else in Processing, we only see the top layer—this is commonly referred to as a "normal" blend mode. When the pixels

have alpha transparency (as they do in the smoke example), Processing uses an alpha compositing algorithm that combines a percentage of the background pixels with the new foreground pixels based on the alpha values.

However, it's possible to draw using other blend modes, and a much loved blend mode for particle systems is "additive." Additive blending is in fact one of the simplest blend algorithms and involves adding the pixel values of one layer to another (capping all values at 255 of course). This results in a space-age glow effect due to the colors getting brighter and brighter with more layers.

IV. CONCLUSIONS

Finally, we develop a demo system based on our algorithm and show the results generated by our simulation algorithm. The simulation algorithm of explosion effects is implemented by Visual C++6.0 and OpenGL graphics library programming. The operation system is Windows XP, Graphics card is NVIDIA NVS 3100M, memory capacity is 1G, RAM 2G. When particle number is 40000, the frame rate is 60FPS.

Experimental results show that our algorithm can produce real time simulation of explosion in consumer PC Platform and the effect of smoke looks realistic. The visual simulation result shown in Figure 5, from the simulation results, we can see the smoke behavior of explosion is accord with the true scene.

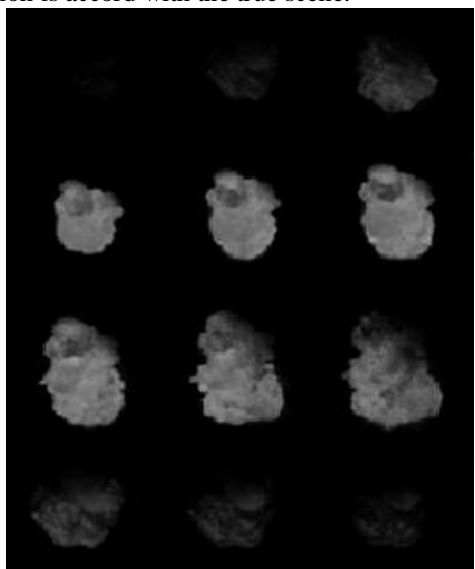


Figure 5. The explosion process based on our algorithm

This text summarized and analogized the modeling method of dynamic irregular fuzzy objects at home and abroad. With a detail introduction to the classification of the works as well as different kinds of methods employed in the field. The methods applied mainly include the particle system method, the mathematic and physics-based method and the texture based method. We studied mathematical model first, then using thought of particle systems to management the spread and dissipation of smoke. Combining the mathematical model and particle system can overcome the randomness of the particles. When rendering, with texture mapping technology to

improve simulation efficiency. The results show that the proposed algorithm is simple and practical. The next step is to consider visual effects, including dissipation, whirlpool and obstacle collisions.

The simulation algorithm based on particle system has the following advantages: can produce more realistic images; The structure of the particles are simple to set up, it is easy to modify parameters; The drawing method is flexible and real-time. The dynamic image of the explosion is produced by using particle system series method, the set of values are the preliminary observations of the particles, and dynamic behavior of the particles is controlled according to the characteristics of the object itself, and mathematical dynamic equation.

Designing and developing a real-time and efficient, simulation method and the practical software has been the people's important work, but is a difficult work. Using computer to simulate the natural phenomenon effectively, there are many problems for us to study deeply and in detail. The next step is needed to use light Tracing Ray technology to establish the illumination model when exploding, and simulate the interaction of the flame and ambient light.

REFERENCES

- [1] WeiX, "Lattice Based Natural Phenomena Modeling", *NewYork: University of Stony Brook*, 2004.
- [2] WANG Ji-zhou, GU Yao-lin, "A Survey on Flame simulation methods", *Journal of image and Graphics*, Vol. 12, No. 11, pp. 1961-1969, 2007
- [3] Zhang Qin, Xie Jun-y, i Wu Hui-zhong, etal, "Overview of methods of modeling for irregular objects", *Journal of Image and Graphics*, Vol. 5, No. 3, pp. 186~190, 2000
- [4] YAN LB, LI ST, "Research of real-time rendering and modeling of dynamic dense smoke", *computer engineering and science*, Vol. 23, No. 1, pp. 68~74, 2001
- [5] LIU Y Q, LIU X H, "Real-time 3D fluid simulation on GPU with complex obstacles", *Journal of software*, Vol. 17, No. 3, pp. 568~576, 2006
- [6] LI H J, LI K J, "The Visual Simulation of Missile' s Attack Course Based on OpenGL", *Journal of system simulation*, Vol. 16, No. 3, pp. 30-533, 2004
- [7] Zuo Lu-me, i Huang Xin-yuan, "Texture mapping and its application in 3D game engine", *Journal of Computer Simulation*, Vol. 21, No. 10, pp. 146~148, 2004
- [8] Lin Xi-we, Yu Jin-hui, "Computer synthesis of fire using particle and texture rendering", *Computer Applications*, Vol. 24, No. 4, pp. 77~79, 2004
- [9] LIU Y Q, LIU X H, WU E H, "Real-Time 3D Fluid Simulation on GPU with Complex Obstacles", *Journa of Software*, Vol. 17, No. 3, pp. 568-576, 2006
- [10] Zuo Lu-me, i Huang Xin-yuan, "Texturemapping and its application in 3D game engine", *Journal of Computer Simulation*, Vol. 21, No. 10, pp. 146~148, 2004
- [11] Losasso F, IrvingG, Guendelman E, etal, "Melting and burning solids into liquids and gases", *IEEE Transactions on Visualization and Computer Graphics*, Vol. 12, No. 3, pp. 343~352, 2006
- [12] ZHOU Shumin, SUN Yamin, LU Ling, CHEN Zhifeng, "Fire Simulation Model Based on Particle System and Its Application in Virtual Reality Proceedings", *the 16th International Conference on Artificial Reality and Telexistence*, 2006.

- [13] Xiaoyong sheng, "Research and Realization of 3D Flame Simulation Model", *Dissertation for Master Degree, East China Normal University*, 2009.
- [14] HUA Ze-xi WANG Ying-chun SUN lian-shun, Study of Explosion Simulation Based on Particle System Computer Science, Vol. 39, No. 4, pp. 278-281, 2012
- [15] Jason S, "Visual Simulation of Shockwaves", *Graphical Mod e-Is*, Vol. 71, No. 4, pp. 126-138, 2009
- [16] LI Zhe, SU Ying, NING Yun-long, "Modeling and rendering of virtual fireworks based on particle system", *Journal Of Changchun University Of Technology(Natural Science Edition)*, Vol. 31, No. 4, 2010
- [17] Chen changbo, "simulation of explosion fragmentation effect based on flow field", *computer engineering*, Vol. 36, No. 11, pp. 271-273, 2010
- [18] Meng xiaoke, "explosive discharge visual simulation and key technology based on OSG", *Computer simulation*, Vol. 27, No. 7, pp. 234-236, 2010
- [19] Lars A. Animating Physically Based on Explosions in Real-time [CJ//Proc. of the 5th International Conference on Virtual Reality, Computer Graphics, Visualization and Interaction. Af ri-graph, South Africa: [s. n. J, 2007.
- [20] Abhinav G. Explosion Simulation Using Compressible Fluids// *Proc. of the 6th IEEE Int'l Conf. on Computer Vision, Graphics and Image Processing: IEEE Press*, 2008.
- GongLin** (1980-), female, received master degree of engineering from science and technology university of Nanjing china. Currently, she is a lecturer at Zhenjiang Watercraft College. Her major research interests include virtual environment simulation and image processing. she has published nearly ten papers in related journals.
- Hu Dingjun** (1976-), male, received master degree of engineering from science and technology university of neimenggu china. Currently, he is a senior engineer at Zhenjiang Watercraft College. His major research interests include virtual environment simulation and image processing.

Reliable Transmission Protocol based on Network Coding in Delay Tolerant Mobile Sensor Network

Luo Kan

Chengdu Technological University, Chengdu 610031, China

Wang Hua

Sichuan Vocational and Technical College of Communications, Chengdu 611130, China

Shyi-Ching Liang

Chaoyang University of Technology, Taiwan, 41349

Email: jerry.scliang@gmail.com

Abstract—According to the random mobility and intermittent connectivity problems in delay tolerant mobile sensor network, the reliable transmission mechanism based on network coding is proposed in this paper. Mainly aiming at the mobility and network coding technique of sensor node in DTMSN, comprehensively considering various factors affecting the quality of service guarantee in DTMSN, the data packet is mapped to network with the cluster as units, and the optimal transmission scheme is determined according to the bit error rate and the relay forwarding node selected by opportunistic. The simulation experiment results demonstrate that compared with the direct transmission and flooding algorithm, the proposed reliable transmission mechanism has better performance in terms of bit error rate, real-time and energy efficiency.

Index Terms—Delay Tolerant Mobile Sensor Network; Data Transmission; Network Coding; Reliability; Bit Error Rate; Flooding Algorithm

I. INTRODUCTION

Wireless sensor networks have great prospect in the applications of military affairs, circumstance observation, disaster relief operation, dangerous area domination, etc., but for the special application fields, such as the information collection of underwater, animal tracking, the node has the very strong random mobility and the network has also intermittent connectivity problem. A non-connection network structure, namely, Delay Tolerant Mobile Sensor Network (DTMSN), is proposed in this paper. The inherent characteristic of DTMSN will pose a great challenge to datum transmission, which network overhead is increased, the reliability of network is reduced, and transmission delay is also prolonged. Therefore, aiming at the resource-constrained sensor node and DTMSN's inherent characteristics, an effective and reliable data transmission mechanism becomes one of the key issues in DTMSN.

In literature [4], aiming at the co-channel interference and lack of a central controller in ad hoc networks, the “cooperative and opportunistic transmission” concept is promoted, and the author propose a new data collection protocol called DRADG (Distance-aware Replica Adaptive Data Gathering protocol), which economizes network resource consumption through making use of an adaptive algorithm to cut down the number of redundant replicas of messages, and achieves a good network performance by leveraging the delivery probabilities of the mobile sensors as main routing metrics. In opportunistic network, the mobile node can establish connections with these nodes with never establishment routing relationship, so Pelusi L et al. summarize and discuss the existing datum transmission mechanism, and a reasonable approach to resolve the connectivity problem in wireless mobile Ad Hoc network is proposed. Muralidhar et al. introduce a discrete network corresponding to any Gaussian wireless network that is obtained by simply quantizing the received signals and restricting the transmitted signals to a finite precision. Since signals in the discrete network are obtained from those of a Gaussian network, the Gaussian network can be operated on the quantization-based digital interface defined by the discrete network. The transmission control protocol will perform badly in Oppnets, because the end-to-end connection always needs to be reconstructed and data need to be retransmitted from the original sender again through the reconstructed connection. To solve this problem, the author in literature [7] investigates the principles of on-time and at-one-time delivery, propose a designed adaptive transmitting platform to improve the end-to-end transmission quality of service. Byung-Gook Kim et al. investigate an opportunistic resource scheduling problem for the relay-based OFDMA cellular network where relay stations (RSs) perform opportunistic network coding with downlink and uplink sessions of a mobile station, which the probability of collisions can be

effectively reduced and resource allocation also can be optimized.

In addition, Joda. R et al. address network coding for robust transmission of sources over an orthogonal two-hop wireless network with a broadcasting relay. The network consists of multiple sources and destinations in which each destination, benefiting the relay signal, intends to decode a subset of the sources. Tae-Won Yune et al. consider a multiuser cooperative transmission scheme with network coding over wireless channels. Network coding based on decode-and-forward protocol is a promising technique to improve network throughput. In multi-hop wireless networks, Argyriou A. et al. propose a cross-layer framework for optimizing the performance of opportunistic network coding. The target scenario considers a wireless ad hoc network (WANET) with backlogged nodes and multiple unicast packet flows. The literature [12] set to define simple extensions for distributed medium access control (MAC) protocols that optimize the transmission of network-coded packets independently of the actual network coding algorithm. Next, Abouelseoud M. et al. develop a technique that expand opportunistic analysis to a broader class of networks, propose new opportunistic methods for several network geometries, and analyze them in the high-SNR regime. For each of the geometries studied in the paper, the paper analyzes the opportunistic DMT of several relay protocols, including amplify-and-forward, decode-and-forward, compress-and-forward, non-orthogonal amplify-forward, and dynamic decode-forward. Selection schemes based on the direct source-destination links are shown to achieve optimal performance. Owing to the inherent characteristic of DTMSN, it is difficult to use traditional sensor network to collect datum. Therefore, a novel data collection method named relative distance-aware data delivery scheme (RDAD) is proposed in literature [14], which the strategy introduces a simple non-GPS method with small overhead to gain the relative distance from a node to sink and then to calculate the node delivery probability which gives a guidance to message transmission. The research of the DTMSN has received great development in some extent, but there still are some problems that deserve deeper research, which have limitations in improving DTMSN performance.

Therefore, According to the existing research results, an opportunity-reliable data transmission mechanism based on network coding called as NCDTMSN which can be applied to Delay Tolerant Mobile Sensor Network is proposed by the in-depth analysis and research of the protocol mechanism of Delay Tolerant Mobile Sensor Network (DTMSN), node's mobility models and network coding algorithm. The major works of this paper are: 1) to establish an adaptive mobility-aware analysis model based on the network properties and packet structures of Delay Tolerant Mobile Sensor Network and give the analytical method of the bit error rate, packet loss ratio and energy efficiency; 2) to analyze opportunity network coding's influence on the performance of network for the mobility-aware models established in step 1) and propose

reliable transmission strategies; 3) to combine mathematical analysis with simulation experiments, analyze and verify the proposed data transmission strategies and existing mechanism for the reliability, connectivity and lifetime of network. The conclusion shows the proposed strategy can effectively improve the validity of performance for the Delay Tolerant Mobile Sensor Network.

II. THE NETWORK MOBILITY-AWARE MODEL

A. The Network Model

Suppose in DTMSN, there are n sensor nodes deployed randomly in the topology rectangular area of $s \times t$. They randomly move to different directions with a specific probability as shown in figure 1. For the ease of analysis with generality, every node in the network has the same communication radius as r , and the clustering management is applied. So every node has the chance to compete to be the cluster head. The moving velocity v and relative distance p of nodes within cluster are recorded by the cluster head node.

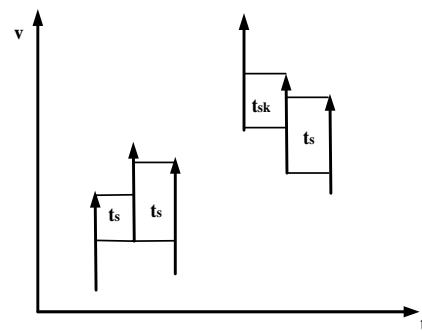
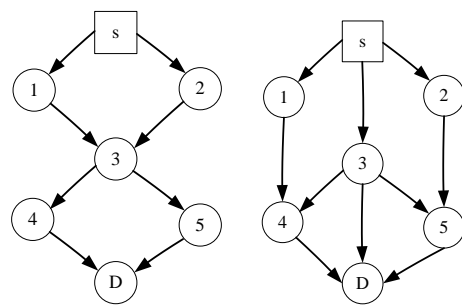


Figure 1. Nodes mobile and DTMSN network connecting model



(a) Sequenced by time delay (b) Sequenced by velocity
Figure 2. Network coding with the unit of cluster

In DTMSN network topology, nodes are always mobile, which means there is no static node. Those nodes within cluster can get the relevant data by connecting with the cluster head node in DTMSN network topology. Otherwise, for analyzing the influence of short-wave channel of the ionosphere and troposphere reflection and the environment with dense and intensive buildings on the performance of DTMSN, Rayleigh fading channel is adopted by wireless channel, which means wireless signals are transmitted through DTMSN Rayleigh fading channel and the envelope obeying Rayleigh distribution. That is to say if fading exists in the channel, each

transmitted hop is assumed to fade independently according to a Rayleigh process.

In order to analyze deeply and conveniently the changing situations of the mobility of sensor nodes and the connectivity of DTMSN network, nodes mobile network connectivity model is established as shown in figure 2. The horizontal axis is time and the vertical axis is velocity. For a certain time and velocity, the time interval of successful connection between nodes and their neighbor nodes is recorded as t . When they are disconnected, nodes can be connected with the cluster head node, and this time interval is recorded as t_{sk} .

B. The Mobility-Aware Model

DTMSN network platform uses Mica2 node. Therefore, the ber (bit error ratio) P_{bit} based on Rayleigh fading channel is given by formula (1):

$$P_{bit} = \frac{1}{2} \exp((\mu(X) - \text{var}(X) - 10\beta \lg(\frac{t_s}{t_e})) \frac{\mu(X)}{\text{var}(X)}) \quad (1)$$

where, β is the path loss parameter; $\mu(X)$ is the average of Rayleigh distribution which can be calculated by formula (2); $\text{var}(X)$ is the Rayleigh distribution variance which can be calculated by formula (3).

$$\mu(X) = t \int_0^{+\infty} x \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx = t\sqrt{2\pi}\sigma \quad (2)$$

$$\text{var}(X) = t \int_0^{+\infty} x^2 \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} dx - \mu^2(X) = t \frac{4-\pi}{2} \sigma^2 \quad (3)$$

where, T is the time interval of successful connection between nodes and neighbor nodes or the cluster head node.

In DTMSN network, when nodes are disconnected, data must be retransmitted. When other nodes come, these data can be received and forwarded again. Thus the packet loss ratio P can be calculated by formula (4):

$$P = (1 - (1 - P_{bit})^N)^{\frac{t_s}{t_e}} \quad (4)$$

where, N is the frequency of retransmission; t_e is the time interval of disconnection. So the ratio of the time interval of connection to that of disconnection, namely, $t_s:t_e$, should be considered when calculating the packet loss ratio.

Aiming at the wireless channel of DTMSN, these parameters for determining mobility-aware are $\langle P_{bit}, N, v, a, t_s, t_e, T_\eta \rangle$, where a denotes the first order derivative of nodes' velocity, so-called acceleration; T_η represents the established valid links between nodes, which are also the energy efficiency of data communication under the condition of connection.

The change of N, v, a , etc makes the variations of t_s, t_e, T_η , which cause the lifetime of DTMSN network, network throughput, reliability and other performances vary with these parameters. Therefore, the mobility-aware analysis model based on node's physical

properties and the random mobility feature of network is given by equations (5), (6) and (7).

According to this mobility-aware analysis model, the network condition can be known anytime. The network's connectivity can be improved and the influence of random mobility on network's performance can be reduced. So this model not only can enhance the quality of data transmission, but also has the ability to maximize the lifetime of DTMSN network.

$$t_s = \frac{S}{v} (1 - P) = \frac{\sqrt{(A_2 - A_1)^2 + (B_2 - B_1)^2}}{v} (1 - P) \quad (5)$$

where, A and B are 2 adjacent nodes. a_1, a_2, b_1 and b_2 are the coordinate values of two nodes which can be queried from cluster head nodes.

$$P_v = \iint_{t_s} f_{v,a}(v, a) d\sigma \quad (6)$$

where, P_v is the probability keeping the connection between mobile nodes. Function f can be obtained by Rayleigh fading distribution.

The intensity of network's connectivity $F(s)$ established by a specific probability in n nodes randomly deployed in the topology rectangular area of an $s \times t$ can be calculated by equation (7).

$$F(S) = (n-1)(1-P) \iint_{t_s} f_{v,a}(v, a) d\sigma \quad (7)$$

Based on the mobility-aware analysis model established by equations (5) to (7), the network's performance can be analyzed and evaluated.

III. THE RELIABLE TRANSMISSION MECHANISM BASED ON NETWORK CODING

A. The Analysis of Network Coding Performance

For further improvement of DTMSN network's performance, the influence of random mobility and failure connectivity on reliability must be reduced, so those data packets on network must be resent to wireless channel after network coding. Suppose the original data packet sequence is $[P_1 \dots P_k]$. To code it with opportunity linear network coding, and then send it to network and select next hop of relay node based on mobility-aware. The process is shown as follows:

On the transmitting nodes, the transmission control unit of the i th data group P_i and the waiting data group of the cluster head node are combined for network coding.

Nodes within cluster are lined up with priority by the cluster head node based on their velocity and connection probability. And step (3) will be executed due to the differences of their priority.

According to node's mobility-aware model, select a random linear network coding method from figure 3, which is to code $[P_1 \dots P_k]$ into $[P_{1X}, P_{2X}, \dots, P_{kX}]$; then establish connections and send them to new groups based on their connection probabilities.

Send those coded new groups to network and wait the next hop of relay node to receive them.

B. Reliable Transmission Mechanism

According to the above analysis, a reliable transmission mechanism NCDTMSN based on network coding, which can apply to DTMSN, is proposed in this section. The description about mobility-aware and random linear network coding mechanism algorithm on data transmission nodes with the unit of cluster and data forwarding delay nodes is shown as follows:

Input: parameters $\langle P_{bit}, N, v, a, t_s, t_e, T_\eta \rangle$

Output: the probability of network connection, lifetime and other performances parameters.

Sending codes:

(1) Send those locating information of sensors within cluster to the cluster head node.

(2) Divide those waiting data packets into different groups; and report information of the total number of groups and sizes to the cluster head node; then monitor channels and wait for the response.

(3) After receiving confirmation, regroup the waiting or forwarding data in those data groups and the cluster head nodes in order to execute random linear coding as figure 3.

Relay node

(4) Establish connections with neighbor nodes by opportunity, and the receiving data groups and its own waiting data groups execute random linear coding again as figure 3. According to the mobility-aware model established in section 1.2, the best next hop transmission node from neighbor nodes is selected.

(5) For those delay nodes in mobility-aware model, step (2) and (3) should be executed when sending data groups; and step (4) should be executed when receiving data.

(6) If data groups are correctly received, send confirmation to the cluster head node; otherwise automatically discard data packets and require the cluster head node or delay nodes to resend data groups again.

IV. THE ANALYSIS AND VERIFICATION OF PERFORMANCE

On DTMSN network, the efficiency of the proposed reliable transmission mechanism NCDTMSN can be verified by the value analysis and simulation experiment from the aspects of real time, reliability, and throughput and so on. Three transmission mechanisms are adopted: (1) Direct transfer algorithm; (2) Flooding algorithm; (3) The proposed reliable transmission mechanism NCDTMSN. The mathematical analysis and simulation experimental parameters are shown as table 1.

On DTMSN network, the throughput ϕ when data groups with random linear network coding can be correctly received after resending for N ($N \geq 0$) times is:

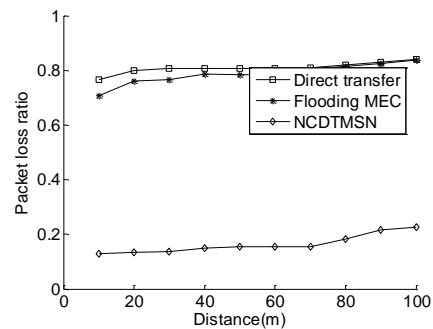
$$\phi = \frac{E(t_s)}{E(t_e) + E(t_s)} = \mu(X) \sum_{i=0}^{\infty} (1-P)^i \quad (8)$$

And the energy efficiency of NCDTMSN mechanism's transmission when DTMSN network is connected validly can be analyzed by formula (9).

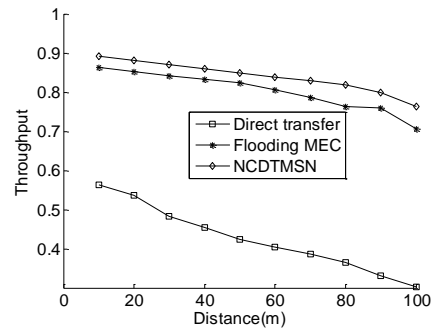
$$T_\eta = \text{var}(X) \frac{\sum_{i=0}^{t_s} \eta}{\sum_{i=0}^{t_s} \eta + \sum_{i=0}^{t_e} \eta} \quad (9)$$

TABLE I. PARAMETER SETTING

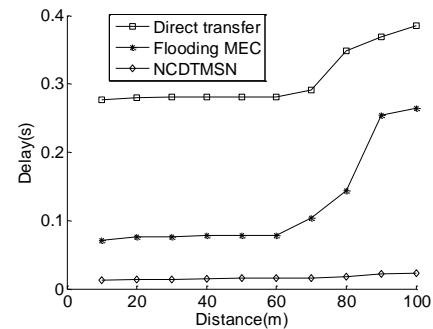
Name	Value
Network topology	1000m×100m
Number of sensor nodes	20~50
Communicating radius(m)	1~200
The velocity of sensor nodes(m/s)	1~10
The pause time of nodes mobility (s)	0
The size of data groups(byte)	512
Total resending time after network coding	0~2
The strength of network coding	0~30
Antenna model	Omni-directional



(a) Packet loss ratio



(b) Throughput



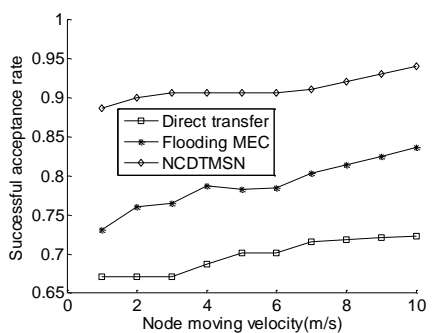
(c) Delay

Figure 3. The influence of distance on network performance

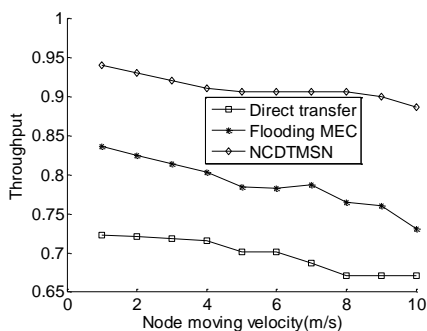
A. The Analysis of the Performance in Different Communication Radiuses

The influence of communication radius on the performance of DTMSN network is tested by the first group of experiments. Figure 4 shows the changes of control mechanisms on the aspects of packet loss ratio,

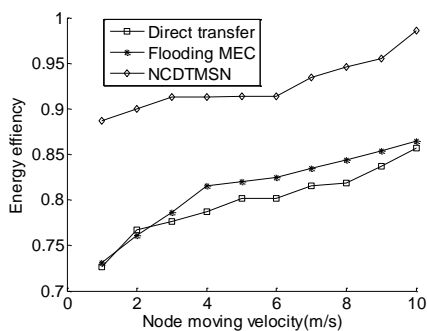
throughput and average delay and so on with different communication radius. It is apparent that the influence of mobility on system's performance is considered by the proposed NCDTMSN mechanism, and not only the throughput is improved but also the reliability of data transmission is enhanced by combining this mechanism with random linear network coding technique. When the communication radius is short, NCDTMSN can maintain a better performance than that of direct transfer and flooding algorithm while keeping a low bit error rate; when the radius is long, it also has the ability to effectively enhance system's performance. Furthermore, figure 3 (c) shows that the proposed NCDTMSN mechanism can enhance long distance communication's delay problems and offer a good guarantee for real time.



(a) The successful acceptance rate of data groups



(b) Throughput



(c) Energy efficiency

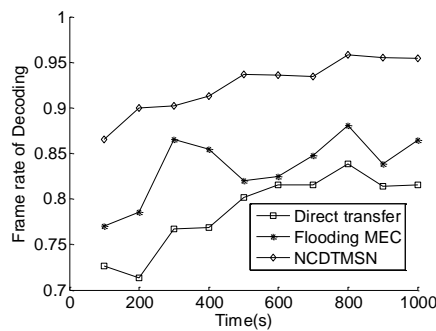
Figure 4. The influence of velocity on network's performance

B. The Influence of Velocity on Performance

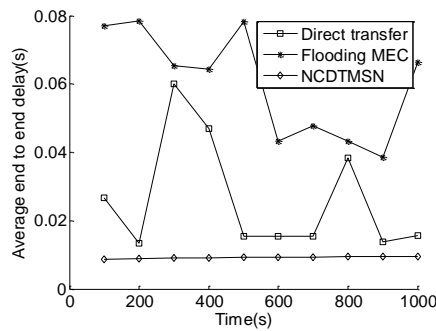
The influence of sensor nodes' velocity on the performance of DTMSN network is analyzed by the second group of experiments. The test results are shown in figure 4. It gives three performance contrasts of transmission control mechanism from the aspects of acceptance rate, throughput and energy efficiency and so

on. The performances of every aspect of NCDTMSN mechanism are all better than those of direct transmission and flooding mechanism. These improvements on performances mainly ascribe the network's connectivity based on mobility-aware and the reliable transmission based on network coding. The data quality on the receiver side will be enhanced while the network's life time is prolonged. Simultaneously, this mechanism can reduce the restrictions degree of high-velocity mobile and dynamic topology in wireless communications.

The third group of experiments considers the playback quality and real time of the multimedia data on receiver side as the main factors and processes performance contrast between three transmission control mechanisms. The result is shown as figure 5. The simulation time is 1000 seconds when playing a medium quality video. According to figure 5 (a), it is hard for the traditional direct transfer and flooding algorithm to guarantee the quality of multimedia communication, due to the intermittent connectivity caused by their nodes' uncertain velocity. And the video playback quality will be reduced on the receiver side. However, the proposed NCDTMSN mechanism has the ability to sense the mobile state of neighbor nodes or nodes within cluster based on the cluster head's location data in order to choose the best next hop transmission nodes. The network connectivity will be improved greatly, which can also make sure a reliable transmission of multimedia data for a long time while enhancing the playback quality on the receiver side. Besides, it is apparent from figure 5 (b) that comparing with the traditional direct transfer and flooding algorithm, the proposed NCDTMSN mechanism can provide a better safeguard for multimedia real time communication.



(a) The frame rate of decoding



(b) The average end-to-end delay

Figure 5. The analysis of multimedia performance

V. CONCLUSION

Owing to the inherent characteristic in delay tolerant mobile sensor network (DTMSN), such as random mobility and intermittent connectivity problems, this will cause the robustness and reliability of the data transmission cannot be guaranteed.

Aiming at this problem, a network mobility-awareness analysis model is firstly constructed. Then, the influence of opportunistic network on DTMSN performance is analyzed, so the reliable transmission mechanism based on network coding is proposed. The mathematics analysis and the simulation results show compared with the direct transmission and flooding algorithm, the proposed reliable transmission mechanism has the advantages in improving network throughput, prolonging network lifetime and extending energy efficiency.

REFERENCE

- [1] Zhang, Zhensheng. "Routing in intermittently connected mobile ad hoc networks and delay tolerant networks: overview and challenges." *Communications Surveys & Tutorials, IEEE 8.1* (2006): 24-37.
- [2] Akyildiz, Ian Fuat, Tommaso Melodia, and Kaushik R. Chowdury. "Wireless multimedia sensor networks: A survey." *Wireless Communications, IEEE14.6* (2007): 32-39.
- [3] Wang, Yu, and Hongyi Wu. "Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering." *Mobile Computing, IEEE Transactions on 6.9* (2007): 1021-1034.
- [4] Cevher, Volkan, Marco Duarte, and Richard G. Baraniuk. "Distributed target localization via spatial sparsity." *European Signal Processing Conference (EUSIPCO)*. 2008.
- [5] Zhong S, Xia K, Yin X, et al. The representation and simulation for reasoning about action based on Colored Petri Net/Information Management and Engineering (ICIME), *2010 The 2nd IEEE International Conference on. IEEE*, 2010: 480-483.
- [6] Le, D., Jin, Y., Xia, K., & Bai, G. (2010, March). Adaptive error control mechanism based on link layer frame importance valuation for wireless multimedia sensor networks. *In Advanced Computer Control (ICACC), 2010 2nd International Conference on* (Vol. 1, pp. 465-470). IEEE.
- [7] Xia K, Cai J, Wu Y. Research on Improved Network Data Fault-Tolerant Transmission Optimization Algorithm. *Journal of Convergence Information Technology*, 2012, 7(19).
- [8] XIA, K., WU, Y., REN, X., & JIN, Y. (2013). Research in Clustering Algorithm for Diseases Analysis. *Journal of Networks*, 8(7), 1632-1639.
- [9] Yao, Yufeng, Jinyi Chang, and Kaijian Xia. "A case of parallel eeg data processing upon a beowulf cluster." *Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on. IEEE*, 2009.
- [10] Kai-jian, Xia, et al. "An edge detection improved algorithm based on morphology and wavelet transform." *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*. Vol. 1. IEEE, 2010.
- [11] Wang, Yu, and Hongyi Wu. "DFT-MSN: The Delay/Fault-Tolerant Mobile Sensor Network for Pervasive Information Gathering." *INFOCOM*. 2006.
- [12] Chang, Tzu-Chien, Kuochen Wang, and Yi-Ling Hsieh. "Enhanced color-theory-based dynamic localization in mobile wireless sensor networks." *Wireless Communications and Networking Conference, 2007. WCNC 2007*. IEEE. IEEE, 2007.
- [13] Suo, Hui, et al. "Issues and challenges of wireless sensor networks localization in emerging applications." *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on*. Vol. 3. IEEE, 2012.
- [14] Gou, Haosong, Younghwan Yoo, and Hongqing Zeng. "A partition-based LEACH algorithm for wireless sensor networks." *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*. Vol. 2. IEEE, 2009.
- [15] Gou, Haosong, Younghwan Yoo, and Hongqing Zeng. "A partition-based LEACH algorithm for wireless sensor networks." *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*. Vol. 2. IEEE, 2009.
- [16] Xia K, Cai J, Wu Y. Research on Improved Network Data Fault-Tolerant Transmission Optimization Algorithm. *Journal of Convergence Information Technology*, 2012, 7(19).

Routing Optimization Based on Taboo Search Algorithm for Logistic Distribution

Yang Hongxue

Beijing Electronic Science & Technology Vocational College, Beijing 100029, China

Xuan Lingling

Zhejiang Institute of Communications, Hangzhou, 311112 China

Abstract—Along with the widespread application of the electronic commerce in the modern business, the logistic distribution has become increasingly important. More and more enterprises recognize that the logistic distribution plays an important role in the process of production and sales. A good routing for logistic distribution can cut down transport cost and improve efficiency. In order to cut down transport cost and improve efficiency, a routing optimization based on taboo search for logistic distribution is proposed in this paper. Taboo search is a metaheuristic search method to perform local search used for logistic optimization. The taboo search is employed to accelerate convergence and the aspiration criterion is combined with the heuristics algorithm to solve routing optimization. Simulation experimental results demonstrate that the optimal routing in the logistic distribution can be quickly obtained by the taboo search algorithm.

Index Terms—Taboo Search; Logistic Distribution; Electronic Commerce; Routing Optimization; Heuristics Algorithm; Aspiration Criterion

I. INTRODUCTION

Logistics is an emerging discipline and distribution is an important element of modern logistics. A good routing for logistic distribution can cut down transport cost and improve efficiency. The efficiency of it has great influence on improving whole efficiency of logistics system and reducing transport cost. The route optimization problem of logistic distribution [1] is a classic combinatorial optimization problem, which is a kind of NPC-hard problems and has high computational complexity. Due to the prosperity of market economy and the speedy development of logistic distribution industry, more and more enterprises recognize that the logistic distribution plays an important role in the process of production and sales of enterprise. The selection of traditional artificial distribution routes needs to cost much time and energy, because it completely depends on the laborers' experiences and wisdom. With the gradual increase of enterprise and business scale and the increasing number of distribution centers and the more complicated distribution routes, the artificial distribution route cannot satisfy the enterprise's business demand, so it is necessary to design routes with computers. In order to solve the problem, many algorithms with intelligence

working are developed. These algorithms can find the optimal or sub-optimal solution efficiently. Simulation experiments show that it is a good method to solve the optimization problems for logistics distribution centers. The routing optimization is developed by using aspiration criterion.

In view of the intrinsic characteristics of routing optimization, many searchers study the optimal solution of the linear programming problems, such as the supplier, manufacturer, and logistics center and so on. As we all understand the routing, from the supplier to the manufacturer and the manufacturer to the logistics centers is complicated. Therefore, the routing algorithm has integrated many neighborhood search methods, and has adopted routing decomposition to solve logistic distribution. Our proposed solution is divided into several subsets for routing optimization, each of which is solved by the search algorithm. In order to promote the convergence speed and the global search ability, there are many optimization algorithms for solving distribution routing problem, such as traveling salesman, dynamic programming, saving method, scanning, partition distribution algorithm, scheme evaluation etc. Although these algorithms can solve such problem, there are also some shortcomings, such as combination point messy and being difficult to combine the edge point for saving method, non-optimization optimization problem for scanning. According to the optimization characteristics of physical distribution routing problem, how to construct the heuristic algorithm with simple operation and good optimization performance is well worth further studying. Simulated annealing and taboo search algorithm has been successfully applied to this issue in recent years. But they also have their own problems, for example, local search ability is poor and feasible solution is not very good in the simulated annealing, and taboo search algorithm is heavily reliant on the initial solution. The hybrid algorithm is a research focus currently, which can partially offset the defects.

Taboo search algorithm is a new evolutionary algorithm, which is a new heuristic algorithm through simulating the process of swarms searching for food and is a stochastic heuristic optimization algorithm proposed by Glover [8, 9], a scholar of Italy, for the first time in 1986. Glover makes full use of the similarity between the

course of swarm searching food and the well known traveling salesman problem, and simulates the course of swarm's searching food to solve this kind of TSP questions, namely, a shortest route between nest and food source is found by the information communication and the mutual cooperation between individual.

The taboo search algorithm is applied to a variety of combinatorial optimization problems, particularly suitable for discrete optimization problem about more points' nondeterministic search in the solution space, such as traveling salesman problem (TSP), quadratic assignment problem (QAP), job-arrangement scheduling problem (JSP) etc. Furthermore, it is widely used in communication network load and water science field. In a word, it is an optimization method on the whole, which has generality and robustness.

The prototype of taboo search algorithm (TSA) is a model for searching the shortest route; therefore it has the congenital advantage in route optimization. Now we have lots of successful applicable examples of TSA on the problem of TSP, for example, Swarm-Q [11], MMAS [12] etc. There is the common ground between the problems of logistic distribution routing optimization and TSP, which search the shortest route among all customer points, but the logistic distribution routing optimization also has its characteristic, namely, it has more complicated constraint conditions and optimization objectives. According to the characteristic, this paper researches a route optimization algorithm based on taboo search algorithm, which can avoid algorithm premature and stagnation in the local search process by introducing the chromosome operator and meanwhile, improve the update mode of candidate and the strategy of selecting customer points and enhance the positive feedback effect of TSA so as to promote the convergence speed and the global search ability, which can receive a relatively good practical effect on the logistic distribution routing optimization problems.

II. MATHEMATICAL MODEL OF LOGISTIC DISTRIBUTION

A. Description of Problems

The general distribution route problems can be described as follows: there are L customer points; the quantity demand and locations of each customer point are known; it can use K vehicles at most to arrive at these demand points from the distribution center and each vehicle has a definite deadweight after completing the distribution tasks and returning to the logistic center. It is required that with the arrangement of vehicle's delivery route, the delivery distance is the shortest and at the same time the following constraint conditions can be satisfied; the sum of quantity demand of each route cannot exceed the vehicle's deadweight; the total length of each distribution route cannot exceed the maximum driving distance of each distribution. The demand of each customer point must be satisfied by only one vehicle. The purpose is to get the minimum general cost (i.e. distance, time, etc.)

B. Establishment of Mathematical Model

1) The Definition of Symbol

L : The total number of customer points;

q_i : The quantity demand of customer i , where $i=1,2,\dots,L$;

d_{ij} : The distance from customer i to j , specially, when $i,j=0$, it denotes the distribution center, for example, $d_{0,3}$ denotes the distance from the distribution center to customer point 3; $d_{2,0}$ denotes the distance from the distribution center to customer point 2; $i,j=0,1,2,\dots,L$;

K : The total number of vehicles;

Q_k : The maximum deadweight of vehicle k , where $k=1,2,\dots,K$;

D_k : The maximum driving distance of vehicle k , where $k=1,2,\dots,K$;

n_k : The total customers number distributed by vehicle k , when $n_k=0$, it denotes vehicle k does not participate in distribution, $k=1,2,\dots,K$;

R_k : A set of customer points distributed by vehicle k . When $n_k=0$, $R_k=\Phi$; when $n_k \neq 0$,

$R_k = \{r_k^1, r_k^2, \dots, r_k^{n_k}\} \subseteq \{1, 2, \dots, L\}$, where r_k^i denotes that the order of this customer point in the distribution routes is i , and $k=1,2,\dots,K$.

Constraint conditions

According to the previous description of the logistic distribution routing optimization problems, we can abstract the following constraint conditions:

1) The sum of customer points' quantity demand on each route cannot exceed the vehicle's deadweight;

$$\sum_{i=1}^{n_k} q_{r_k^i} \leq Q_k, n_k \neq 0$$

2) The total length of each distribution route cannot exceed the maximum driving distance of each distribution;

$$\sum_{i=1}^{n_k} d_{r_k^{i-1}, r_k^i} + d_{r_k^{n_k}, 0} \leq D_k, n_k \neq 0$$

3) The demand of each customer point must be satisfied by only one vehicle.

$$R_{k1} \cap R_{k2} = \Phi, k1 \neq k2$$

4) The distribution routes must cover all customer points:

$$\bigcup_{k=1}^K R_k = \{1, 2, \dots, L\}$$

$$0 \leq n_k \leq L$$

$$\sum_{k=1}^K n_k = L$$

2) The Optimization Objective Function

According to the optimization objective of logistic distribution routing optimization problems in the paper, the equation of the optimization objective function is given as follows:

$$\min \left[Z = \sum_{k=1}^K \left(\sum_{i=1}^{n_k} d_{r_k^{i-1}, r_k^i} + d_{r_k^{n_k}, 0} \right) \cdot \text{sgn}(n_k) \right]$$

$$\text{sgn}(n_k) = \begin{cases} 0, & n_k \geq 1 \\ 1, & \text{Others} \end{cases}$$

$$\rho(t+1) = \begin{cases} \max[\lambda \cdot \rho(t), \rho_{\min}] & r = r_{\max} \\ \rho(t) & otherwise \end{cases}$$

III. DISTRIBUTION ROUTE OPTIMIZATION BASED ON TABOO SEARCH ALGORITHM

A. Basic Taboo Search Algorithm

The taboo search algorithm (TSA) is a “nature” algorithm inspired by the nature creatures’ behavior, which is generated from the behavior study of swarm colony. The biggest characteristic of TSA is the indirect asynchronous contact way with “candidate” as medium of swarms in colony. When the swarms are in action, for example, searching food or finding the route back home, they will leave some chemical substances (these are called “message”). These substances can be felt by the latter swarms in the same colony and influence the latter swarms’ action as one kind of signals (it is concretely expressed that the latter coming swarms are much more possible to choose the routes with substances than these without substances.), and the messages left by the latter coming swarms will reinforce the previous candidate and such cycle will continue. In this way, the route chosen by more swarms will be more possible to be chosen by the latter coming swarms and that is because the left information is of high concentration. For the shorter route will be visited by more swarms in a certain time, there will be more left messages, and the route will be more likely to be chosen by other swarms in the next time. This process will not continue until all swarms choose the shortest route.

Vehicles can be replaced by artificial swarms to make distribution on customer points. When the swarm serving in customer point i selects the next customer point j , it should mainly consider two elements, namely, one is the intimacy between the two customer points i and j , which is called visibility η_{ij} , the other is the feasibility from i to j showed in the cycled route scheme so far, called as candidate density τ_{ij} . The probability that the swarm k will transfer from customer i to j in t time is:

$$p_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha \cdot [\eta_{ij}]^\beta}{\sum [\tau_{ik}(t)]^\alpha \cdot [\eta_{ik}]^\beta} & \text{if } j \in allowed_k \\ 0 & otherwise \end{cases}$$

where $allowed_k = \{0, 1, \dots, n-1\} - taboo_k$ denotes the customer point not yet served by the swarm k . The visibility is:

$$\eta_{ij} = \frac{1}{d_{ij}}$$

When the next customer point to be served will make the total carrying capacity exceed the vehicle deadweight or make the delivery distance exceed the maximum driving distance for one time, it will return to the distribution center and the artificial swarm will start to continue distribution in replace of the next vehicle. After a complete cycle, the swarm traverses all customer points and completes one distribution. When all swarms finish one cycle, according to the good or bad target function value, the increment of candidate will be calculated and

the candidate in the relevant route will be updated. The updating principle is:

$$\tau_{ij}(t+n) = \rho \cdot \tau_{ij}(t) + \Delta\tau_{ij}$$

$$\Delta\tau_{ij} = \sum_{k=1}^m \Delta\tau_{ij}^k$$

B. Improved Taboo Search Algorithm

The operator is the nuclear content of simulated annealing (SA). We introduce the aspiration criterion, including neighborhood, taboo list and taboo length into taboo search algorithm, in order to promote the convergence speed and the global search ability.

In SA, the primary thought of neighborhood is that the excellent individuals of parent generation can be very close to the global optimal solution and should be inherited in the filial generation and get continuously evolved. Neighborhood can save the excellent individuals of parent generation in the filial generation and avoid the loss of excellent individual in the population caused by taboo list and taboo length, etc.

In TSA, after completing the search of each generation, we reproduce the optimal solution of the present parent generation into the filial generation and make the optimal individual continue to accumulate candidate in filial generation, which can enhance the convergence speed of algorithm.

The taboo list and taboo length of SA are based on the chromosome swap. Therefore, we firstly make the swap of the logistic distribution model before introducing taboo list and taboo length.

Suppose there are L customer points and K distribution vehicles and the swap method in this paper is these L customer points will be labeled from 1 to L separately, which are all natural numbers; when the first vehicle starts from the distribution center, it will be labeled with 0, and other vehicles will be denoted as $L+1, L+2, \dots, L+K-1$ respectively. The same vehicle can make distribution for several times, therefore, the vehicle which has made distribution for over 2 times will still be denoted as $L+K, L+K+1$. When a new vehicle starts from the distribution center or the Swap is completed, that means the previous vehicle’s route is over and it should return to the distribution center. Then, one distribution can be denoted as the swap composed by 0 and natural numbers, for example, there are 6 customer points denoted from 1 to 6 separately and 3 of them should be responsible for transportation, then the Swap is:

$$0, 1, 2, 3, 7, 4, 5, 8, 6$$

The distribution routes of 3 vehicles are respectively: vehicle1 $[0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0]$, vehicle 2 $[0 \rightarrow 4 \rightarrow 5 \rightarrow 0]$, vehicle 3 $[0 \rightarrow 6 \rightarrow 0]$.

Another example:

$$0, 1, 2, 3, 8, 4, 5, 9, 6$$

The distribution routes are: vehicle 1 $[0 \rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 0]$, vehicle 3 $[0 \rightarrow 4 \rightarrow 5 \rightarrow 0]$, vehicle 1’s second distribution $[0 \rightarrow 6 \rightarrow 0]$.

C. Taboo List

Taboo list is the operation of SA which can enhance the group diversity and avoid algorithm premature and

stagnation. Introducing the taboo list into TSA can effectively enhance the search space and avoid the algorithm's falling into the trap of the local optimal solution.

After the search of each generation in TSA, we implement the Swap and taboo list between the optimal solution and the suboptimal solution and the taboo list principle is as follows:

1) Suppose the two groups of Swap are respectively S1 and S2. The length and the start position of taboo list section will be generated randomly at first;

2) Find the taboo list section between S1 and S2 and suppose S1: $P_1|P_2|P_3$, S2: $Q_1|Q_2|Q_3$, and P_2 and Q_2 are respectively the taboo list sections between S1 and S2; insert Q_2 into S1, which is placed in front of P_2 , and then get the new Swap S3: $P_1|Q_2|P_2|P_3$;

3) In S3, delete the repetitious particle of P_1 , P_2 , P_3 against Q_2 and then get the taboo list particle S3;

4) Use the same method upon S2, and get the new particle S4;

5) Compare the results of S1, S2, S3, S4 and select the two groups of optimal particle to be saved.

Taboo length is an evolution method to enhance the population's diversity. The moderate taboo length can not only maintain the individual diversity in the population but also improve the algorithm efficiency.

In TSA, after the taboo list, the taboo length operation will be executed in the excellent individual in the population, and the operation method is:

1) Generate the taboo length times N randomly;

2) Generate two different natural numbers n_1 , $n_2 > 1$ randomly (The first number will not change in order to make sure the Swap starts from the logistic center.);

3) In the particle S of the optimal individual, exchange the particle in positions number n_1 and n_2 with each other;

4) Repeat the step 2) and 3) for N times and generate the new particle S';

5) Compare the results of S and S' and save the optimal solution.

After improved by the introduced SA, TSA has got improvement in the convergence speed and the global search ability. Then we will improve the candidate's updating way and the strategy of selecting customer point in the following in order to improve the TSA's self-adaptability.

D. Selection of the Candidate Transfer Parameter ρ

According to the basic TSA, ρ is a constant. If ρ is too large, it will relatively reduce the probability of the selected route which has not been searched, which will influence the global search ability; if ρ is too small, it will influence the algorithm's convergence speed. Therefore, we will make some appropriate adjustments on ρ in the improved algorithm. In the initial period of algorithm, we hope to find the sub-optimal solution of the algorithm as soon as possible, so ρ should be relatively large in order to increase the influence of information concentration and promote the algorithm's convergence speed; When the algorithm is in stagnation, we should decrease ρ in order to reduce the influence of candidate on swarm colony and

increase the swarm colony's search ability in the solution space and get out of the constraint of the local optimal solution.

$$\rho(t+1) = \begin{cases} \max[\lambda \cdot \rho(t), \rho_{\min}] & r = r_{\max} \\ \rho(t) & otherwise \end{cases}$$

In the equation above, r denotes the cycle times where it has not been in evolution; r_{\max} is a constant; $\lambda \in (0,1)$ is also a constant which controls the attenuation speed ρ ; ρ_{\min} is the minimum value of ρ in order to avoid the too small value of ρ will influence the convergence speed. When r gets to a preset value r_{\max} , we will reduce ρ and then r will be calculated again; the cycle will not continue until ρ gets to the preset minimum value ρ_{\min} .

F. Selection of the Deterministic Search and the Explorative Search

Convergence acceleration is to make evolution as fast as possible so as to get a better solution on the basis of having got the near-optimum solution. Because the TSA is a heuristic algorithm, the unceasing "exploration" is a necessary method to evolve for TSA; it is only the "exploration" that restrict the convergence speed of TSA, for example, when the algorithm gets a sub-optimum solution which can be likely to be further evolved, but the "exploration" scope is very large, which relatively reduces the probability to choose this route for swarms, then the candidate concentration in this route will be gradually attenuate and this route is gradually "forgot".

The value of q_0 is also discussed. When $q < q_0$, the algorithm adopts the deterministic search and the swarm chooses the shortest route with the probability q_0 at this moment; when $q \geq q_0$, the algorithm adopts the explorative search and the swarm will choose the route randomly with the probability $1 - q_0$. In the initial iteration of the algorithm, q_0 chooses the relatively large initial value and make the deterministic search with relatively large probability, which can accelerate the speed to find the local near-optimum route; in the middle range of the algorithm, q_0 chooses the relatively small initial value and executes the explorative search with relatively large probability, which can enlarge the search space; In the later stage of algorithm, q_0 returns to the initial value in order to accelerate the convergence speed.

Combining with the improved TSA, the algorithm flow chart of the logistic distribution routing optimization problems based on the improved TSA is shown as follows.

IV. EXPERIMENT AND CALCULATION

The literature [13] adopts the improved SA to get the solution of the logistic distribution routing optimization problems, and we will use the examples in this literature to calculate the result and compare their performance.

Example 1: a certain distribution center uses 2 vehicles to make distributions for 8 customers. Suppose the deadweight of vehicle is 8,000kg and the maximum driving distance for each time is 40km. The distances between distribution center and customer and the distances among customers are in the following table (0

denotes the distribution center, 1~8 denote the ID of 8 customer points, respectively.):

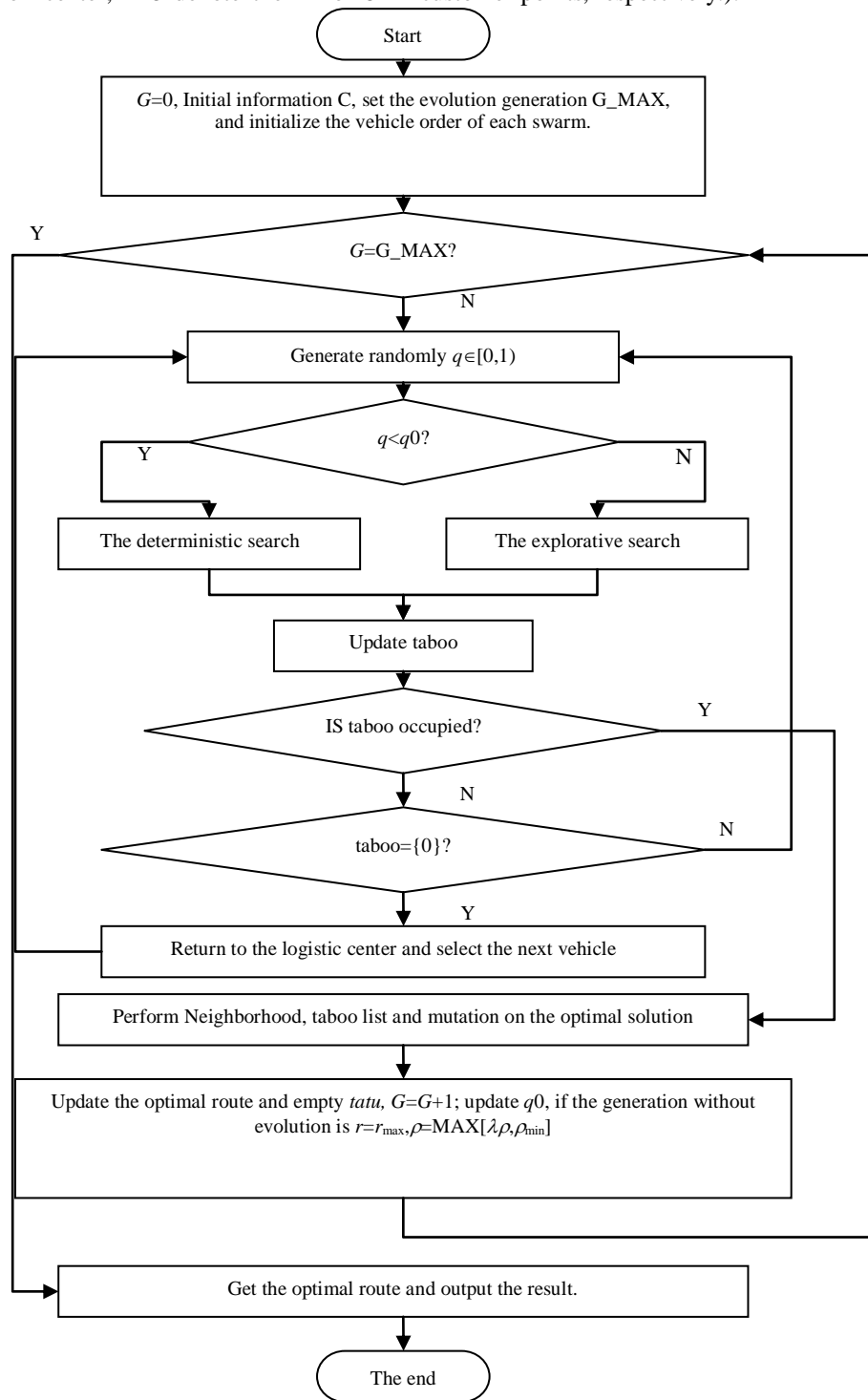


Figure 1. The algorithm flow chart

TABLE I. THE DISTANCES AMONG THE DISTRIBUTION CENTER OR CUSTOMERS(UNIT: KM)

	0	1	2	3	4	5	6	7	8
0	0	4	6	7.5	9	20	10	16	8
1	4	0	6.5	4	10	5	7.5	11	10
2	6	6.5	0	7.5	10	10	7.5	7.5	7.5
3	7.5	4	7.5	0	10	5	9	9	15
4	9	10	10	10	0	10	7.5	7.5	10
5	20	5	10	5	10	0	7	9	7.5
6	10	7.5	7.5	9	7.5	7	0	7	10
7	16	11	7.5	9	7.5	9	7	0	10
8	8	10	7.5	15	10	7.5	10	10	0

TABLE II. THE COORDINATES OF CUSTOMERS' POSITION AND THE QUANTITY DEMAND

D	X axes coordinate(km)	Y axes coordinate(km)	quantity demand(T)	ID	X axes coordinate(km)	Y axes coordinate(km)	quantity demand(T)
1	12.8	8.5	0.1	11	6.7	16.9	0.9
2	18.4	3.4	0.4	12	14.8	2.6	1.3
3	15.4	16.6	1.2	13	1.8	8.7	1.3
4	18.9	15.2	1.5	14	17.1	11.0	1.9
5	15.5	11.6	0.8	15	7.4	1.0	1.7
6	3.9	10.6	1.3	16	0.2	2.8	1.1
7	10.6	7.6	1.7	17	11.9	19.8	1.5
8	8.6	8.4	0.6	18	13.2	15.1	1.6
9	12.5	2.1	1.2	19	6.4	5.6	1.7
10	13.8	5.2	0.4	20	9.6	14.8	1.5

The results of 10 times calculation are as follows:

TABLE III. THE CALCULATION RESULTS AND THE CONCRETE SCHEME

ORDER	1	2	3	4	5	6	7	8	9	10
The total distance	113.0	109.6	110.2	111.7	110.4	111.2	109.1	109.6	107.8	110.4

The average calculation result for ten times is 110.3083km, which is higher than the average result 122.0km in literature[6] and 112.5km in literature[14] and the optimal solution is 107.84km. The corresponding concrete scheme is:

- 0→4→3→17→11→20→0
- 0→8→19→15→16→13→6→0
- 0→5→14→2→12→9→10→7→1→0
- 0→18→0

Comparing with the optimal solution 108.6 km in literature [14], it has also got improved. The optimal result figure of this paper is shown as follows:

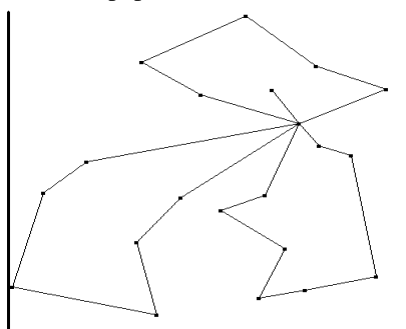


Figure 2. The optimal trajectory result

V. CONCLUSION

According to the characteristics of the logistic distribution routing optimization problems, an optimization route algorithm based on taboo search algorithm is proposed. Through introducing aspiration criterion, this algorithm can avoid algorithm premature and stagnation in local search process and meanwhile, improve the candidate's updating way and the strategy of selecting customer point and enhance the positive feedback effect of taboo search algorithm so as to promote the convergence speed and the global search ability. The experimental results show that the improved taboo search algorithm can get the optimal solution or the optimal approximate solution of the logistic distribution routing optimization fast and effectively. The research in this paper has a certain reference value for the study of

taboo search algorithm and logistic distribution routing optimization problems.

REFERENCES

- [1] Donati, Alberto V., et al. "Time dependent vehicle routing problem with a multi swarm colony system." *European journal of operational research* 185.3 (2008) pp. 1174-1191.
- [2] Meuleau, Nicolas, and Marco Dorigo. "Swarm colony optimization and stochastic gradient descent." *Artificial Life* 8.2 (2002) pp. 103-121.
- [3] Silva, Carlos A., et al. "Distributed supply chain management using swarm colony optimization." *European Journal of Operational Research* 199.2 (2009) pp. 349-358.
- [4] Zhen, Tong, et al. "Hybrid taboo search algorithm for the vehicle routing with time windows." *Computing, Communication, Control, and Management, 2008. CCCM'08. ISECS International Colloquium on*. Vol. 1. IEEE, 2008.
- [5] Jianrong, Liu Lin Zhu. "Study of the Optimizing of Physical Distribution Routing Problem Based on Mixed Swarms Algorithm." *Computer Engineering and Applications* 13 (2006) pp. 061
- [6] Zhong S, Xia K, Yin X, et al. The representation and simulation for reasoning about action based on Colored Petri Net// *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*. IEEE, 2010 pp. 480-483.
- [7] Le, D., Jin, Y., Xia, K., & Bai, G. (2010, March). Adaptive error control mechanism based on link layer frame importance valuation for wireless multimedia sensor networks. *In Advanced Computer Control (ICACC), 2010 2nd International Conference on* (Vol. 1, pp. 465-470). IEEE.
- [8] Xia K, Cai J, Wu Y. Research on Improved Network Data Fault-Tolerant Transmission Optimization Algorithm. *Journal of Convergence Information Technology*, 2012, 7(19).
- [9] XIA, K., WU, Y., REN, X., & JIN, Y. (2013). Research in Clustering Algorithm for Diseases Analysis. *Journal of Networks*, 8(7), 1632-1639.
- [10] Zhuojun, Li. "Mixed Taboo search algorithm Solving the VRP Problem." *Journal of Wuhan University of Technology (Transportation Science & Engineering)* 2 (2006) pp. 033.
- [11] Yao, Yufeng, Jinyi Chang, and Kaijian Xia. "A case of parallel eeg data processing upon a beowulf cluster." *Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on*. IEEE, 2009.
- [12] Kai-jian, Xia, et al. "An edge detection improved algorithm based on morphology and wavelet transform." *Computer*

- and Automation Engineering (ICCAE), 2010 the 2nd International Conference on.* Vol. 1. IEEE, 2010.
- [13] Jianrong, Liu Lin Zhu. "Study of the Optimizing of Physical Distribution Routing Problem Based on Mixed Swarms Algorithm." *Computer Engineering and Applications* 13 (2006) pp. 061.
- [14] Yoo, Terry S. *Insight into images: principles and practice for segmentation, registration, and image analysis.* Vol. 203. Wesley^ eMassachusetts Massachusetts: AK Peters, 2004.
- [15] Christensen, Gary E., and Hans J. Johnson. "Consistent image registration." *Medical Imaging, IEEE Transactions on* 20.7 (2001) pp. 568-582.
- [16] Gonzalez, Rafael C., Richard E. Woods, and Steven L. Eddins. *Digital image processing using MATLAB.* Vol. 2. Tennessee: Gatesmark Publishing, 2009.
- [17] Guizar-Sicairos, Manuel, Samuel T. Thurman, and James R. Fienup. "Efficient subpixel image registration algorithms." *Optics letters* 33.2 (2008) pp. 156-158.
- [18] Transformation-Alpert, Principal Axes. "The principal axes transformation-a method for image registration." *J Nucl Med* 31 (1990) pp. 1717-1723.
- [19] Rohde, Gustavo K., Akram Aldroubi, and Benoit M. Dawant. "The adaptive bases algorithm for intensity-based nonrigid image registration." *Medical Imaging, IEEE Transactions on* 22.11 (2003) pp. 1470-1479.

Opportunistic Cooperative Reliable Transmission Protocol for Wireless Sensor Networks

Hua Guo

College of information science and engineering, Shandong University of Science and Technology, Qingdao 266510, China

Email: Guohua197701@163.com

Yu Sheng-Wen

Institute of surveying and mapping, Shandong Science and Technology University Shandong Qingdao 266510, China

Douglas Leith

Hamilton Institute, National University of Ireland Maynooth, Ireland

Email: doug.leith@nuim.ie

Abstract—The high reliability requirement of data transmission in wireless sensor networks has been a large challenge due to the random mobility and network topology instability, as well as restriction of energy consumption and system life time. In this work, the relationship matrix of distance and speed was created at first. Then the binary tree of mobility and reliability was proposed based on the distance relationship matrix. The optimal relay nodes with high reliability were selected by seeking the binary tree. The high reliable opportunistic cooperative data transmission scheme was presented finally. The outcome of both mathematical analysis and NS simulation indicate that the proposed mechanism is superior to the traditional data transmission mechanism such as the reliability, system throughput and energy efficiency and so on.

Index Terms—Wireless Sensor Networks; Relay Selection; Opportunistic Cooperative Communication; Reliability

I. INTRODUCTION

Wireless sensor network is a special kind of wireless communication network [1] [2], which is considered the most important technology because of its huge number of nodes, low cost, can be deployed quickly and does not depend on any fixed infrastructure, real-time, accurate and comprehensive collection of information manner on many occasions. Wireless sensor network has been changing the way of people interacting with the nature [3] [4].

There are many research results [5] [6] [7] [8], which are mainly about some important issues such as architecture design, routing protocol design, energy conservation algorithm, locationing schemes and so on. Because wireless sensor network is very different from traditional wireless network or wireless Ad Hoc network, how to guarantee the reliable data transmission is still wanted largely explored research [9]. Since, it is important to describe the reliable control algorithm and develop guarantee approaches for reliability in wireless sensor network.

In order to address the reliability requirement, Akan O. B. et. al. [10] proposed a new reliable transmission approach scheme, which is an event-to-sink reliable transport protocol developed to achieve reliable event detection with minimum energy expenditure and includes a congestion control scheme, as well as achieving reliability and conserving energy in wireless sensor networks. Literature [11] proposed a novel routing protocol for wireless industrial sensor networks, which provides real-time, reliable delivery based on energy awareness and estimates the energy cost, delay and reliability of a path from the node to the sink node. Hao Jiang et. al [12] proposed a braided reliable cooperative reliable transport algorithm, which maintains one-hop reliability to ensure end-to-end reliability for guaranteeing the robust, as well as to adjust the data rate to improve the sensing reliability. In literature [13], the author examined the reliability and security and discussed the unique characteristics to elaborate on security and reliability, as well as analyzed the traditional architectures and standards of wireless sensor networks.

At the same time, Bletsas A. et. al [14] analyzed the performance of some opportunistic relaying protocols according to employ simple packet level feedback and strictly orthogonal transmissions and obtained the result, which gave the diversity-multiplexing tradeoff of the proposed protocols either matches or outperforms of the multi-input-single-output zero-feedback performance. The article [15] proposed an energy-efficient hybrid opportunistic cooperative transmission protocol for single-carrier frequency division multiple access cooperative networks, which improves the energy efficiency by selecting the most energy-efficient cooperative transmission protocol according to current channel state information. An opportunistic energy-efficient cooperative communication method was presented by Amin O. et. al [16], which is different from classical relaying selection schemes and adopt either end-to-end signal-to-noise ratio or capacity as selection

metrics, as well as an energy efficiency metric for selecting the best relay or resorting to direct transmission.

Zhiguo Ding et. al [17] studied the channel state information based on the architecture of cooperative transmission protocols, and cooperative diversity comes at the price of the extra bandwidth resource consumption. An opportunistic and cooperative algorithm was proposed in literature [18], which takes advantage of topology and architecture in wireless ad hoc networks in order to research the multiuser or spatial diversity for supporting different transmission priorities, improving real time performance, as well as ensuring fair transmissions among nodes. In article [19], the authors presented a new high-throughput, reliable multicast protocol for wireless Mesh networks, which integrates four building blocks—namely, tree-based opportunistic routing, intraflow network coding, source rate limiting, and round-robin batching—to support high-throughput, reliable multicast routing. Hemachandra K.T. and Beaulieu N.C.A [20] researched the closed-form lower bound which is derived for the outage probability when the relay and the destination nodes are affected by interference simultaneously and derived the asymptotic outage probability results to obtain useful insights on the effects of interference and feedback delay. Above all, it is necessary to explore and design the cooperative scheme and opportunistic technology for supporting reliable communication in wireless sensor networks.

On the basis of the above researches, we study opportunistic cooperative reliable transmission protocol, which consider the opportunistic approach and relay selection schemes for ameliorate the system performance of wireless sensor networks. First, the relationship matrix of distance and mobile speed between sensor nodes was created. Second, based on the distance relationship matrix, the binay tree of mobility and reliabiliy was proposed. Third, the optimal relay nodes with high reliabiliy were selected by seeking the binay tree. The high reliable opportunistic cooperative data transmission scheme was presnted finally. The simulation and mathematics analyses show that the proposed protocol obtain better perform than other traditional mechanisms such as throughput, reliability, data delivey ratio, averagy delivery delay and system energy consumption significantly.

The rest of the paper is organized as follows. Section 2 describes the system models and mobility relationship matrix. Section 3 proposes the binay tree struct and opportunistic relay selection approach and reliable data transmission mechanism (BTOC), followed by the details of implementation. Experiment results are given in Section 4. we give the conclusion of this work in Section 5 finally.

II. MOBILITY AWARE COOPERATIVE PROTOCOL

A. System Model

Based on the existing study results [21] [22], the system model with wireless sensor network is composed of several sensor nodes, which store the location information of adjacent nodes and maintain the routing

information of the network. In the routing information table, the packet reception ratio would be obtained and updated by flooding the beacon packets or link quality indicator, as well as the distance measurement technology such as received signal strength indicator.

As shown in Figure 1, node S send data packets to node D, R set contains several relay nodes, which are nearer to receiving node D than node S. Let D_{Si} denote the one hop distance from node S to relay node R_i selected form Rset, which could be calculated by formula (1) or (2).

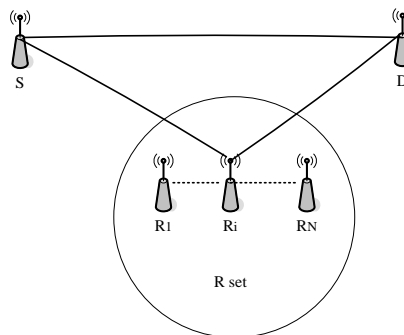


Figure 1. Network model

$$D_{Si} = dis(S, D) - dis(i, D) \tag{1}$$

$$D_{Si} = \sqrt{(x_i, y_i) - (x_s, y_s)} \tag{2}$$

Here, function $dis()$ is used to calculate the one hop distance information between one node and another node. If the sending node S, Relay node and receicing node D are at one line, the one hop distance is obtian by formula (1).

Otherwise, we use formula (2) calculate the one hop distance. (x_i, y_i) denote the location information of relay node R_i , (x_s, y_s) denote the location information of sending node S.

We define the D_{Si} and bit error rate P_b as channel state. Every sensor node and its neighbour nodes stroe the channel state $\langle D_{Si}, P_b \rangle$ parameters. Let C set denote the candidate relay nodes, which are one sensor node's neighbour nodes. So, C set is subset of R set.

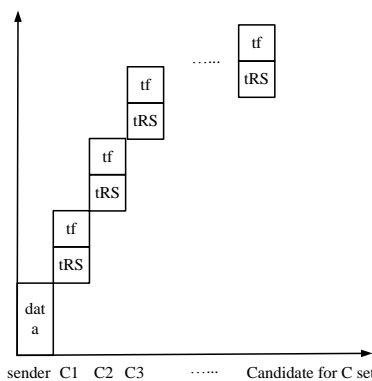


Figure 2. Delay composition of candidate node

Let t_i denote the one hop medium delay of some relay node, which is in the C set. That means data sent by

sending senso node to one relay node, which declare that it has received the data packets and would forward them. Hence, t_i is comprised of two parts. One part is the receiving delay, another is the opportunistic cooperative forwarding delay. The first part is defined by data packets transmission delay and broadcasting delay. The second delay is decided by opportunistic relay selection delay and forwarding delay. As shown in Figure 2, t_i can be calculated by formula (3).

$$t_i = \sum_{j=0}^i (t_{RS} + t_f) \quad (3)$$

where, let t_{RS} be the delay of opportunistic relay selection. The forwarding delay t_f is determined by the channel state of wireless sensor networks.

B. Mobility Aware Relayship Matrix

According to the research results of bit error rate and delay, the performance characteristics of wireless sensor networks with mobility is researched and evaluated by some relationship matrixs.

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} V_{t_1} \\ \vdots \\ V_{t_n} \end{bmatrix} - \begin{bmatrix} V_1 \\ \vdots \\ V_n \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} \quad (4)$$

$$\begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix} = \begin{bmatrix} V_1 \\ \vdots \\ V_n \end{bmatrix} + 0.5 \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} \quad (5)$$

Here, matrix $[a_1, a_2, \dots, a_n]$ is the accelerated speed of every sensor node in the wireless sensor networks. Matrix $[V_1, V_2, \dots, V_n]$ is the mobile initialization speed of sensor nodes. Matrix $[V_{t_1}, V_{t_2}, \dots, V_{t_n}]$ is the mobile speed of sensor nodes at t_i ($0 < i < n$) time. The accelerated speed matrix could be obtained by formula (4). Matrix $[D_1, D_2, \dots, D_n]$ is the distance after t_i ($0 < i < n$) time.

So, combine the formula (4) and (5), the distance matrix could be calculated.

If we can get the mobile speed information and location information, the accelerated speed matrix should be created with algorithm as shown in figure 3. The process of relay selection is as follows:

- (1) Obtain initialization speed values of every sensor node with real-time detection.
- (2) denote the initialization speed as matrix V_0 .
- (3) Analyze the matrix tRS values of every sensor node.
- (4) Gain the matrix tf values of every sensor node.
- (5) If sensor nodes (N_1, N_2, \dots, N_i) is subset of R set,
- (6) Calculate the delay t_i by formula (3).
- (7) Obtain the speed values of every sensor node with real-time detection at t_i time.
- (8) Calculate the accelerated speed matrix $[a_1, a_2, \dots, a_n]$

The pseudocode for the proposed relay selection approach is summarized as Algorithm 1.

Algorithm 1: Algorithm of accelerated speed matrix

- Input:** R set
Output: C set
- (1) obtain initialization speed values of every sensor node with real-time detection.
 - (2) denote the initialization speed as matrix V_0 .
 - (3) analyze the matrix tRS values of every sensor node.
 - (4) Gain the matrix tf values of every sensor node.
 - (5) If sensor nodes (N_1, N_2, \dots, N_i) is subset of R set,
 - (6) for $j=0 \dots i$
 - (7) $sum = tRS + tf$;
 - (8) **endfor**
 - (9) **endif**
 - (10) obtain the speed values of every sensor node with real-time detection at t_i time.
 - (11) calculate the accelerated speed matrix $[a_1, a_2, \dots, a_n]$
 - (12) output C set and matrix a

The distance matrix should be created with algorithm as shown in Algorithm 2. The process of relay selection is as follows:

- (1) obtain initialization speed values of every sensor node with real-time detection.
- (2) denote the initialization speed as matrix V_0 .
- (3) obtain accelerated speed matrix $[a_1, a_2, \dots, a_n]$ by algorithm as shown in Figure 3.
- (4) Gain the matrix t_i values with formula (3).
- (5) calculate the distance matrix $[D_1, D_2, \dots, D_n]$

Algorithm 2: Algorithm of distance matrix

- Input:** R set
Output: C set
- (1) obtain initialization speed values of every sensor node with real-time detection.
 - (2) denote the initialization speed as matrix V_0 .
 - (3) obtain accelerated speed matrix $[a_1, a_2, \dots, a_n]$ by algorithm as shown in Figure 3.
 - (4) Gain the matrix t_i values with formula (3).
 - (5) If sensor nodes is from C set,
 - (6) for $i=0 \dots n$
 - (7) $D_i = (V_0 + 0.5a_i * t_i) t_i$;
 - (8) **endfor**
 - (9) **endif**
 - (10) calculate the distance matrix $[D_1, D_2, \dots, D_n]$
 - (11) output C set and matrix D

According to formula (1), (2), (3), (4) and (5), the mobility aware relationshipa matrix would be set up, which would be used to create the wireless sensor network model.

III. OPPORTUNISTIC COOPERATIVE SCHEME BASED ON BINAY TREE

A. Binay Tree Based on Mobility

On the basis of section II, we use biny tree structure to define the sensor node with mobility. The following is the binary tree definition.

There are n nodes in the set.

There are only two child nodes and the order of them can not be changed.

The structure would be stored and constructed by link structure.

Hence, the end to end data transmission routing in wireless sensor network could be described as one binay tree, which is shown in Figure 6.

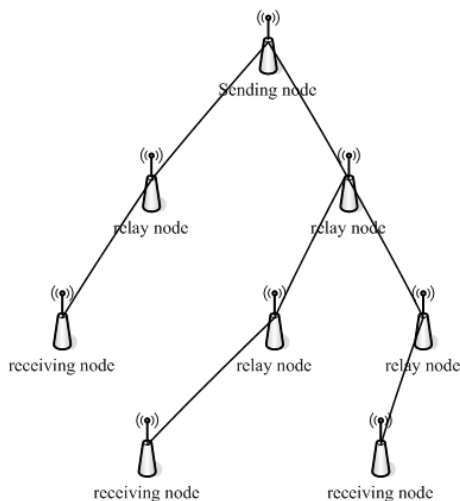


Figure 3. Structure of Binay Tree in wireless sensor network

In Figure 3, there are one sending data node, some relay nodes and receiving node. The source node sends data to one of the receiving nodes baed on the routing path, which could be created with the traversing algorithms as shown in Algorithm 3 and 4.

```

Algorithm 3: Non-recursive algorithm of preorder traverse
void PreorderTraverse(TriWsnNode p){
TriWsnNode stack [MAX],q;
int top=0,i;
for(i=0;i<MAX;i++) stack [i]=NULL;
q=p;
while(q!=NULL){
printf("%c",q->data);
if(q->rchild!=NULL) stack [top++]=q->rchild;
if(q->lchild!=NULL) q=q->lchild;
else
if(top>0) q=stack [--top];
else q=NULL;
}
}

```

```

Algorithm 4: Recursive algorithm of preorder traverse
void PreOrderTraverse(TriWsnNode p){
if ( p!= NULL ) {
Visit (TriWsnNode);
PreOrder ( p->lchild );
PreOrder ( p->rchild );
}
}

```

The Non-recursive algorithm of preorder traverse is given by Algorithm 3, the implementation workflow of which is as follows:

- Step (1): Set the node stack pointer to the current root.
- Step (2): If the node is not empty, to access the node.
- Step (3): If the right tree node is not empty, then the right tree node would be push into the stack.
- Step (4): Current pointer to the left tree node repeat (2) - (3) until the child is left empty.
- Step (5): Turn back the stack, the stack pointer to the current node.
- Step (6): If the stack is not empty or the current pointer is non-empty, continue (2), until the end of the binay tree of the wireless sensor networks.

The pseudocode for the recursive algorithm of preorder traverse is summarized as Algorithm 4.

B. Opportunistic Relay Selection

The sending sensor node would be as the root node of the binay tree. Then, the data transmission routing and relay selection nodes would be initialized based on opportunistic coopretive scheme. According to the dynamic topology and mobility, the optimal rouring and relay nodes have to be updated adaptively with the above mentioned analytical results and models.

```

Algorithm 5: Recursive algorithm of preorder traverse
void InitializeTriWsnTree(TriWsnTree *t){
Location s (x, y);
TriWsnTree q;
printf("\nplease input data:(exit for 0)");
s (x, y)=getLocation (x, y);
if(Location(x, y)==null){ *t=NULL; return;}
q=(TriWsnTree) malloc (sizeof (struct TriWsnTree));
if(q==null){
exit(0);}
q->data=s(x, y);
*t=q;
createBiTree(&q->LHopNode);
createBiTree(&q->RHopNode);
}

```

Algorithm 5 shows the creating progress of the binary tree based on PreorderTraverse algorithm detailed on Algorithm 4 or 5. Here, the receiving nodes have to be as the leaf nodes because the data transmission routing is used to be transport data from source node them.

In addition, the optimal forwarding routing path could be created with the following algorithm, which is dealt with the receiving nodes. The main idea of the algorithm based on binary tree is as follows and Algorithm 6. The hops of opportunistic cooperative data transmission could be calculated by the recursive algorithm, which is given by the Algorithm 7.

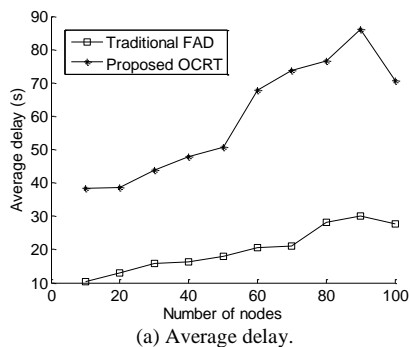
Algorithm 6: Optimal opportunistic coopeartive path creation algorithm

```

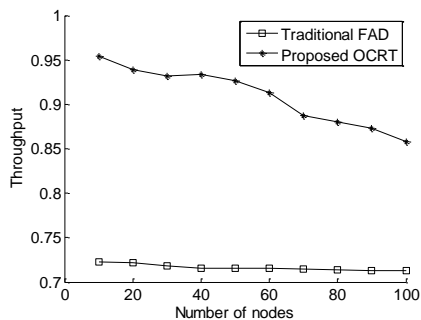
typedef struct OptimalBinayTreeWsn
{
float Pb;
OptimalBinayTreeWsn PreNode, LHopNode,
RHopNode, RelayNode;
}
void OpAU(OptimalBinayTreeWsn T)
{
for(i=0;i<n;i++)
{
TNode.PreNode=-1;
TNode.LHopNode=TNode.RHopNode=-1;
}
for(i=0;i<n;i++)
{
TNode.Pb;
}
for(i=n;i<m;i++)
{
for(j=0;j<=i-1;j++)
{
if(TNode.PreNode!=-1) continue;
if(TNode.Pb <= 0.05)
{
RelayNode=TNode;
TNode=TNode.LHopNode;
}
else if(TNode.Pb <= 0.1)
{
RelayNode=TNode;
}
}
}
}
}

```

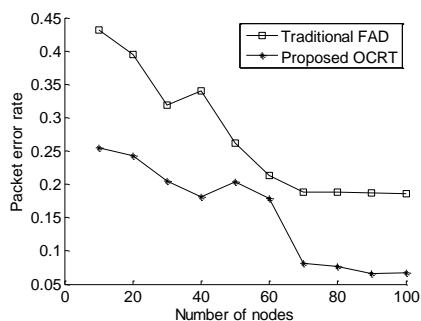

with binay tree. From Figure 5 (d), the energy efficiency of the proposed scheme is better than traditional one.



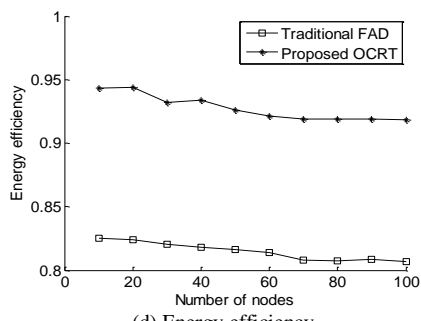
(a) Average delay.



(b) Throughput.



(c) Packet error rate.



(d) Energy efficiency.

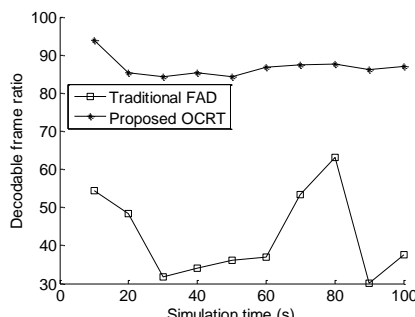
Figure 5. Performance analysis with number of nodes

TABLE II. PARAMETER AND SIMULATION SETTINGS

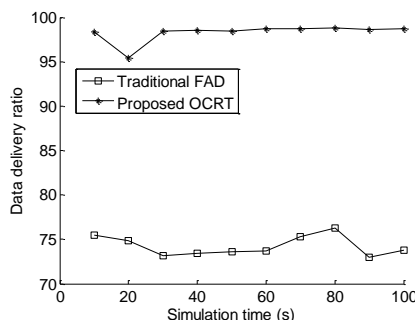
Parameters	Value
Video	Foreman
Format	QCIF
I frame	45
B frame	89
P frame	266
Total frames	400

For verifying the multimedia communication provisioning ability of the proposed OCRT, a medium

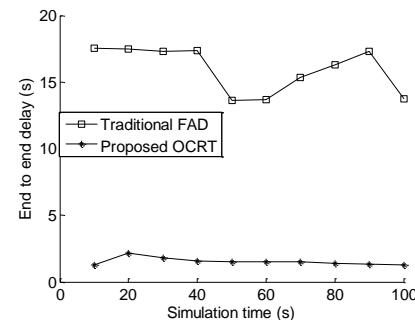
quality MPEG4 video was sent to one receiving node in experiment 3, which is shown in Figure 6. The simulation environment and parameter settings is given by Table II.



(a) Decodable frame ratio.



(b) Data delivery ratio.



(c) End to end delay

Figure 6. Multimedia performance analysis

The real media stream transmission performance is compared with Figure 6 (a) and (b). The proposed OCRT protocol can guarantee the reliable multimedia communication in wireless sensor networks than traditional FAD.

V. CONCLUSIONS

This work proposed an opportunistic cooperative reliable data transmission scheme with mobility aware model and binay tree. The aim of our paper is to guarantee the reliability of data communication in wireless sensor networks.

The protocol analyses the characteristics of performance with mobility and gives the opportunistic adaptive cooperative reliable mechanism, which is able to select the optimal relay nodes and chooses the optimal data transmissio routings. On the other hand, we combine the binay tree structure and wireless sensor networks communication, to support the reliability for wireless sensor networks. The numerical analysis and simulation

results illustrate the proposed protocol can significantly improve wireless data transmission reliability, strengthen real time performance and system throughput as well as energy efficiency.

REFERENCES

- [1] Jennifer Yick Author Vitae, Biswanath Mukherjee Author Vitae, Dipak Ghosal. "Wireless sensor network survey," *Computer Networks*, Vol. 52, no. 12, pp. 2292–2330, 2008.
- [2] Stankovic J.A. "Wireless Sensor Networks," *Computer*, Vol. 41, no. 10, pp. 92-95, 2008.
- [3] Yang Zhang, Meratnia N., Havinga P. "Outlier Detection Techniques for Wireless Sensor Networks: A Survey," *IEEE Communications Surveys & Tutorials*, Vol. 12, no. 2, pp. 159-170, 2010.
- [4] Feng Wang, Jiangchuan Liu. "Networked Wireless Sensor Data Collection: Issues, Challenges, and Approaches," *IEEE Communications Surveys & Tutorials*, Vol. 13, no. 4, pp. 673-687, 2011.
- [5] Raymond David R., Marchany, R.C., Brownfield, M.I. et. al. "Effects of Denial-of-Sleep Attacks on Wireless Sensor Network MAC Protocols," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 1, pp. 367-380, 2009.
- [6] Euisin Lee, Soochang Park, Fucai Yu, et. al. "Communication model and protocol based on multiple static sinks for supporting mobile users in wireless sensor networks," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1652-1660, 2010.
- [7] Chi-Tsun Cheng, Tse C.K., Lau F.C.M. "A Delay-Aware Data Collection Network Structure for Wireless Sensor Networks," *IEEE Sensors Journal*, vol. 11, no. 3, pp. 699-710, 2011.
- [8] Ning Wang, Yong Zhu, Wei Wei et.al. "One-to-Multipoint Laser Remote Power Supply System for Wireless Sensor Networks," *IEEE Sensors Journal*, vol. 12, no. 2, pp. 389-396, 2012.
- [9] Feng Xia. "QoS Challenges and Opportunities in Wireless Sensor/Actuator Networks," *Sensors*, vol. 8, no. 2, pp. 1099-1110, 2008.
- [10] Akan, O.B., Akyildiz, I.F. "Event-to-sink reliable transport in wireless sensor networks," *IEEE/ACM Transactions on Networking*, Vol. 13, no. 5, pp. 1003-1016, 2005.
- [11] Junyoung Heo, Jiman Hong, Yookun Cho. "EARQ: Energy Aware Routing for Real-Time and Reliable Communication in Wireless Industrial Sensor Networks," *IEEE Transactions on Industrial Informatics*, vol. 5, no. 1, pp. 3-11, 2009.
- [12] Hao Jiang, Lijia Chen, Jing Wu, et.al. "A Reliable and High-Bandwidth Multihop Wireless Sensor Network for Mine Tunnel Monitoring," *IEEE Sensors Journal*, Vol. 9, no. 11, pp. 1511-1517, 2009.
- [13] Islam K., Weiming Shen, Xianbin Wang. "Wireless Sensor Network Reliability and Security in Factory Automation: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 42, no. 6, pp. 1243-1256, 2012.
- [14] Bletsas A., Khisti A., Win M.Z. "Opportunistic cooperative diversity with feedback and cheap radios," *IEEE Transactions on Wireless Communications*, vol. 7, no. 5, pp. 1823-1827, 2008.
- [15] Hassan, E.S. "Energy-efficient hybrid opportunistic cooperative protocol for single-carrier frequency division multiple access-based networks," *IET Communications*, Vol. 6, no. 16, pp. 2602-2612, 2012.
- [16] Amin, O., Lampe, L. "Opportunistic Energy Efficient Cooperative Communication," *IEEE Wireless Communications Letters*, Vol. 1, no. 5, pp. 412-415, 2012.
- [17] Zhiguo Ding, Yu Gong, Ratnarajah T. et.al. "On the Performance of Opportunistic Cooperative Wireless Networks," *IEEE Transactions on Communications*, Vol. 56, no. 8 pp. 1236-1240, 2008.
- [18] Shan Chu, Xin Wang. "Opportunistic and cooperative spatial multiplexing in MIMO ad hoc networks," *IEEE/ACM Transactions on Networking (TON)*, Vol. 18, no. 5, pp. 1610-1623, 2010.
- [19] Koutsonikolas, D., Hu, Y.C., Chih-Chun Wang. "Pacifier: High-Throughput, Reliable Multicast Without "Crying Babies" in Wireless Mesh Networks," *IEEE/ACM Transactions on Networking*, Vol. 20, no. 5, pp. 1375-1388, 2012.
- [20] Hemachandra K.T., Beaulieu, N.C. "Outage Analysis of Opportunistic Scheduling in Dual-Hop Multiuser Relay Networks in the Presence of Interference," *IEEE Transactions on Communications*, Vol. 61, no. 5, pp. 1786-1796, 2013.
- [21] Dardari D. , Conti A., Buratti C. et. al. "Mathematical Evaluation of Environmental Monitoring Estimation Error through Energy-Efficient Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, Vol. 6, no 7, pp. 790-802, 2007.
- [22] Razi A., Afghah, F., Abedi A. "Power Optimized DSTBC Assisted DMF Relaying in Wireless Sensor Networks with Redundant Super Nodes," *IEEE Transactions on Wireless Communications*, Vol. 12, no. 2, pp. 636-645, 2013.
- [23] Lin Yang, Gang Zheng, Hengtai Ma. "A Random Network Coding-based ARQ Scheme and Performance Analysis for Wireless Broadcast," *Journal of Information and Computing Science*, Vol. 4, No. 2, 2009, pp. 124-130.

Hua Guo, male, be born in Laiwu, Shandong in China January, 1971.

On 2006 he received his graduate degree in computer application from Shandong University of Science and Technology, China. From 2007 until today he is working as a lecturer at College of Information Science and Technology at Shandong University of Science and Technology, China. His research interests include embedded system and wireless communication, especially IOT (Internet of Things).

An Improved Channel Estimation Method based on Jointly Preprocessing of Time-frequency Domain in TD-LTE System

Yang Jianning

Network Center Director, Yunnan Normal University Business School, Kunming 650106, China

Lin Kun

Admissions Office, Yunnan Normal University Business School, Kunming 650106, China

Zhao Xie

Network Center Director, Yunnan Normal University Business School, Kunming 650106, China

Abstract—As we know, LTE is the transition between the 3G and 4G and is also the global standard of 3.9G, which has improved and enhanced the air interface technology of 3G. LTE is characterized into two cases: LTE-FDD and LTE-TDD. Since LTE-TDD is researched dominantly in China, all the studies in this paper are discussed under the TD-LTE system. However, in the TDD system, the reciprocity is usually assumed since the uplink and the downlink work at the same frequency, that is to say, the estimated channel in uplink can be used to guide some downlink transmission, such as power allocation and so on. In order to estimate the uplink channel precisely, an appropriate estimation scheme with good performance is needed. In this paper, we have studied various channel estimation algorithms based on the characteristic pilot structure in uplink. And then an improved channel estimation method based on the jointly preprocessing in time-frequency domain is proposed. This method is mainly used in the following cases: the uplink channel is not estimated accurately via the successive pilot due to the interference and noise; the data processing will introduce delay and the user moves quickly, which will cause the estimated channel is not the real downlink channel. We further process the estimated channel both in time domain and in frequency domain. The final channel is used to design the beam forming vector, thus the performance of the system will be improved. Simulation results show that, the proposed method has improved the system performance in terms of bit error rate. .

Index Terms—Link Budget; Precoding Scheme; SLNR; Interference; Iteration; BER Performance

I. INTRODUCTION

Long Term Evolution (LTE) is the evolution of 3 G, proposed in the 3GPP conference in Toronto in 2004. LTE is not the 4 G technology with widespread misunderstanding, but the transition between the 3G and 4G, which is global standard of 3.9 G [1]. LTE has improved and enhanced the 3 G air interface technology, using OFDM and MIMO as unique standard of its

wireless network evolution, which can provide downside under 326 Mbit/s and about 86 Mbit/s peak rate in the 20 MHz spectrum bandwidth [2]. Moreover, LTE has improved performance of the cell-edge user, increased the cell capacity and reduced the delay system. LTE is characterized into two cases: LTE-FDD and LTE-TDD. China mobile are promoting the development of TD-LTE at full tilt and all the studies in this paper are discussed under the TD-LTE system [3].

Since the cyclic prefix (CP) is introduced in OFDM which is the key technology of LTE, we can utilize frequency equalization to eliminate the multi-path interference at the receiver [4]. As we know, frequency equalization requires that the accurate frequency response is available on each subcarrier, therefore the channel need to be estimated before equalization and the accuracy of channel estimation will directly determine the performance of the receiver. In addition, in most cases, if channel information is available, the performance of the system will be promoted. In TD-LTE system, various preprocessing techniques in downlink based on the channel information, such as power allocation, precoding and so on, need to know the estimated channel information in uplink to be completed via the reciprocity between uplink and downlink. Usually, the accuracy of the channel will determine the performance of the system which utilizes these techniques [5].

Channel estimation is to estimate the wireless channel response from the transmitter to the receiver. Common channel estimation methods contain blind channel estimation and the channel estimation based on pilot signal. Although blind channel estimation needs not transmit pilot which can improve the data transmitting rate, it is necessary to collect a large number of data for ensuring the reliability [6, 7]. In TD-LTE system the pilot signal is continuous in the frequency domain. Therefore the estimated channel in the uplink holds the jump characteristic in amplitude and phase when the interference and noise exist. However, this paper

proposes a new method called filter in frequency domain to smooth the jump between adjacent subcarriers [7].

TD-LTE can largely reduce the communication delay, increase user data rates, improved system capacity and coverage, and reduces operator costs, making it a major mobile operator of choice for building next-generation communications system standard. TD-LTE features a variety of advantages thanks to their use of advanced communications technologies, including OFDM, MIMO, adaptive modulation and coding (AMC) and hybrid automatic repeat request (HARQ) and so on. In the test personnel to TD-LTE communication system, when tested according to the different needs of different test cases written test plan, modify parameters associated with these technologies, and in the end of the test based on those parameters need to check the test results, it is in this brief introduction to TD-LTE systems involved in communications technology.

In this paper, the estimated channel in uplink will be used to calculate the beam forming vector in downlink. In the present some beam forming method, such as Zero-forcing (ZF) beam forming method, the Minimum Variance Distortionless Response (MVDR) beam forming method, the Minimum Mean Square Error (MMSE) and so on, all need to use the original channel information to realize beam forming.

TD-LTE system utilizes the estimated channel at present moment in uplink to compute the beam forming vector at the next moment in downlink, which is based on the assumption that the feedback delay between downlink and uplink is zero [8]. When the channel change slowly, this kind of practice is feasible. But in practice the receiver need to accomplish various signal processing technologies and the delay exists when the channel information is feedback to transmitter. Under the circumstances, if we use the estimated channel at current moment in uplink to compute the beam forming vector at the next moment in downlink, the error will be introduced due to the variation of channel. Therefore it is not very reasonable that we use the estimated channel at current moment in uplink to compute the beam forming vector at the next moment in downlink. The channel is caused to be vibrational as a result of the movement of the mobile station or other reasons, in this case, using such channel information for the beam-forming of downlink will introduce some degree of error, the bit error rate (BER) performance will decrease and the communication quality cannot be guaranteed [9]. Therefore we need to forecast the channel which is closer to actual situation.

Based on the advantage of the existing channel estimation technologies, we propose a new channel preprocessing scheme termed the improved channel estimation method based on the jointly preprocessing in time-frequency domain. Simulation results show that, the proposed channel preprocessing scheme has improved the system performance in BER compared with pioneering channel estimation technologies [10].

The remainder of this paper is organized as follows. In Section 2, we outline the system model of the channel estimation [11]. In Section 3 we describe the pioneering

channel estimation technologies. In Section 4, we present an improved channel estimation method based on the jointly preprocessing in time-frequency domain. In section 5, we provide the simulation results and performance analysis, and Section 6 concludes the paper. Notations: $(\cdot)^H$, $(\cdot)^{-1}$, and $E(\cdot)$ denote, conjugate transpose, inverse, and expectation, respectively [12].

II. RELATED WORKS

A. The Basic Block Diagram of Modern Mobile Communication

The transmitting data goby channel encoding and interleaving is processed with SC-FDMA technology, and CP is added to it, the final signal is transmitted in the transmitter. Channel estimation utilizes the received pilot to estimate the frequency response of the channel. Therefore channel estimation results will directly affects the system performance. The receiver decomposes the data in inverse process accordingly. For the subcarrier k , let h denote the channel from the transmitter to receiver and x is the transmitting data. Figure 1 shows the basic block diagram of modern mobile communication base station [13]. The signal at the receiver is given by

$$y = hwx + n + s \quad (1)$$

where w is the beam forming vector, s denotes the sum of all interference terms and n is the additive white Gaussian noise with $E[nm^H] = \sigma^2$.

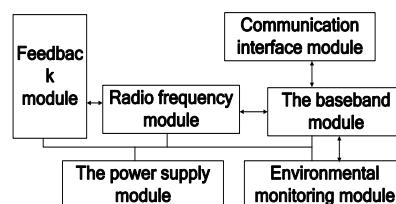


Figure 1. The basic block diagram of modern mobile communication base station

The mobile communication base station has experienced from analog to digital, from narrowband to broadband development sharp change, the system architecture with the evolution and development of the function of the system constant. At present, the base station of the latest generation of 3G mobile communication with multi-carrier, high-efficient pre-distortion digital power amplifier, a high performance HSDPA, open architecture features. The base station is in the form of the future, will undoubtedly toward functional macro base station, more powerful and integrated volume more portable, micro base station network more flexible distributed base station 3 direction; at the same time the base station in the framework, the evolution will be open, modular products. As the base function is stronger, the composition of the base station also at continuously perfect, but the basic functional modules of the base station from the digital communication are basically unchanged [14].

Software testing is to verify whether the difference between target design requirements, in accordance with the proposed to identify actual output and output theory exists to design, locating and correcting incorrect to reduce risk. However, only for the purpose of checking and correcting the software defects are not test, through the analysis of the abnormal output positioning reasons for this defect, can let testers find out the test rules and effective strategies, and improve the effectiveness and efficiency of test. This analysis can also detect deficiencies in the software development process, so that relevant personnel quickly to improve. Even if not detected any defect, test analysis can also provide reference for software quality evaluation [15-17].

According to (1), we can estimate the frequency response of channel using the pilot by

$$h = \frac{y}{x} \tag{2}$$

The channel gain on other symbols in one sub-frame can be obtained via the interpolation of the channel gain on the pilot subcarrier. The development of the communication technology of RF system require, which also contributed to the improvement and optimization of RF transceiver system architecture [18]. Put each new structure is a breakthrough of the previous theory, also caused new problems brought new advantage at the same time, people also gradually overcome these problems in promoting the system architecture evolution to update, better direction [19]. Of course the evolution direction is not change, namely, high reliability, low cost, low power, high integration and so. Undeniable, has a deep impact of the rapid development of integrated chip for RF system architecture evolution. In the same RF system, usually the receiver and transmitter is reciprocal symmetric architecture, this paper will select several typical architecture of receiver architecture appears in the evolution process as an example, to analyze the characteristics and differences of different RF system architecture [20].

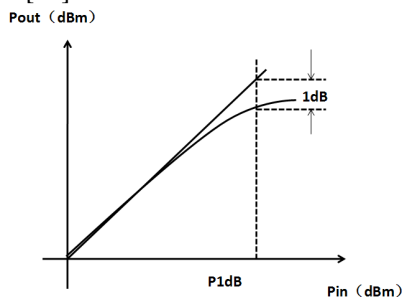


Figure 2. The power output of the 1dB compression point

The LMMSE channel estimation algorithm has good performance of channel estimation and can remove the effect of noise by using the statistical correlation of channel [21, 22]. However, the statistics of channel is hard to obtain and the complexity is large, therefore the availability of the LMMSE algorithm is poor. In most systems of opposite sign, so gain will system with input signal amplitude increases, usually gain dropped to define

than the linear output power gain and low 1dB values for the power output of the 1dB compression point as shown in Figure 2. When the power is more than P1dB, the gain will decline rapidly and reach the maximum output power, the value is generally greater than that of P1dB 3dB-4dB. This is to say the 1dB compression point is larger, the dynamic range of linear larger RF system.

B. Various Channel Estimation Schemes

LTE protocol provides a specific pilot structure used for channel estimation for multiple antennas system. We can accomplish the channel estimation with the auxiliary of pilot. The DMRS (Demodulation Reference Signal) of PUSCH (Physical Uplink Shared Channel) in TD-LTE system is only transmitted on the frequency band where the UE (User Equipment) is transmitted; therefore the pilot length is restricted. For example in extreme cases, when the UE transmits PUSCH on only one RB (Resource Block), if DMRS using massive pilot patterns, the sequence's length is the frequency 12, or if using pectinate or other patterns, the sequence's length frequency is insufficient to 12. Since the pilot length determines the number of available pilot, the pilot length is not ought to be too little. Therefore uplink will utilize the massive pilot as depicted in Fig. 3.

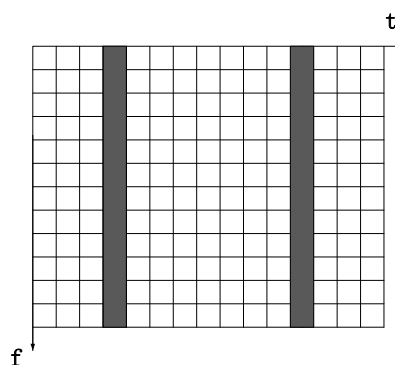


Figure 3. The pilot structure of uplink

Using such pilot pattern we can estimate the channel response on each subcarrier. The channel estimation of TD-LTE system is usually divided into two parts, one is the channel response estimation of the pilot subcarrier, where we mainly finish the corresponding processing on the pilot and eliminating the interference effects and finally obtain the channel response estimation of the pilot subcarrier; The other part is part of interpolation, through the interpolation we can get complete channel response estimation. However, there are several channel estimation schemes.

C. The LS Channel Estimation Algorithm

The least square estimation (LS) algorithm is one of the classical channel estimation algorithms.

According to the minimum variance criteria, the cost function is defined as:

$$J_{LS} = (Y_p - \hat{H}_p X_p)^H (Y_p - \hat{H}_p X_p) \tag{3}$$

where Y_p denotes the received signal on the pilot subcarriers, \hat{H}_p refers to the channel response estimation on the pilot subcarriers and X_p is the transmitted pilot signal. Let J_{LS} in (3) is zero, from the equation (3), we calculate the \hat{H}_p according to

$$H_p = \frac{Y_p}{X_p} = \frac{H_p X_p}{X_p} + \frac{N_p}{X_p} = H_p + \frac{N_p}{X_p} \quad (4)$$

where N_p is the noise and interference on the pilot subcarrier.

For this kind of channel estimation scheme, all operations can be accomplished in the frequency domain and the structure is very simple. Therefore this kind of channel estimation scheme is used widely. However, as expressed in (4), since the LS criterion does not consider the influence of the noise, the channel estimation results will be affected by noise seriously.

The MMSE channel estimation algorithm

In order to remove the effect of noise and improve the accuracy of channel estimation, the MMSE criterion is considered to design channel estimation algorithm. According to MMSE criteria, the cost function is defined as:

$$J_{mmse} = E[(H_p - H_p')^* (H_p - H_p')] \quad (5)$$

Let M denote the number of pilot subcarrier and $Y_p = [Y_1, Y_1, \dots, Y_M]^T$ denote the set of the received signal at all pilot subcarriers. Then $R_{YY} = E[Y_p Y_p^H]$ and $R_{HY} = E[H_p Y_p^H]$ are the self-correlation matrix of the received signal at each pilot subcarrier and the cross-correlation matrix of the received signal at each pilot subcarrier with the channel response. After deriving, the channel estimation based on MMSE criterion is:

$$\hat{H}_p = W^T Y_p = R_{HY} R_{YY}^{-1} Y_p \quad (6)$$

where $\hat{H}_p = [\hat{H}_1, \hat{H}_2, \dots, \hat{H}_M]^T$, $W = [w_1, w_2, \dots, w_M]^T$ is the tap of the filter.

Apparently, in order to estimate the channel response at the pilot subcarrier by using the received pilot signal via MMSE criterion, the cross-correlation matrix of the received signal at each pilot subcarrier and statistic characteristics of channel are needed to be available. However, in practice, the computing complexity is very large and the algorithm is hard to be carried out. Therefore some simple algorithms will be utilized, such as linear MMSE, which is proceeded only in frequency domain. The channel estimation can be expressed as

$$\hat{H}_{MMSE} = R_{HH} (R_{HH} + \sigma_n^2 (XX^H)^{-1})^{-1} \hat{H}_{LS} \quad (7)$$

where \hat{H}_{LS} is the result of the LS algorithm, σ_n^2 is the variance of the additive white Gaussian noise and $R_{HH} = E[HH^H]$ is the self-correlation matrix of channel.

From the expression (7), we know that the LMMSE channel estimation algorithm is based on the LS channel estimation algorithm.

III. PROPOSED SCHEME

In TD-LTE system, since time differences exist between uplink and downlink, transmission channel downlink information of uplink and downlink is not exactly the same. Under the circumstances, if we use the current estimated channel information to design the beam forming vector at next time, the error will be introduced due to the channel change. In this paper, a time prediction scheme of channel is proposed. In addition, the channel information on each subcarrier can be directly estimated without interpolation because of the inherent successive pilot structure in uplink and the estimated channel information on each subcarrier is accurate. When noise and interference exist, the channel information of adjacent subcarriers is relatively independent and the jump between adjacent subcarriers will be introduced. However, TD-LTE system utilize SCME (Spatial Channel Model Extended) channel to model the system and the SCME channel changes slowly, that is to say, the channel information between adjacent subcarriers is highly correlated. Therefore the jump characteristic and slow degeneration of adjacent subcarriers are contradictory. In this paper, we will use some smooth processing technologies to smooth the channel through the prediction in time domain to improve the system performance.

To accomplish our proposed channel estimation method by utilizing the prediction in time domain and sooth processing, the following steps will be included.

A. Using the Received Pilot Signal to Estimate the Uplink Channel

According to TD-LTE protocol, uplink pilot pattern adopts the massive pattern, that is, the pilot signal is inserted on each subcarrier in the middle symbol of each time slot. Therefore the channel information on each subcarrier can be estimated at the receiver. This kind of pilot pattern is suitable for frequency selective channel. For simplicity, the LS channel estimation algorithm is considered in this paper. By using pilot signal, we can estimate the channel response on each subcarrier. For one time slot, the estimated channel information can be expressed as:

$$\mathbf{H} = [H_1, H_2, \dots, H_N]^T \quad (8)$$

where N is the number of the subcarriers and H_k is the channel information on the k th subcarrier.

B. Predicting the Estimated Channel in Time Domain

Considering the slow change characteristics of the space channel in time domain, the downlink and uplink are time division duplex in TD-LTE system. Compared with the FDD system, TDD system is more flexible to configure proportion downlink and uplink resources, in order to support different business types. No matter which kinds of configuration, they all have a common

characteristic that the downlink transmission is the time division, interlock together. Channel in adjacent time slot is always relevant. Considering the correlation we use the channel information at current time to predict the channel information at the next moment.

The transmission time interval (TTI) is 1 ms, containing a sub frame (2 slot). The pilot is on the middle SC-FDMA symbol at each time slot, each sub-carrier have guided frequency, which covers the whole bandwidth. Then through the channel estimation, we will get the channel information of two slots in frequency domain:

$$\begin{aligned} \mathbf{H}_1 &= [H_{11}, H_{12}, \dots, H_{1N}]^T \\ \mathbf{H}_2 &= [H_{21}, H_{22}, \dots, H_{2N}]^T \end{aligned} \quad (9)$$

In order to get the channel information of next frame, we make the forecast on each sub-carrier get the channel information at current moment. The prediction contains many ways. If the channel changes slowly, the effect of the forecast has inconspicuous effect; but if channel changes fast, the choice of prediction is especially important. Fig. 3 compares the system performance under the circumstances that the time delay between uplink and downlink exist or not. Fig. 1 show that time delay has certain effects on the bit error rate (BER) performance of the system, the higher signal-to-noise ratio (SNR) is, the more obvious this effect is. Therefore it is necessary to predict the estimated channel information.

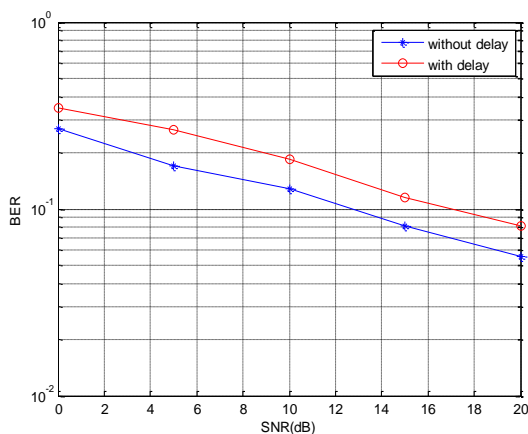


Figure 4. The comparison of system performance with delay and without delay

Since the SCME channel in our system changes slowly, the prediction in this paper makes a fine adjustment for the estimated channel. When the estimated channel in one sub frame is known, for the k th subcarrier, the channel information of the next sub frame is predicted linearly:

$$\begin{cases} H_{3k} = (\alpha_1 H_{1k} + \beta_1 H_{2k}) / (\alpha_1 + \beta_1) \\ H_{4k} = (\alpha_2 H_{1k} + \beta_2 H_{2k}) / (\alpha_2 + \beta_2) \end{cases} \quad (10)$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are preset parameters. Based on certain amount of experiments, $\alpha_1, \alpha_2, \beta_1, \beta_2$ are chosen as the following values which can achieve better effects.

$$\begin{cases} \alpha_1 = 1, \beta_1 = \frac{SNR}{20} * 0.1 \\ \alpha_2 = \frac{SNR}{20} * 0.1, \beta_2 = 1 \end{cases} \quad (11)$$

C. Smooth the Predicted Channel in Frequency Domain

By using the massive pilot pattern in uplink, the channel information on each subcarrier can be directly estimated without interpolation and the estimated channel information on each subcarrier is accurate. When noise and interference exist, the channel information of adjacent subcarriers is relatively independent and the jump between adjacent subcarriers will be introduced. The correlation of the real channel is destroyed. If the estimated channel is directly used for the designing the beam forming vector, the system performance will decrease.

However, the above-mentioned prediction contributes a fine adjustment to the estimated channel has improved the system performance weakly, which do not change the jump in terms of magnitude and phase caused by successive pilot structure and therefore we need make some processing in frequency domain. Due to the correlation of the adjacent subcarriers, we can sooth the channel in frequency domain. Smooth's role is to remove or weaken the high frequency components of channel estimation, and its essence is the low pass filtering, that is, through the low pass filtering, the high frequency component is removed.

There are many kinds of smooth, we list the main two ways: arithmetic average method and low-pass with window method. For arithmetic average method, many adjacent subcarriers are considered; we use the average value of these subcarriers to replace the original sub-carrier channel coefficient. value here refers to take a subcarrier, and it before and after the adjacent several subcarrier channel coefficient, in fact the average operation plays a smooth role, each sub-carrier will consider its adjacent subcarriers and the high coefficients is weakened. For the low-pass with window method, using low-pass filter for channel information in frequency domain filtering is another smooth way. Here we use digital filters and the design of the filter contains the following steps: designate the expected frequency response of the filter through the Fourier inverse transform; select a window function which meet band pass or attenuation index, and then determine the order of filter using the relationship between the length of filter and transitional bandwidth; get the filter coefficients from the selected the window function.

In this paper, the simple arithmetic average method is considered and the average with L subcarriers is considered. The predicted channel is $\mathbf{HH} = [HH_1, HH_2, \dots, HH_{N_c}]^T$ (N_c is the number of subcarriers). For the k th subcarrier, the channel through smooth is

$$H'_k = \begin{cases} (H_{k-\frac{L-1}{2}} + \dots + H_{k-1} + H_k + H_{k+1} + \dots + H_{k+\frac{L-1}{2}}) / L, & k = \frac{L+1}{2}, \dots, N_c - \frac{L+1}{2} \\ H_k & \end{cases}$$

Compared the magnitude and phase after smooth with original channel

The upper two figures are the magnitude of smoothed channel and original channel, and the under two figures are the phase of smoothed channel and original channel. Obviously, the channel through smooth processing is smooth. Through such a simple way, we would weaken the influence of the noise and interference in channel estimation and prediction. The magnitude and phase of channel information diagram after smooth are smoother obviously than the corresponding quantities of original channel, which is more in line with the real channel characteristics. And if the smooth scheme is used for designing the beam forming vector, the system performance will be better.

The low-pass with window method is also considered in this paper. The window functions mainly contain cosine window, Hamming window and Kaiser window etc.

D. Use the Optimized Channel to Compute the Beam Forming Vector

The optimized channel through above-mentioned prediction and smooth is used to compute the beam forming vector. There are many kinds of beam forming methods, such as ZF, MMSE and so on.

IV. SIMULATION RESULTS

In this section, we provide the simulation results to demonstrate the effectiveness of our proposals for the improved channel estimation method based on the jointly preprocessing in time-frequency domain. We assume the channel is a quasi-static flat fading channel in our system. Without loss of generality, we consider the BER as the performance measurement to verify the advantage of our scheme in this paper.

In our TD-LTE system, the number of antennas at base station is 4 and the user equipment (UE) is equipped a single antenna and the antenna distance is 0.5λ . The smooth mean is the average with 5 subcarriers and ZF is chosen as the beam forming method. With 3km/h of UE speed, we compare the BER performance between the beam forming method with the proposed scheme in this paper and the beam forming method without the proposed scheme in Fig. 5.

According to the simulation results, the proposed estimation method based on the jointly preprocessing in time-frequency domain in this paper has improved the system performance in terms of BER. That is to say, the proposed estimation scheme based on the jointly preprocessing in time-frequency domain in this paper is better than the original scheme.

However, the proposed estimation method based on the jointly preprocessing in time-frequency domain in this paper can be used to design beam forming vector of the MMSE and MVDR beam forming or other beam forming

schemes. Systematic test on the receiver, is whether the receiver RF link design is necessary for successful, more important is the important source of further improvement design. The design of the test is divided into two parts: LNA single board test and whole test. In order to separate to test the noise coefficient and gain of low noise amplifier, the design will be specially low noise amplifier of independent board, Figure 6 is a low noise amplifier board testing environment.

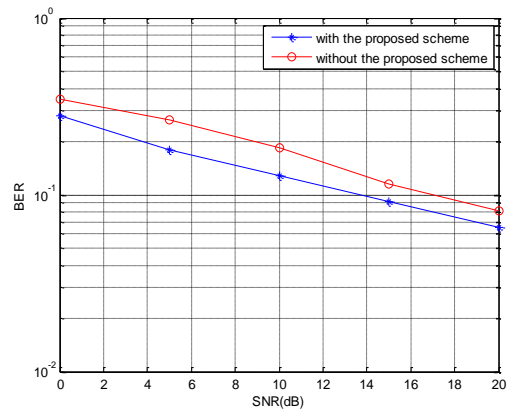


Figure 5. The comparison between the system without and with the proposed scheme

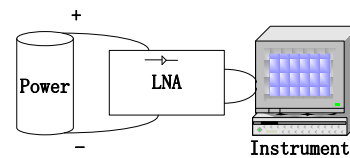


Figure 6. The low noise amplifier board testing environment

In the receiver system is complete, the design of the final output board should provide analog input ADC, after ADC sampling by FPGA analysis to assess the quality of signal, data and then, determine the performance of receiver. In view of this, the output signal now through the fly line will receive link RF unit to the evaluation board ADS6445 evaluation board, analog to digital converter on the input signal to digital signal to PC, PC to analyze the data through dedicated software.

The present communication industry concern LTE as the research object, using 3GPP for the LTE specification and requirements as the basis, combined with the master the RF knowledge tempted to realize RF circuit of TD-LTE base station receiver. Adhere to the theory research and analysis to guide the engineering practice in the design process, through the engineering practice to deepen understanding and cognition theory. In this paper, base station system and RF system structure is introduced as the breakthrough point, discuss several receiver structures for mainstream, and select zero if architecture design direction as the. Then, the related protocol with 3GPP extracted from the RF characteristics of TD-LTE base station receiver characteristic. Then, based on the

RF characteristics quantitatively calculate a specific radio frequency index TD-LTE base station receiver, combined with ADS2008 link simulation control these RF index to each function module of the specific circuit. Then, in strict accordance with the circuit module of the requirement for selection and circuit design, and create physical.

Affected by the broadband mobile networks, IP trend, TD-LTE system is designed for the whole IP network to provide better service based on network communication, it and traditional cellular mobile wide area network, broadband wireless access metropolitan area network, wireless transmission network and the traditional fixed access network and access based on the same IP core network, realization of seamless connection service. And relative to the 3G communication system simplifies the network architecture, a flat structure. TD-LTE air interface protocol, including the physical layer (PHY), media access control (MAC) layer, the radio link control (RLC) layer, data aggregation protocol (PDCP) layer and the radio resource control (RRC) layer. Control surface air interface protocol stack of TD-LTE system and user plane respectively, as shown in Figure 7.

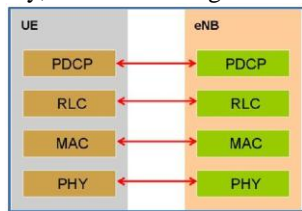


Figure 7. Control surface air interface protocol stack of TD-LTE system

TD-LTE communication protocol stack is usually divided into 3 layers. The PHY layer is often referred to as the layer 1 (L1); PDCP, RLC and MAC layer is often referred to as layer 2 (L2); non access stratum (NAS) and the RRC layer is often referred to as layer 3 (L3). TD-LTE system information is encapsulated into a master information block (MIB) and multiple system information block (SIB), associated with the various functions of the system parameters are stored in MIB, which is an important information for the UE initial access, usually including the use of frequent parameter; SIB1 included with the district selection related parameters, including system information block scheduling information other; SIB2 includes channel information related to sharing; SIB3-SIB11 contains frequency, frequency, wireless access technology (RAT) between the cell reselection parameters, earthquake and tsunami warning system and other related.

The RRC connection control including RRC connection establishment, retention and release, safety management and data radio bearer (DRB) to establish, modify and release. UE RRC connection state decide the process and the operation of the implementation of the access layer, the RRC state has two kinds: RRC idle state and the RRC connection state. In the RRC idle state: UE will monitor the paging channel, check whether there is a call; monitor the broadcast channel, in order to obtain system information; specific discontinuous reception;

implementing cell selection or reselection. In the RRC connection state: to interact with the network data; reporting cache state and the channel quality can be controlled by eNB; cell switching. RRC connection establishment initiated by UE, eNB detects the connection establishment request after the corresponding configuration and return, the specific process is shown in Figure 8.

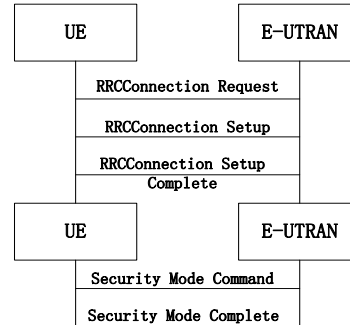


Figure 8. The connection establishment request after the corresponding configuration and return, the specific process

In the RRC idle state of LTE system under the mobility control refers to the UE perform cell reselection. Cell reselection on each frequency corresponding to the corresponding priority, and the priority given by the system information, the frequency of cell reselection according to these equal, then reordered according to the channel quality residential. In the RRC system connection, LTE mobility control for cell switching of E-UTRAN execution state. E-UTRAN chooses to receive cell switching to UE, to maintain the link connection. E-UTRAN measurement report requires UE to submit candidate receiving district before switching; because UE is always connected with a small, therefore the handoff in LTE is a hard handoff.

In the area of the selection process, UE estimates of the support RAT, support carrier, and find out the strongest signal from the cells of the. UE choose a good plot, in order to cell reselection, cell reselection according to the absolute priority of each plot, UE channel quality. The amount of sorting, then UE test whether target cell can access. If there is more than one candidate, UE District second ranking. If the service area of the order of R_s , the cells adjacent to the order of R_n , once an adjacent cell sorting than the service area, and maintain a certain time, then UE for the small cell reselection purposes.

V. CONCLUSION

In this contribution, LTE is considered as a considerably promising technology for the next generation mobile communication system. Since China mobile are promoting the development of TD-LTE at full tilt, we consider the TDD mode of LTE. However, the TD-LTE system is also an OFDM system and frequency equalization requires that the accurate frequency response is available on each subcarrier. In addition, various preprocessing techniques based on the channel information in downlink need know the estimated channel information in uplink which can be completed via the reciprocity between uplink and downlink.

Therefore the channel estimation is necessary in practical application.

In TD-LTE system, since time differences exist between uplink and downlink, transmission channel downlink information of uplink is not exactly the same with downlink. Under the circumstances, if we use the estimated channel information at current time to design the beam forming vector at next time, the error will be introduced due to the channel variability. In this paper, a prediction scheme of channel is proposed in time domain.

In addition, the channel information on each subcarrier can be directly estimated without interpolation because of the inherent successive pilot structure in uplink of TD-LTE system and the estimated channel information on each subcarrier will be accurate. When noise and interference exist, the channel information of adjacent subcarriers is relatively independent and the jump between adjacent subcarriers will be introduced. However, TD-LTE system utilize SCME channel to model the system and the SCME channel changes slowly, that is to say, the channel information between adjacent subcarriers is highly correlated. Therefore the jump characteristic and slow degeneration of adjacent subcarriers are contradictory. In order to solve the contradiction, in this paper, we will use some smooth processing technologies to smooth the channel through the prediction in time domain to improve the system performance.

In this paper we propose the estimation method based on the jointly preprocessing in time-frequency domain which outperforms the original channel estimation scheme.

REFERENCES

- [1] 3GPP TS 36. 300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved UTRA (E-UTRA). 2008.
- [2] L Tong, G Xu, B Hassibi, and T Kailath, "Blind channel estimation based on second-order statistics: a frequency-domain approach", *IEEE Trans. Inform. Theory*, vol. 41, pp. 329-334, 1995.
- [3] Y Zhao, A Huang, "A novel channel estimation methods for OFDM mobile communication systems based on pilot signals and transform domain processing", in *Pro. IEEE 47th Vehicular Technology Conference*, pp. 2089-2093, 1997.
- [4] Theodore S. Rappaport etc. "Wireless Communications Principles and Practice." *Publishing House of Electronics Industry*, 2003.
- [5] Sinern Estimation Coleri, MustafaErgen, AnujPuri, Ahmad Bahai, "A Study of Channel in OFDM Systems", *IEEE VTC, Vancouver, Canada. September*, 2002.
- [6] D. B. Van, O. Edfors, and M. Sandell, "On channel estimation in OFDM systems", *Proc. IEEE Vehic. Tech. Conf*, pp. 815-819, 1999.
- [7] H. Landau, H. O.. Prolatespheriodal wave functions, "Fourier analysis and uncertainty-III: The dimension of the space of essentially time and band-limited signal", *Bell Syst. Tech.* pp. 41-44, 2002.
- [8] Bingyang Wu; Shixin Cheng; Ming Chen; Haifeng Wang. "Analysis of decision aided channel estimation in clipped OFDM", *Vehicular Technology Conference, VTC-2005-Fall*, pp. 1030-1033, 2005.
- [9] EffridedDustin, KeffRasjka, John Paul, "Automated Software Testing Introduction, Management, and Performance", *Addison-Wesley* 2002.
- [10] G. Patel, S. Dennett. "The 3GPP and 3GPP2 Movements Toward an All-IP Mobile Network", *IEEE Personal Communications*, vol. 7, no. 4, pp. 66-64, 2002.
- [11] Chernak Y. "Validating and improving test-case effectiveness", *IEEE Software*, vol. 18, no. 1, pp. 81-86, 2001.
- [12] QinqunFeng, Zhang Wuguang, Peng Yan, "Research on software testing tool based on Internet", *Computer applications and software*, vol. 23, pp. 133-135, 2006.
- [13] poolBaoyong, Yu Zhiping stone, "a review of. CMOS RF integrated circuit analysis and design", vol. 32, pp. 31-35, 2007.
- [14] Li Genqiang, Kuang Wang, Wen Zhi-cheng, "RF and wireless technology", *Beijing: Electronic Industry*, vol. 12.. pp. 13-15, 2009.
- [15] Cao Peng, Qi Wei, "Broadband wireless communication transceiver technology", *Beijing: mechanical industry*, vol. 80, pp. 33-35, 2012.
- [16] Ozaki, K.; Tomitsuka, K.; Okazaki, A.; Sano, H.; Kubo, H., "Channel estimation technique for OFDM systems spread by chirp sequences", *2012 IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, pp. 2125-2130, 2012.
- [17] Vinogradova, J.; Sarmadi, N.; Pesavento, M, "Subspace-based semiblind channel estimation method for fast fading orthogonally coded MIMO-OFDM systems ", *2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 153-156, 2011.
- [18] M. Sadek, A. Tarighat, A. H. Sayed, "A Leakage-based Precoding Scheme for Downlink multi-user MIMO Channels", *IEEE Transactions on Wireless Communications*, vol. 26, no. 8, pp. 1505-1515, 2008.
- [19] A. Tarighat, M. Sadek, A. H. Sayed, "A multi User Beamforming Scheme for Downlink MIMO Channels based on Maximizing Signal-to-Leakage Ratios", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1129-1132, 2005.
- [20] J. van de Beek, O. Edfors, M. Sandell, S. Wilson, P. Borjesson, "On Channel Estimation in OFDM System", in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 815-819, 1995.
- [21] K. Wong, R. Cheng, K. B. Letaeif, R. D. Murch, "Adaptive antennas at the mobile and base stations in an OFDM/TDMA system", *IEEE Transactions on Communications*, vol. 49, no. 1, pp. 195-206, 2001.
- [22] M. Sadek, A. Tarighat, A. H. Sayed, "Active Antenna Selection in multi-user MIMO Communications, " *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1498-1510, 2007.

Dynamic Routing Algorithm Based on the Channel Quality Control for Farmland Sensor Networks

Dongfeng Xu

College of Informatics, South China Agricultural University, Guangzhou, 510642, China
Email: xdf123@scau.edu.cn

Abstract—This article reports a Dynamic Routing Algorithm for Farmland Sensor Networks (DRA-FSN) based on channel quality control to improve energy efficiency, which combines the distance and communication characteristics of farmland wireless sensor network. The functional architecture of the DRA-FSN algorithm, routing establish the mechanisms, the communication transmission mechanism, the global routing beacon return mechanism, abnormal node handling mechanism and sensor networks timing control mechanisms were designed in detail in this article. This article also evaluates and simulated the performance of DRA-FSN algorithm in different conditions from energy efficiency, packet energy consumption and packet distribution balance by comparing DRA-FSN algorithm with DSDV, EAP algorithm. Simulations showed that the DRA-FSN was more energy efficient than EAP and DSDV, the DRA-FSN algorithm overcame the shortcoming that capacity and bandwidth of the routing table correspondingly increase as more and more nodes joining the network. It has better performance in scalability and network loading balance.

Index Terms—Wireless Sensor Networks; Dynamic Routing Algorithm; Channel Quality; Energy Efficiency

I. INTRODUCTION

Wireless sensor networks (WSNs) are providing tremendous benefit for a number of industries. The ability to add remote sensing points, without the cost of running wires, results in numerous benefits including energy and material savings, process improvements, labor savings, and productivity increases. Power cast takes the capabilities of wireless sensors a step further by allowing them to be powered without wires and without the need to change batteries [1, 2]. Wireless sensors networks are being widely deployed and power cast's technology can provide benefit for many applications including: Farms, forests, mountains, oceans, and so on [3, 4].

In WSNs, information is transmitted through the wireless radio channel which is susceptible to signal attenuation, noise, reflection, diffusion, and other factors. These factors result in the signal transmission distance and the packet loss rate and the error rate is very unstable, so that the application of wireless sensor networks subject to considerable restrictions [5, 6]. Wireless sensor channel more vulnerable to the impact of changes in the

external environment than wired networks. Ananastasi's studies confirmed that the communication distance of wireless sensor influenced by rain and fog [7, 8, 9]. The transmission distance of Mica2dot node was 120m in 2.4GHz frequency tests, but the transmission distance is only about 10m in fog and heavy rain test. Kang et al tested the transmission of wireless sensor signals influenced by the weather situation. They pointed out that the wind can make the antenna vibrate, and thus the signal was impacted significantly. Also they pointed out that rain, snow and fog signal cannot be ignored [6]. Information is vulnerable to outside interference in the process of information transmission in the wireless sensor, and an error occurs. Jeong J, et al. carried out two types of experiments indoor and outdoor, and found the wireless sensor error bit is mostly one and two bits. Also they pointed out that the scheme of 1 and 2-bit linear error-correcting codes is efficient because the multi-hop nature of information transmission in wireless sensor error correction code method the proportion of passes to save energy [10, 11]. Ahn et al also designed a method of data error correction on FECA algorithm for mobile wireless networks from the packet error rate distance function [12, 13].

In farmland sensor network environment, there are many uncertain factors, such as lower communication quality and higher error packet. Based on these considerations, this paper drew on the idea of directed diffusion protocol and the Destination Sequenced Distance Vector (DSDV) algorithm [14, 15], combined with the characteristics of farmland wireless sensor network communication channel quality. In order to improve the effectiveness of network energy, this paper proposed an energy efficient farmland dynamic routing algorithm (DRA-FSN) for farmland wireless sensor networks. The DRA-FSN routing has been simplified initialization phase algorithm for the fixed characteristics of the location of the base station. The DRA-FSN algorithm took into account the channel quality of communication between nodes, the global routing information, the remaining nodes energy, and a number of factors, so that the next hop node had the dynamic and strong channel adaptability for prolonging the network life cycle. The DRA-FSN had better scalability, for its non-corresponding increase in capacity and bandwidth of

the routing tables while more and more nodes joined the network.

The rest of this paper is organized as follows, section 2 gives Functional architecture of the DRA-FSN algorithm, section 3 describes The establishment mechanisms, the communications transport mechanism, the global routing beacon return mechanism and time control mechanism of the sensor network, the section 4 simulates the performance comparisons evaluation of the DRA-FSN algorithm with DSDV and EAP routing algorithm from the energy efficiency, the average energy consumption of the packet and the packet distribution.

II. THE FUNCTIONAL ARCHITECTURE OF DRA-FSN ROUTING ALGORITHM

There are six functional modules in the DRA-FSN routing algorithm. Which are the Route Setup Module (RSM), the Data Transmission Module (DTM), the Next Hop Selection Module (NHSM), the Global Route Beacon Module (GRBM), the Packet Handle Module (PHM) and Table Handle Module (THM) shown in figure 1. The DRA-FSN routing Algorithm layered sensor network in the whole network flooding Hello packets to establish the gradient field and obtain the node information from a neighbor.

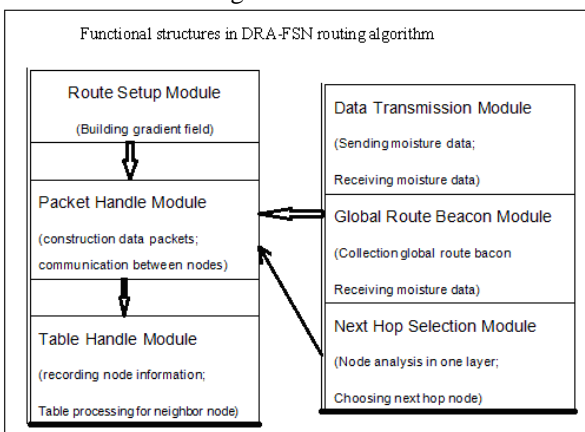


Figure 1. Functional Structure of DRA-FSN

The responsibility of the RSM was initiating routing. The PHM was responsible for construction data packets and communication between nodes. The THM mainly used for recording node information and providing data processing for a neighbor node. The DTM drove moisture packet sending. When the local node clock Threshold arriving or receiving water packets from the neighbor, the DTM wake up the next hop, constructs packets and sends out. The GRBM's main function was to complete the global routing beacons collections, when the water packets reached the base station node, the base station node would send a global routing beacon, and the beacons reversed transmission along the transmission route of the water packets. The NHSM was the core of the DRA-FSN routing, the module NHSM gave full consideration to various factors including the status of the quality of communication between nodes, node moisture data packet transmission to the base station node capacity which expressed as the number of global routing beacons,

nodes bear the ability to send tasks, and the next hop node to acted as likely to transit nodes.

The key characteristics of DRA-FSN routing didn't require the base station node while processing node abnormalities. The NHSM adopted the historical value of the N rounds. The historical value of the abnormal nodes within the past N round would be zero if some node fails. So the NHSM could avoid these abnormal nodes while choosing the next hop. Over a period of time, the terminal could find those abnormal nodes because there were no packets from them.

III. DETAILED DESIGN OF DRA-FSN ROUTING ALGORITHM

This section described the details of DRA-FSN routing in establishing mechanisms, transmission mechanism, the global routing beacon post back mechanism and a sensor network timing control mechanism. The processes of the various mechanisms were specified one by one.

A. Routing Establishment

Route establishment process was shown in Algorithm 1:

*%Algorithm1: DRA-FSN routing establishment algorithm %

Step1: In the system initialization phase, the sink flooded hello packet in the whole network through its closed node a_{00} , and built networking gradient level field. At first, the initial state of the sink node was "Level = 0". All other nodes of the network level were initialized as "Level = ∞ " and were set in the receiver model, and all entries in the neighbor node information table of the node neighbour_nodeInfo were not activated. The sink node broadcasted Hello packets in the form of flooding that contained the node level, node ID number, the node number of global routing beacon, residual energy and the maximum possible number of network layers.

Step2: If the node clock reached the sleep time, then turned to step 12.

Step3: If the node clock reached the time of send packets or received packets from other nodes, then turned to step 11; if node clock has reached to the time of sending Hello packets, then turned to step 9.

Step4: If the node a_{zw} hadn't received a hello packet from the other node a_{ij} , then turned to step 2.

Step5: The node checked the destinations ID number of the Hello packet reception, if the destination ID was not its own ID number, then turned to step 2.

Step6: First of all, find corresponding node of a_{ij} entries in the information table neighbour_nodeInfo in the neighbors of a_{zw} according to the node ID number. If these corresponding entries were found, then updated this entry node level, the number of the node global routing beacon, the remaining energy and other information, and activated these entries. If did not find the appropriate entries, then created a new entry into this table, recorded the node level of the node a_{ij} level, node ID, the number of the node global routing beacon, residual energy information and activate this entry.

Step7: if $Level_{zw} > Level_{ij} + 1$, then set $Level_{zw} = Level_{ij} + 1$, recorded the time of receiving Hello packet time t_0 , calculated the node Hello packets sending delay t_1

using the formula (1), and turned to step9. if $Level_{zw} \leq Level_{ij} + 1$, then turned to step 8.

$$t_1 = Mod(i_d, C_0) * T \quad (1)$$

where, t_1 was the transmission delay, T was a constant which means the maximum delay according to the size of the network setting by the system, C_0 was a constant, i_d was the ID number of the node, and a transmission delay was obtained through the remainder.

Step8: a_{zw} did not send Hello packets, and discarded the Hello packet

Step9: if the node clock time $t < (t_0 + t_1)$, then turned to step3.

Step10: updated $Level_{zw}$, and wrote it into Hello packet, and then broadcasted Hello packet, turned to step 3.

Step11: The Hello packet would stop flooding after a certain delay. In all the active entries of a_{zw} 's neighbor nodes in table neighbour_nodeInfo, all nodes that satisfied $Level = Level_{zw} - 1$ or $level = Level_{zw}$ would be as a next hop candidate nodes.

Step12: The initialization phase algorithm was ended, all nodes began to sleep.

In addition to the sink node broadcasted a Hello packet when the system began running, the other source nodes broadcasted Hello packets only when their levels were updated. The update of the level was no infinite loop because of its monotonically decreasing. If there were some nodes with communication delay, the nearest node of the sink would be conducted to the level gradient field, and each node could build its level gradient field only sending a Hello packet. With the operation of the network, the node communication capability decreased with energy reduction. After a certain operating time, re-create network routing process may be necessary.

B. Routing Transmission

The farmland sensor has its own characteristics, such as the majority of information is periodically collecting. The time delay t_2 of the node transmission packet data was calculated by the formula (2).

$$t_2 = (M - N + 1) \times (M - 1) \times T \quad (2)$$

where, M represented the maximum number of layers of the whole network. It was obtained by the sink node which usually was greater than or equal to the maximum number of layers of the actual network. N was node gradient layers. T was same as formula (1), which means that the delay maximum. The whole data transmission process was described in Algorithm 2:

Algorithm 2: DRA-FSN routing transmission algorithm %

Step1: After a certain delay, the whole Hello data packet turned to the end of the flooding. The node a_{zw} began collection data, and calculated transmit packet delay t_2 according to the equation (2).

Step2: If the node clock arrived time $t_0 + t_2$, then turned to step 8.

Step3: If the node clock reached the sleep time, then turned to step 9.

Step4: If the network nodes a_{zw} hadn't received a packet from the other node a_{ij} , then turned go to step 2.

Step5: If the gradient layer of node a_{ij} was the same with the other same layer's gradient layers, then turned to step7.

Step6: Ignored all nodes in activated table neighbour_nodeInfo all entries in the same layer and all the nodes of the lower layer calculated their values P by equation (3). The maximum P value of the node was selected as the next hop, and modified the data packets.

Step7: Forwarded the data packet, recorded source node ID number, serial number and the ID number of the node a_{ij} , and turned to step 2.

$$P = \alpha_1 A_{one} + \alpha_2 A_{route} + \alpha_3 E + \alpha_4 C + [\alpha_5] \quad (3)$$

Step8: Recording source node ID number, serial number and a_{ij} ID number of the data packet

Step9: Node a_{zw} calculated the P value for all activated neighbours node of table neighbour_nodeInfo in the same layer and the lower layer of nodes by the formulas (3), and selected the node with maximum P value as the next hop.

Step10: Constructed the structural data packet and sent out, then turned to step 2.

Step11: Turned to sleep node.

C. Global Routing Beacon Return

The routing algorithm usually route in accordance with two types: based on the global information and based on local information.

DRA-FSN routing algorithm was based on local information. In order to increase the global performance of the DRA-FSN routing algorithm, the DRA-FSN routing algorithm used global beacon nodes to estimate. If a_{ij} received more global routing beacon, then the node has better global routing performance. It stood for a_{ij} had the higher successful rate to the sink.

There were two fields of ID number and serial number of the source node in packets. The source node was responsible for collecting the packet that was the original sender of this packet. The serial number was an incremental source node label which means the number of rounds in information collection. Two fields were uniquely identifies in a data packet, and these two fields would not change during the transmission in the entire sensor network.

The global routing beacon return included following steps, shown by algorithm 3.

Algorithm 3: DRA-FSN routing beacon return algorithm %

Step1: The sink would obtain sending node ID, source node ID number, and serial number, after it received a data packet. Then the sink constructed a global routing beacon packet and sent to the destination node with ID number of the sending node.

Step2: If the node a_{zw} received a global routing beacon packet, and the destination node ID was not equal to its ID number, then accepted this beacon packet. Or the node a_{zw} would discard this global routing beacon packet, then turned to step 7.

Step3: If the node a_{zw} received a global routing beacon packet, and the destination node ID was equal to its ID number, and then added one for the node number of global routing beacon, and turned to step 7.

Step4: the node a_{zw} found records from a neighbor information table neighbour_nodeInfo according to the source node ID number and serial number in the global routing beacon packet.

Step5: if the corresponding record had not been found, and then discarded this global routing beacon packet, and turned to step 7.

Step6: the node a_{zw} constructed a new global routing beacon packet and sent to the destination node that was the sending node in records, and turned to step 7.

Step7: the node a_{zw} continued to run the interrupted operation before receiving this global routing beacon packet.

D. Abnormal Node Handling

When wireless sensor networks appeared the failed node, such as energy depletion, failure, etc., the node was not working properly. In addition to collecting error data, the presence of the failed node would make the node was not working normally if it was as the next hop node. If the failure node was responsible for forwarding a large number of nodes, so that the affected nodes would be more. Therefore, how to deal with the problem of the failed node was an important issue to be considered in the routing algorithm.

DRA-FSN routing algorithm was based on local information and judgment for abnormal nodes was also based on local information. There are two reasons for the node a_{ij} if it had not received a single-hop response packet during last N times from the node a_{ij} data packets. One was the poor quality of the communication channel resulting in the loss of data packets, the other was nodes abnormal. The A_{one} parameters in the formula (3) could show whether the node the a_{ij} was abnormal in a certain extent, but it was not absolute. Whatever the reason, the node a_{zw} should not be the next hop of a_{ij} . In DRA-FSN routing algorithm, the node a_{zw} could judge whether the node was abnormal from its local collected information. Therefore, it could avoid the abnormal nodes, and reduced unnecessary data transmission to save energy.

E. Timing Control Mechanisms

Farmland wireless sensor networks has its particularity such as many farmland data was periodically transmitted. So the network nodes should be in a dormant state to conserve energy in the non-working period. In the routing initialization phase, the sink estimated the maximum gradient of the number of layers for the entire network, and sent out by the Hello packet. All nodes receiving hello packets calculated sending delay within a maximum delay value T . Many time delay parameters could be obtained by following status.

(1) The sink node sent hello packets in t_{init} , so that the nodes in the first layer could receive this package at time t_{init} .

(2) The nodes in the second layer could receive this packet between time t_{init} and $t_{init}+T$.

(3) The nodes in the m^{th} layer could receive the hello packet between time t_{init} and $t_{init}+(m-1)T$.

The routing initialization phase would end no later than $t_{init}+(m-1)T$. Due to the low-level node would be used as a forwarding node for the upper node, so the node number of n^{th} layer could send its own data packet after time $(m-n+1) \times (m-1) \times T$. Totally, the entire network could send out all packets during $t_{init}+m \times (m-1) \times T$. After that, the entire network at the moment entered into sleep mode until the next round of transmission.

IV. DRA-FSN ROUTING ALGORITHMS SIMULATIONS

In this paper, the simulation environment was set to $N \times N$ nodes in the network that the entire network into a square distribution. In this section, a comparative evaluation was conducted among DRA-FSN, DSDV and EAP [13] routing algorithms from energy efficiency, single data packet average power consumption and the packets distribution balance.

The energy of the base station was supposed to infinite, the base has no longer received any packet as the basis for wireless sensor networks death. The simulation results on various aspects of the network were run multiple times. The average of the obtained data was used to estimate by the simulation program. Energy efficiency was expressed by the number of packets which could be handled in a limited energy condition. The most important task in farmland water sensor networks is to maximize the water data collected from each node. The base station node receives the more packets, the more strong ability of the routing algorithm to transfer data packets.

A. Simulation of Energy Efficiency

Because the EAP routing was a clustering protocol and it would cause increasing packet length while fusing cluster data. Additional operation of the DSDV routing was to exchange routing table between the nodes, the larger the network size, the greater the routing table, so the more the extra energy consumed. DRA-FSN routing took full advantage of the characteristics of the farmland sensor network. It simplified the network initialization process, deletion of unnecessary operations, made an additional reduction in energy consumption, thereby enhancing the effectiveness of the network energy. An experiment in figure 2 showed that the DRA-FSN route was better than the EAP routing, and EAP route was better than DSDV routing. Additional operating in the DRA-FSN routing were the return packet broadcast Hello packets periodically and global routing beacons. Clustering would bring additional consumption.

The numbers of packets received by the three kinds of routing algorithm base station node were shown in Figure 2 without node abnormalities. From the horizontal view in figure 2, the number of packets received from the base station node was less affected by the size of the network. The received packet number was about 35,000 by the base station under different network size in running DRA-FSN routing protocol. The packet number was

about 12000 in DSDV routing protocol and was about 29000 in the EAP routing protocol.

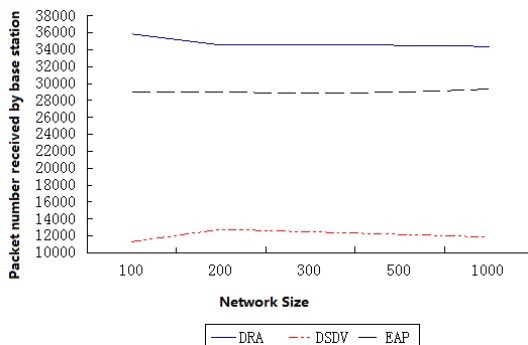


Figure 2. Amount of packets contrast received by the Base station

B. The Average Energy Consumption

Except for the energy consumption for the normal transmission of moisture packets, there were a series of auxiliary operations, such as the establishment of routing, packet acknowledgment, packet retransmission, packet transfer, and so on. These additional energy consumptions also related to the performance of the routing protocols. The average energy consumption of a single packet of routing protocols can reflect this performance to some extent, and its formula was given by equation (4).

$$E_{packet_ave} = \frac{E_{total}}{N_{receive}} \tag{4}$$

where, E_{packet_ave} was the average packet energy consumption, E_{total} was total energy consumption, $N_{receive}$ was the number of packets received in base station node.

The total energy consumption of the network was the total energy consumption of all nodes until the base station node could not receive packets in the network. The average energy consumption of the three routing protocols in a single package was shown in Figure 3 under different network size.

The average energy consumption of a variety of routing algorithm was increased with the network size increasing in Figure 3.

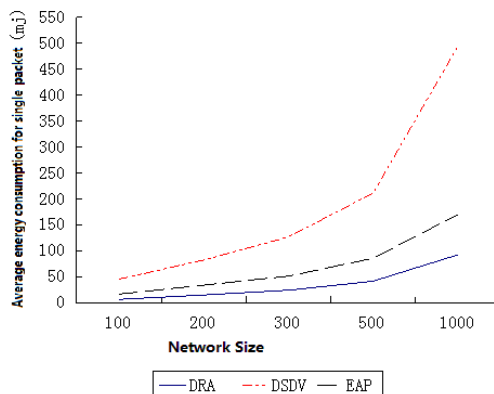


Figure 3. The average energy consumption per packet

Figure 3 indicated that the routing would be collected away from the base station node data back to a base station to consume more energy. Under different network size, the average energy consumption of a single packet the DRA-FSN route than the other two routing protocols. In the case of 1000 nodes, the average energy consumption of the DSDV routing a single packet rose rapidly, this is the result in the routing table increases due to the node increased between nodes exchange the sake of increased energy consumption of the routing table. This article by farmland sensor network deployment model analysis shows that all nodes in the data simply transmitted to the base station node, and do not need to know the routing of the other nodes. And farmland sensor networks take full advantage of the many-to-one feature to simplify the initialization process of the DSDV routing.

C. Packets Distribution Balance

The packet transmission can not guarantee without error, so the packets by multi-hop transmission were easy to lose. Data was difficult to reach the base station from the away moisture node, so that the packets distribution balance of moisture packets could be used as the routing protocol performance evaluation. The better balance of the routing protocol showed its better network performance. The packet distribution was shown in Figure 4, 5 in network size of 100, 300 nodes. In testing, data packet distribution had the biggest variable in the EAP route, and the DSDV routing had the smallest variance. This analysis showed that DRA-FSN routing in the node layer within a single hop factors played a role in.

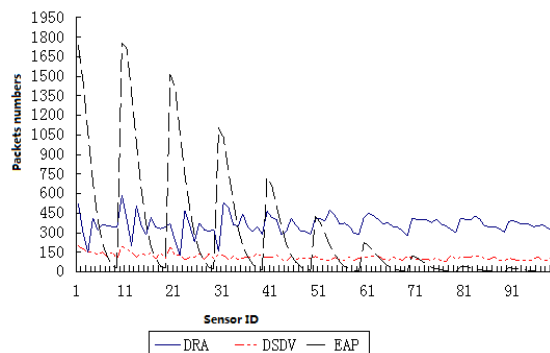


Figure 4. Packets distribution in 100 sensors

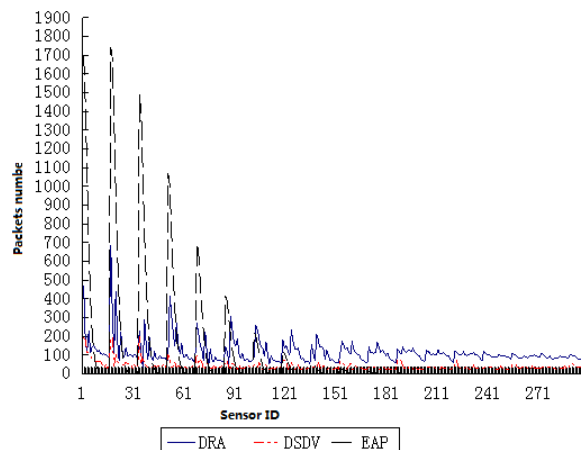


Figure 5. Packets distribution in 300 sensors

From Figure 4, 5 although the number of packets in the EEDRP - FMSN protocol was less than the EAP protocol, EAP routing protocol packets was over-concentrated in certain node. Most nodes appeared packet loss in the network in EAP protocol, so its balance was poor. In the farmland sensor network, the EAP routing protocol is the most desirable of the three routing protocols, and the DRA-FSN protocol was better than DSDV protocol under the same conditions. The DRA-FSN protocol extended the network life cycle than DSDV protocol.

V. CONCLUSIONS

In this paper, a Dynamic Routing Algorithm for Farmland Sensor Networks (DRA-FSN) was reported basing on channel quality control to improve energy efficiency, which combines the distance and communication characteristics of farmland wireless sensor network. The functional architecture of the DRA-FSN algorithm, routing establish the mechanisms, the communication transmission mechanism, the global routing beacon return mechanism, abnormal node handling mechanism and sensor networks timing control mechanisms were analyzed in detail in this article. The experiments showed that the DRA-FSN routing algorithm had high energy efficiency while reducing the average energy consumption of the data packet. The DRA-FSN routing algorithm was suitable for application farmland sensor network environment.

However, simulations were supposed to an agreed mode in transmitting power and error rate. Some deviation may be related to the real environment. So carrying out the DRA-FSN application into the real farmland will be the next major work of this project. And how to further improve the routing algorithm balance performance of the DRA-FSN routing is also an important work according to the operation of the network environment in the future.

REFERENCES

- [1] Sun Liming et al. *Wireless Sensor Networks*, Qinghua University Press, Beijing, 2005
- [2] Rong Z, Rappaport T S. *Wireless Communications: Principles and Practice*. Prentice Hall, 2002.
- [3] Santi P. Topology control in wireless ad hoc and sensor networks. *ACM Computing Surveys (CSUR)*, 2005, 37(2) pp. 164-194.
- [4] Vieira M A M, Coelho Jr C N, Da Silva Jr D C, et al. Survey on wireless sensor network devices, 2003.
- [5] Guo L Q, Xie Y, Yang C H, et al. Improvement on LEACH by combining Adaptive Cluster Head Election and Two-hop transmission., 2010
- [6] Jeong J, Ee C T. Forward error correction in sensor networks. *University of California at Berkeley*, 2003.
- [7] Ganesan D, Krishnamachari B, Woo A, et al. Complex behavior at scale: An experimental study of low-power wireless sensor networks. *Technical Report UCLA/CSD-TR 02*, 2002.
- [8] Shah R C, Rabaey J M. Energy aware routing for low energy ad hoc sensor networks, 2002.
- [9] Woo A, Tong T, Culler D. Taming the underlying challenges of reliable multihop routing in sensor networks., 2003
- [10] Ananstasi G, Conti M, Falchi A, et al. Performance Measurements of Mote Sensor networks. *ACM*, 2004
- [11] Kang W, Stankovic J A, Son S H. On Using Weather Information for Efficient Remote Data Collection in WSN., 2008.
- [12] Ahn J, Hong S, Heidemann J. An adaptive FEC code control algorithm for mobile wireless sensor networks. *Journal of Communications and Networks*, 2005, 7(4) pp. 489.
- [13] Tong M, Tang M. LEACH-B: An Improved LEACH Protocol for Wireless Sensor Network, 2010.
- [14] Willig H K A, Karl H. Protocols and architectures for wireless sensor networks. *John Wiley, New York*, 2005.
- [15] Liu M, Cao J, Chen G, et al. An energy-aware routing protocol in wireless sensor networks. *Sensors*, 2009, 9(1) pp. 445-462.

Dongfeng Xu received his B.Sc. degree in Computer Science and Technology in 1986 from Wuhan University of Technology. He is now an associate professor in the college of Informatics, South China Agricultural University, Guangzhou, China. His research interest includes network technology and Application.

DCSK Multi-Access Scheme for UHF RFID System

Keqiang Yue

Institute of VLSI Design, Zhejiang University, Hangzhou, Zhejiang, 310027, China
Email: yuekeqiang@163.com

Lingling Sun*, Bin You, and Shengzhou Zhang

Key Laboratory of RF Circuits and Systems, Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China
*Corresponding author, Email: (sunll, youbin)@hdu.edu.cn, zhangsz@sina.com

Abstract—DCSK modulation in chaos communication is a robust non-coherent modulation scheme. In this paper, the multiple-access DCSK scheme based on the OVFS code is proposed. Using the multiple-access DCSK scheme in RFID system, a DCSK-RFID system is presented. In the presented DCSK-RFID system, we use the DCSK for tag modulation for its low complexity and the simple receiver of the DCSK scheme is applied in reader part. The tag's BER performance of the proposed DCSK-RFID system is carefully generalized both in theoretic analysis and in simulations. From the simulation results, the theoretical and simulation values match closely with each other. Then, we design an anti-collision MAC protocol based on multi-access DCSK-RFID scheme. We theoretically analyze the throughput in given number of tags. The simulation shows that the proposed algorithm has better throughput than S-Aloha system.

Index Terms—RFID; Differential Chaos Shift Keying (DCSK); Multiple Access; PHY Layer Modulation; MAC Protocol

I. INTRODUCTION

The application of Radio Frequency Identification (RFID) technology is rapidly growing among different industries like supply chain management, inventory control, supermarket checkout process [1]. In these applications, RFID tags are attached to thousands of objects and their backscattered unique IDs signal are measured, analyzed, and identified by readers [2]. Unfortunately the RFID technique is suffering from several disadvantages. It is very sensitive to multi-tag interference, multiple tags collisions and privacy and security issues problem. Therefore, alternative MAC protocol and directly PHY layer designs will inevitably emerge and need to be evaluated [3].

A flexible MAC implementation means that anti-collision algorithm can be readily manipulated and effectively improve the system throughput. The exiting anti-collision schemes are built upon four multiplexing access technologies: Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Space Division Multiple Access (SDMA) and Code Division Multiple Access (CDMA). The ALOHA and

binary tree anti-collision algorithm [4] based on TDMA are the most popular multiple access method for RFID. In RFID applications with large tag populations, TDMA technique suffers from large number of collisions. The result is an increase in tag identification time and lower system efficiency. Due to its good multiple-access efficiency and high immunity against interference, CDMA technology is very attractive for RFID application [5-6]. In CDMA/RFID system, each tag is assigned an orthogonal spreading code, the reader distinguish different tags by utilizing the orthogonality of the spreading code. The application of CDMA technology in RFID is investigated as a mean to improve the system throughput.

The modulation techniques in RFID physical (PHY) system [7] are based on conventional digital modulation schemes like Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK), Phase Shift Keying (PSK) and Quadrature Amplitude Modulated (QAM). Reference [8] presents optimized ON/OFF states for the ASK modulated passive RFID tags, considering both the reader receiver sensitivity and the tag antenna mismatch conditions. Zaid Al-Amir [9] designs the RFID system using the concept of FSK modulation as a technique to transmit and receive the signal. Reference [10] designs an integrated circuit implementation of a BPSK backscatter modulator for passive RFID tags. References [11-12] exploit multi-state complex-valued load dependent scattering to yield QAM backscatter for passive or semi-passive tag that is compatible with the homodyne reader architecture.

Chaos-based communication systems with information embedded in chaotic signals are wideband, deterministic, non-periodic [13-14]. By contrast with the conventional digital modulation schemes, the chaotic modulation [15-16] can be implemented with extremely simple circuitry and produce different non-periodic waveform segments with low cross correlation properties. At the same time, the multiple access capability of chaos-based communication systems has been looked into [17-18]. The numerous features of chaotic communication are very attractive for RFID or WSN applications [19]. The single-user Differential Chaos Shift Keying (DCSK) [20]

technique is a robust non-coherent modulation scheme, which only requires frame or symbol rate sampling instead of channel estimation and the demodulation can be performed without synchronization. In this paper, a DCSK multiple-access scheme based on OVFS code is first proposed. Because the multi-access becomes an essential feature for practical implementation of the RFID system, using the multi-access DCSK scheme in RFID system is presented. In the proposed multi-access DCSK-RFID system, we construct the tag-to-reader PHY communication and design a new MAC protocol to perform the tag identification.

Our contributions can be summarized as follows:

1) We first present the multiple-access DCSK scheme based on the OVFS code;

2) Using the multiple-access DCSK in RFID tag-to-reader PHY communication, a multiple-access DCSK-RFID system is proposed. In the presented scheme, we use the DCSK modulation in RFID system for tag modulation because of the low complexity and the simple receiver of the DCSK scheme is applied in reader part. We construct the transmitter and receiver structure and deduce the BER performance of the proposed multiple-access DCSK-RFID system.

3) Because of the multi-access capability of the proposed scheme, we design a new MAC protocol based on multiple access DCSK for RFID to perform the parallelizable tag identification. We theoretically analyze the system throughput in given number of tags.

The rest of this paper is organized as follows. In Section 2, a new DCSK multi-access scheme is studied and the multi-access DCSK-RFID system is proposed. The parallelizable tag identification based on the DCSK multiple-access combined DFSA in UHF RFID is discussed in Section 3. Section 4 shows the simulation of the BER performance and anti-collision performance in DCSK-RFID system. Finally, we conclude the paper in Section 5.

II. DCSK-RFID PHY COMMUNICATION

A. Multi-access DCSK Scheme

With a differential shift keying modulator, DCSK uses a chaotic signal as the carrier for transmission. The chaotic signal is generated by the simple Logistic chaotic map chaotic circuit. The Orthogonal Variable Spreading Factor (OVFS) code is used as spreading because of its complete orthogonality and easy generation. A new multiple access DCSK scheme based on the OVFS is proposed. Fig. 1 shows the block diagram of the multiple access DCSK transmitter system. As can be seen from the figure 1, the binary DCSK modulation unit transmits a reference segment of the chaotic signal in the first half of the symbol duration. The second part of the bit serves as an information-bearing signal, depending on whether bit “-1” or “1” is being transmitted. The second part is spread by a unique OVFS code sequence. In multiple access DCSK modulation, the bit information is mapped to $b \in \{-1, 1\}$. Its signal before spreading can be written as:

$$g_m(k) = \begin{cases} x(k) & k = 2(l-1)N + 1 \dots (2l-1)N \\ b_l x(k-N) & k = (2l-1)N + 1 \dots 2lN \end{cases} \quad (1)$$

In this paper, the signal $g_m(k)$ is assigned a unique OVFS code sequence to spread, and the assigned codes are mutually orthogonal.

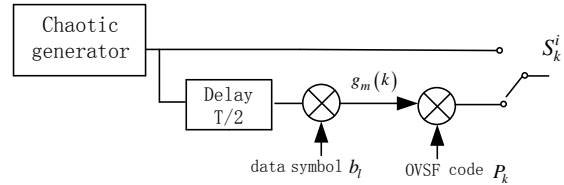


Figure 1. The multiple access DCSK transmitter

where the chaotic signal $x(k)$ is sent in first half bit; N is the spreading factor; $b_l \in \{-1, 1\}$ is the l th data symbol being transmitted.

At the multiple access DCSK receiver, the received signal is first correlated with its delayed version by $T/2$ and de-spread by the OVFS code. Then the information signal can be recovered from the sign of correlation measured at the output of the correlator. Fig. 2 shows the block diagram of the DCSK receiver system.

$$C_k = \sum_{k=(2l-1)N+1}^{2lN} r_k r_{k-N} \quad (2)$$

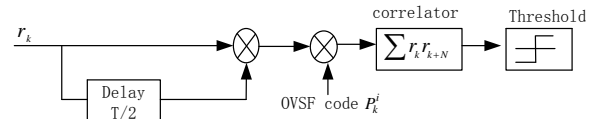


Figure 2. The multiple access DCSK receiver

B. DCSK-RFID Communication System

The typical RFID communication system uses the asymmetrical traffic loads between the uplink and the downlink. The commands and data broadcasted to all the tags are small from the reader in downlink, but in uplink, a great of number of tags transmit the rather heavy traffic to the reader. The multiple access DCSK modulation can be used in the tag-to-reader link to detect multiple tags simultaneously.

Considering RFID system employs the multiple-access DCSK scheme in tag-to-reader uplink communication to perform identification, in this paper, a DCSK-RFID system is designed. The proposed DCSK-RFID system consists of a reader and many tags, and assigns the mutually orthogonal OVFS codes for different tags in the same band simultaneously. On the basis of the considerations above, the proposed DCSK-RFID system architecture is illustrated in Fig. 3. The reader transmits commands to tags firstly. The tags respond their data frames with their system identifier field (SYS), the tag identifier field (ID), and the checksum field (CRC) individually to the reader by proposed DCSK modulation.

The DCSK transmitter integrated on the RFID uplink provides a robust, against multi-path and high security communication with low power and low complexity. We

use the DCSK modulation in the tag's modulation. A tag replies the information by DCSK modulation with the fixed data rate and either FM0 or Miller encoding is utilized to encode the backscattered data from tags back to readers. Because the simple demodulation of the DCSK scheme, it is applied in reader part. Transmissions from tags are independent of each other.

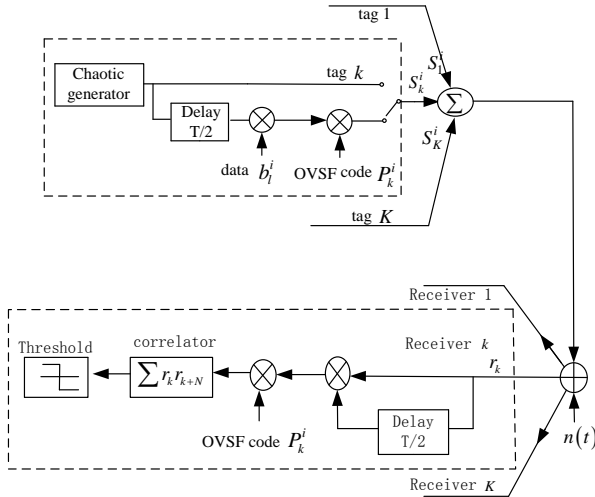


Figure 3. DCSK-RFID multiple access system

We formulate the signals transmitted by tags and the signal received by the reader herein after. We will consider the signals being transmitted in one symbol. The i th tag transmitter output function is as follow:

$$S^i(k) = \begin{cases} x^i(k) & k = 2(l-1)N+1 \dots (2l-1)N \\ b_i^j x^i(k-N) P_{k-(2l-1)N}^i + \xi_k & k = (2l-1)N+1 \dots 2lN \end{cases} \quad (3)$$

In this scheme, each tag is equipped with an OVSF code. By multiplying the OVSF code in the transmitter, the receiver data symbol in the system is capable of code-division multiple-access, which can be written as:

$$r_k = S_k + \xi_k = \sum_{i=1}^K S_k^i + \xi_k = \begin{cases} \sum_{i=1}^K S_k^i + \xi_k & k = 2(l-1)N+1 \dots (2l-1)N \\ \sum_{i=1}^K b_i^j x_{k-N}^i P_{k-(2l-1)N}^i + \xi_k & k = (2l-1)N+1 \dots 2lN \end{cases} \quad (4)$$

where ξ_k is Additive White Gaussian Noise (AWGN) with two-sided power spectral density $N_0/2$.

After being delayed for half of a bit period time, the signals multiply with the delayed signal and the de-spreading OVSF code at the i th tag receiver, then the i th user correlator output is:

$$C_l^i = \sum_{k=(2l-1)N+1}^{2lN} r_k r_{k-N} P_{k-(2l-1)N}^i \quad (5)$$

We replace the r_k with the (4), then (5) the can be rewritten as:

$$C_l^u = \sum_{k=(2l-1)N+1}^{2lN} \left(\sum_{j=1}^K b_j^j x_k^j P_{k-(2l-1)N}^j + \xi_{k+N} \right) \cdot \left(\sum_{i=1}^K x_k^i + \xi_k \right) P_{k-(2l-1)N}^u \\ C_l^u = \sum_{i=1}^K \sum_{j=1}^K \sum_{k=(2l-1)N+1}^{2lN} b_j^j x_k^i x_k^j P_{k-(2l-1)N}^i P_{k-(2l-1)N}^u \\ + \sum_{i=1}^K \sum_{k=(2l-1)N+1}^{2lN} x_k^i \xi_{k+N} P_{k-(2l-1)N}^u + \sum_{k=(2l-1)N+1}^{2lN} \xi_k \xi_{k+N} P_{k-(2l-1)N}^u \\ + \sum_{j=1}^K \sum_{k=(2l-1)N+1}^{2lN} b_i^j x_k^j P_{k-(2l-1)N}^j \xi_k P_{k-(2l-1)N}^u \quad (6)$$

The decoded data symbol from the u th user threshold is determined according to:

$$\hat{b}_l^u = \begin{cases} +1 & \text{if } C_l^u \geq 0 \\ -1 & \text{if } C_l^u < 0 \end{cases} \quad (7)$$

The BER performance [17] of the system can be derived as:

$$BER = \frac{1}{2} P(C_l^u < 0 | b_l^u = +1) + \frac{1}{2} P(C_l^u > 0 | b_l^u = -1) \\ = \frac{1}{2} \operatorname{erfc} \left(\frac{E\{C_l^u | (b_l^u = +1)\}}{\sqrt{\operatorname{var}\{C_l^u | (b_l^u = +1)\}}} \right) \\ = \frac{1}{2} \operatorname{erfc} \left(\left(\frac{K^2 - K}{N} + 4K \left(\frac{E_b}{N_0} \right)^{-1} + 4N \left(\frac{E_b}{N_0} \right)^{-2} \right)^{-\frac{1}{2}} \right) \quad (8)$$

where $E_b = 2N \operatorname{var}\{x\}$ is the bit energy, $\operatorname{erfc}(\cdot)$ represents the complementary error function.

III. THE DCSK MULTI-ACCESS MAC PROTOCOL

In RFID system, the specification of the data communication protocol determines the energy efficiency as well as the system throughput. Consequently, proper MAC protocol is essential for channel access and tag communication schedule in massive tags environment. In this paper, the next issue is to design an efficient MAC scheme to fully exploit DCSK multi-access capability.

In this section, an efficient anti-collision of UHF RFID system based on multiple access DCSK-RFID scheme is studied. In the proposed algorithm, the reader sends Query/ReQuery/AdjustQuery command (the first frame begins with a Query command, and the following frames begin with a ReQuery/AdjustQuery command) command to all tags to initiate the current frame identification. A Query command uses a Q value to set the number of slots in a frame. After receiving a Query, tags randomly choose a slot with the value in $[0, 2^Q - 1]$. When the slot value expires, the tag uses the DCSK modulation containing the OVSF codes to reply the reader, which allows the simultaneous acknowledgment of multiple tags.

When reader receives the data from the tags, there are three cases: 1) If there is an empty slot without any tag responses, the reader sends a non-acknowledgment (NAK)

restarts the identification process; 2) If tags are simultaneously responding to a reader's query in current time slot, Because of each tag equips with a unique quasi-orthogonal OVFS code, reader performs parallel identification process by using OVFS code correlation operations. The reader allows reception of the signals from different tags that overlap in time. This means that the tags are transmitting data within the same time range and frequency band, which can achieve high system throughput; the reader sends an acknowledgment (ACK) .3) If some reply tags use the DCSK scheme in the same OVFS codes, which means tags are beyond the identification ability, indicating that there is a collision, the collision occurs. The reader restarts the identification process;

When the identification process in current frame is over and the other tags are not identified, the reader issues AdjustQuery or ReQuery command to restart the identification process

IV. PERFORMANCE ANALYSIS

A. BRR of the Proposed DCSK-RFID

In this section, the performance of the proposed DCSK-RFID multi-tags communication system is studied. We compare the computer simulation with the theoretic results. In simulation, a great number of bits are transmitted for each user. Fig. 4 shows the BER performance of the DCSK-RFID system.

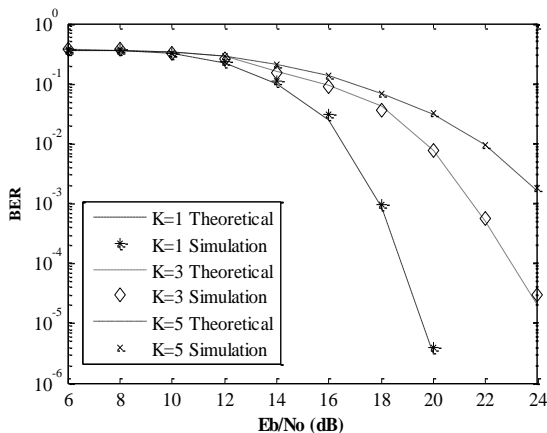


Figure 4. BER performances Verse SNR

We set the spreading factor $N = 128$ and the number of users $K = 1, 3, 5$ respectively. From the figure above, the consistency between the analytical BER and the simulated BER is clearly evidenced. For each tag data, as expected, the BER generally improves with the average E_b / N_0 increasing and converges to a constant $1/2\text{erfc}(N/(K^2 - K))$ which is determined by N and K .

Fig. 5 plots the BER performances against the number of the tags under the specific spreading factor N ($N = 64, 128, 256$ respectively), where we set $E_b / N_0 = 18$ in the DCSK-RFID system.

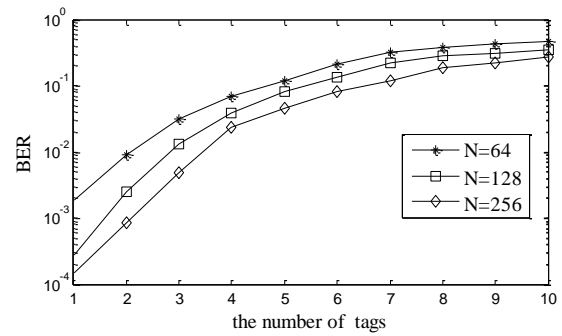


Figure 5. BER performance Verse the number of tags

From the figure, we can see that for a given spreading factor N , the BER performance decreases with the number of users K increasing; with the increase of spreading factor N , the BER performances improves.

B. The Throughput of the DCSK Multi-access MAC Protocol

In this section, we derive the throughput of the proposed protocol. When the tags use different spreading codes in DCSK modulation, they can simultaneously communicate with the reader. Considering M tags communicate with the reader in a certain time slot and available use N spreading codes, and the reader can successfully de-spread the average number of the tag as follow:

$$M_{\text{success}} = M \cdot \left(1 - \frac{1}{N}\right)^{M-1} \quad (9)$$

The Poisson distribution can be used as the approximation of a binomial distribution when the number of trials goes to infinity and the expected number of successes remains fixed [4]. When the appearing tags are Poisson distribution, in current time slot the probability of M tags is:

$$P_M = \left(\frac{G^M}{M!}\right) e^{-G} \quad (10)$$

The throughput of the proposed DCSK-RFID protocol can be expressed as:

$$S = \sum_{n=0}^{\infty} M_{\text{success}} \cdot P_M \quad (11)$$

Making the equation (9) and (10) into equation (11);

$$\begin{aligned} S &= \sum_{n=0}^{\infty} M \cdot \left(1 - \frac{1}{N}\right)^{M-1} \cdot \left(\frac{G^M}{M!}\right) e^{-G} = G e^{-G} \sum_{n=0}^{\infty} \left[\left(1 - \frac{1}{N}\right) \cdot G\right]^M / M! \\ &= G e^{-G} e^{\left(1 - \frac{1}{N}\right)G} = G e^{-\frac{G}{N}} \end{aligned} \quad (12)$$

Let $\partial S / \partial G = 0$, the maximum throughput S can be achieved when $G = M$, and the max S is $S_{\text{max}} = N e^{-1}$. The maximum throughput of the proposed DCSK-RFID scheme is N times than S-ALOHA. For example, when the number of OVFS code is 32 in the multi-access DCSK scheme, the maximum throughput is 11.78. The

maximum throughput of the proposed DCSK-RFID scheme increases with the increase of the number of the OVSF code.

The Fig. 6 gives the throughput comparison between the proposed multi-access DCSK-RFID scheme and the S-ALOHA. In the proposed multi-access DCSK scheme, the number of the OVSF code is 2, 4, 8 respectively.

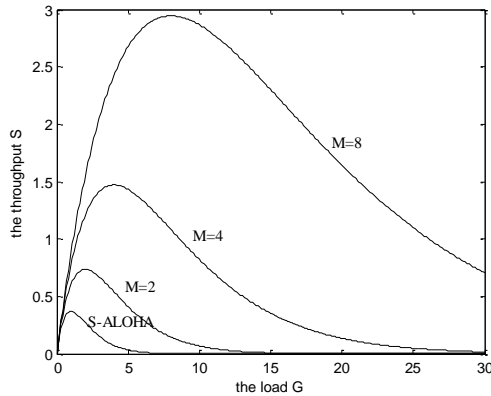


Figure 6. The throughput comparison between the proposed DCSK scheme and the S-ALOHA

From the Fig. 6, we can see that when the number of the OVSF code is 2, the throughput is 0.74; the number is 4, the throughput is 1.47; the number is 8, the throughput is 2.94. The throughput of the proposed DCSK-RFID scheme is better than the slot ALOHA algorithm. The figure 7 also shows that when the $G < N$, the throughput of DCSK-RFID protocol increases; when $G > N$, the throughput of DCSK-RFID protocol begins to decrease. This because the number of tags is larger than the number of the number of the OVSF code, the possibility of tag collisions increases.

V. CONCLUSION

In this paper, due to the easiness of generation and random like nature, a multiple-access DCSK modulation scheme based on OVSF code is presented. Using the multiple-access DCSK modulation in RFID system, a DCSK-RFID digital communication system is proposed. We evaluate the tag's BER performance both in theoretic analysis and in simulations in DCSK-RFID system. Based on the proposed DCSK-RFID system, we study an anti-collision MAC protocol supporting multiple tag identification. We theoretically analyze the throughout of the presented protocol in given number of tags. The simulation result shows that the maximum throughput of the proposed protocol is N times than S-ALOHA.

VI. ACKNOWLEDGMENT

This work was supported by Major State Basic Research Development Program of China (973 Program, No.2010CB327403), The Key Technology Research of Miniaturization UHF RFID Reader (2012C21043).

REFERENCES

[1] K. Finkensteller, RFID handbook: fundamentals and applications in contactless smart cards and identification, 3rd ed. John Wiley & Sons Ltd, 2006.

[2] Hung, Jason C. "Using Active RFID to realize Ubi-Media system". *Journal of Networks*, vol. 6, no. 5, pp. 743-749, MAY 2011.

[3] Michael Buettner, David Wetherall. "A Software Radio-based UHF RFID Reader for PHY/MAC Experimentation". *2011 IEEE International Conference on RFID*, pp. 134-141, APR. 2011.

[4] Dheeraj K, Klair, K. -W. C., and Raad. "A Survey and Tutorial of RFID Anti-Collision Protocols", *IEEE Communications Surveys & Tutorials*, vol. 12, no. 3, pp. 400-421, 2010.

[5] Carlo Mutti, Christian Floerkemeier. "CDMA-based RFID Systems in Dense Scenarios: Concepts and Challenges", *2008 IEEE International Conference on RFID. The Venetian, Las Vegas, Nevada, USA*, pp. 215-222, 2008.

[6] Gustaw Mazurek. "Active RFID System with Spread-Spectrum Transmission", *IEEE Transactions on Automating Science and Engineering*, vol. 6, no. 1, pp. 25-32, 2009.

[7] Shenchih Tung and Alex K. Jones. "Physical Layer Design Automation for RFID Systems". *IEEE International Symposium on Parallel and Distributed Processing., IPDPS 2008.*, pp. 1-8, 2008.

[8] Yao Xi, Sungwook Kwon, Hyungchul Kim. etl. "Optimum ASK Modulation Scheme for Passive RFID Tags Under Antenna Mismatch Conditions". *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 10, pp. 2337-2343, OCTOBER 2009.

[9] Shipu Zheng, Fengqi Yu, and Yuesheng Zhu. "A Novel RFID Transceiver Architecture with Enhanced Readability". *International Conference on Wireless Communications, Networking and Mobile Computing, WiCom 2007*. pp. 2074 – 2077, 2007

[10] Nowshad Amin, Ng Wen Jye, and Masuri Othman. "A BPSK Backscatter Modulator Design for RFID Passive Tags". *RFIT 2007-IEEE International Workshop on Radio-Frequency Integration Technology, Singapore*. pp. 262-265 Dec. 2007.

[11] Stewart J. Thomas, Eric Wheeler, Jochen Teizer, and Matthew S. Reynolds. "Quadrature Amplitude Modulated Backscatter in Passive and Semi-passive UHF RFID Systems". *IEEE Transactions on Microwave Theory and Techniques*, vol. 60, no. 4, pp. 1175-1182, APRIL 2012

[12] Colby Boyer, Sumit Roy. "Coded QAM Backscatter Modulation for RFID", *IEEE Transactions on Communications*, vol. 60, no. 7, pp. 1925-1934, JULY 2012.

[13] Farmer, Michael E. "A chaos Theoretic analysis of motion and illumination in video sequences", *Journal of Multimedia*, vol. 2, no. 2, pp. 53-64, 2007.

[14] Liu, Rui Tian, Xiao-ping. "A space-bit-plane scrambling algorithm for image based on chaos", *Journal of Multimedia*, vol. 6, no. 5, pp. 458-466, 2011.

[15] F. C. M. Lau, M. M. Yip, C. K. Tse, and S. F. Hau. "A Multiple-Access Technique for Differential Chaos Shift". *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 49, no. 1, pp. 96-104, JAN. 2002.

[16] Hui LI, Xuchu DAI, Peixia XU. "A CDMA Based Multiple-Access Scheme for DCSK". *IEEE 6th CAS Symp. an Emerging Technologies: Mobile and Wireless Comm.*, pp313-316, June, 2004.

[17] Zhibo Zhou, Tong Zhou and Jinxiang Wang. "Performance of a Chaotic Binary Sequence Based Multiple- Access DCSK Communication System". *Proceedings of the 2008 IEEE International Conference on Information and*

Automation, Zhangjiajie, China. pp. 1242-1245, June 2008.

- [18] Ying Shi; Yiping Chen; Yigang Zhou; Youyong Liu. "An improved scheme for multiple access differential chaos-shift keying system", *4th IEEE International Conference on Circuits and Systems for Communications, ICCSC 2008*. pp. 358-361, 2008.
- [19] Chia-Chin Chong, Su Khiong Yong. "UWB Direct Chaotic Communication Technology for Low-Rate WPAN Applications". *IEEE Transactions on Vehicular Technology*, vol. 57, No. 3, pp. 1527-1536, 2008.
- [20] Weikai Xu, Lin Wang, Guanrong Chen. "Performance of DCSK Cooperative Communication Systems Over Multi-path Fading Channels". *IEEE Transactions on Circuits and Systems-I: Regular papers*, vol. 58, No. 1, pp. 196-204, JAN. 2011.



Keqiang Yue was born in Henan Province, China, in 1984.

He received the B.S. degree in electronic engineering from the Anyang Normal University, Anyang, China, in 2007; the M.S. degree in communication and information system, Hangzhou Dianzi University, Hangzhou, China, in 2010, and is currently working toward

the Ph.D. degree in circuit and systems at the Zhejiang University, Zhejiang, China

His research interests include wireless communication and RFID anti-collision.



Lingling Sun (SM'01) received the B.S. degree in microwave telecommunications from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982, and the M.S. degree in circuit and system from the Hangzhou Institute of Electronics and Engineering, Hangzhou, China, in 1985.

She is currently a Professor with the Institute of Microelectronics, Hangzhou Dianzi University, Hangzhou, China. She has authored or coauthored over 100 technical papers. Her major research interests include the design and CAD of RF/microwave integrated circuits (ICs) and the research of ICs and systems such as the RF/microwave device modeling, the RF/microwave broadband power amplifier design, the research of the RFIC CAD technology, and the development of an electronic design automation (EDA) tool for RFICs.

Bin You was born in China, in 1974. She received the Ph.D. degrees from Shanghai University. She is currently a vice Professor in Hangzhou Dianzi University. Her research interests include the modeling of passive/ active devices and the design of RF/MMICs.

Shengzhou Zhang was born in Shangdong Province, China, in 1983. He received M.S. degree in circuits and system from Nanjing university of Science and Technology, Nanjing, China.

His research interests include the modeling and design of passive devices in RF CMOS.

An Effective Scheme for Performance Improvement of P2P Live Streaming Systems

Xiaosong Wu, Xingshu Chen*, and Haizhou Wang

Network and Trusted Computing Institute, College of Computer Science, Sichuan University, Chengdu, China

*Corresponding author, Email: wuxiaosong126@126.com, chenxsh@scu.edu.cn, whzh.nc@qq.com

Abstract—To solve the problems of long start-up latency, low playback continuity and data distribution rate in P2P live streaming system, this paper proposes two optimization strategies. One is an adaptive peer selection algorithm based on peers' real-time service ability. In the algorithm, a simple but effective method is proposed to calculate the peers' ability. And source peer adjusts the amount of requested data and the priority of destination peers according to the destination peers' abilities and the limit of out-degree. Thus a load balance and efficiently system can be constructed. The other one is a push-pull combination data distribution mechanism. This mechanism adaptively chooses push or pull model to distribute data fast and efficiently. The results show that the strategies have a significant effect on reducing start-up latency and control message overhead and improving the playback continuity and data distribution rate.

Index Terms—P2P; Live Streaming; Data Distribution; Peer Selection; Quality of Service

I. INTRODUCTION

With the popularity of the Internet and the development of communication technology, video streaming services are more and more popular [1]. But the traditional Client/Server(C/S) model does not meet the requirement of video streaming service due to the limitation of the Server's performance and bandwidth. Compared to C/S model, P2P [2] has the features of off-center, scalability, high performance-price ratio, strong robustness and load balancing. What's more, P2P has many other advantages, such as the higher utilization of network resources, the elimination of bottleneck caused by central servers and no single-point failure. Therefore, various P2P live streaming systems [3] emerge in endlessly, such as PPStream, PPLive and UUSee, which have achieved a huge success. However, the problems of long start-up latency and playback lag, low playback continuity and data distribution rate in P2P live streaming system need to be further studied to satisfy the high quality of service (QoS) requirements.

Many researches have been done to solve the problems mentioned above. And the peer selection algorithm, data distribution mechanism, network topology and the matching between the P2P layer and the physical layer are hot research area. For example, references [4] and [5] proposed a dual mix live streaming architecture, which mixes content distribution network (CDN) with P2P and

mixes tree structure with mesh structure to reduce the long start-up latency. Ref. [6] proposed an improved simulated annealing algorithm to optimize traditional data distribution strategy to satisfy the demand of the playing and the download of the neighbors. Ref. [7] proposed a cache replacement algorithm, which replaces data block based on the requests of other peers, to improve the data request hit rate and avoid the data redundancy. Ref. [8] combined with IP multicast to reduce the load of streaming servers and backbone network and increase the scalability and availability. Ref. [9] designed a new anomaly detection mechanism to detect the abnormal state of the peers and Ref. [10] researched the maximum download speed of the system in the limit of peer out-degree and developed a stochastic flow model to solve the churn problem in P2P live streaming system. Ref. [11] proposed a frame of P4P (provider portal for applications) to reduce the huge network overhead. The P4P framework allows ISP and P2P applications collaborate to achieve a more efficient network traffic control and to reduce the cost, achieving a win-win situation.

The rest of this paper is organized as follows; section II describes a P2P live streaming system named NTLive, which was developed by the Network and Trusted Computing Institute of Sichuan University. Section III gives two optimization strategies to reduce the start-up latency and control message overhead, improve the playback continuity and data distribution rate. Section IV applies the optimization strategies to the NTLive system and compares the performance before and after the optimization. The experimental results show the two optimization strategies have a significant effect on reducing start-up latency and control message overhead and improving the playback continuity and data distribution rate. Section V summarizes this paper and show the future work.

II. INTRODUCTION OF NTLIVE

NTLive is a free P2P live streaming system. It is developed by the Network and Trusted Computing Institute of Sichuan University. It provides stable and smooth live and video-on-demand streaming services to the China Education and Research Network users. Compared to the traditional streaming media system based on C/S model, NTLive has the following advantages: the more users there are, the smoother the system is, and it supports large-scale simultaneous access

due to the P2P-streaming technology. NTLive has been deployed in the Sichuan University education network. By November 21, 2012, it had been downloaded 1675 times and the website of NTLive had been visited 9715 times.

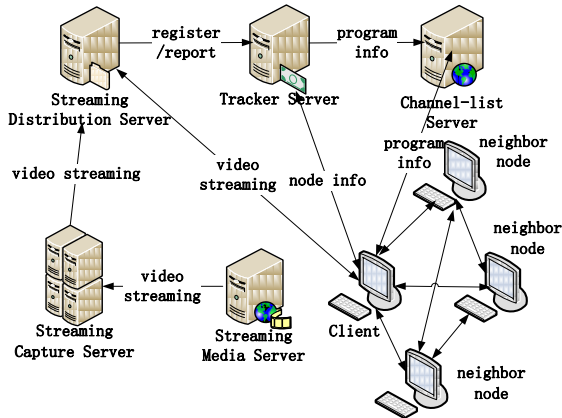


Figure 1. Architecture of NTLive

NTLive is mainly composed of 5 modules, including Streaming Capture Server, Streaming Distribution Server, Tracker Server, Channel-list Server and Clients, as shown in Fig. 1. There is a centralized directory server named Tracker in this system and the Tracker Server maintains the information of all programs and peers in the system. The client register to Tracker when it startups and then it obtain programs' information from the Channel-list Server. When client chooses a channel to play, Tracker returns a list of peers which are playing the same channel and the Streaming Distribution Server is in the list. Then the client requests data from the peers in the list. The peers exchange their neighbor information and data information during the play, and they also report to Tracker to update their information periodically.

NTLive is high robust and stable and it has a random stability mesh topology because of the random peer selection algorithm. But, this algorithm ignores the differences of stability and service abilities among peers. For instance, some peers nearly can't upload data due to the low bandwidth, but there may be many peers request data from them randomly, resulting in a lot of request failure. On the other hand, there are not enough requests to the high upload-bandwidth peers, wasting their upload ability. Thus, there will be low bandwidth utilization, high rejection rate of data request, long playback lag and low speed of data distribution in the system. What's more, NTLive uses pull model to distribute data. This mode can avoid distributing duplicate data, but it will increase the control message overhead and consume bandwidth. In addition, pull model is slower than push model, leading to longer start-up latency and playback lag.

So, we proposed the following strategies to solve the problems caused by random peer selection algorithm and pull model data distribution mechanism in the P2P live streaming systems.

III. OPTIMIZATION

A. An Adaptive Peer Selection Algorithm

Many P2P living streaming systems used random peer selection algorithm and ignored the different abilities among peers like NTLive. So these systems were faced with the problems of unreasonable resources distribution and low resource utilization. Some improved peer selection algorithms [12-16] had been proposed, but they almost only focused on the peers' inherent abilities and greedily chosen the highest service peers and ignored the dynamic changes of the network. As a result, these algorithms can't adapt to the reduction of service ability caused by sudden network congestion, and just let the source peer choose another higher service ability peer.

In this paper, we proposed an adaptive peer selection algorithm based on the peer real-time service ability to solve the problems above mentioned. In this algorithm, a simple but effective method is proposed to calculate the peers' ability. And the source peer dynamically adjusts the number of requested data blocks and the priority of destination peer according to the destination peers' abilities and the limit of out-degree. Thus the system can adapt to the change of network environment and the service ability can be taken full use of. The algorithm is described as follows:

Definition 1, in our system, we divide video data into small blocks with the same size to distribute. We call each block as *data block* and give a unique identifier to each *data block*.

Definition 2, set the service ability of peer B to peer A as V_{B-A} . The initial value of V_{B-A} is 0. If A requests N_{A-B} data blocks from B, then $V_{B-A} = V_{B-A} - 2 * N_{A-B}$.

Definition 3, set T_{B-A} as the time from peer A sends request to peer B to peer B returns data to A. If B doesn't return data, then $T_{B-A} = -1$.

Definition 4, set two time threshold T_0, T_1 ($T_0 < T_1$) and three number threshold N_1, N_2, N_3 to request data blocks.

If $0 < T_{B-A} < T_0$, then $V_{B-A} + 3$, and if $N_{A-B} < N_3$, then $N_{A-B} + 1$;

If $T_0 \leq T_{B-A} \leq T_1$ then $V_{B-A} + 2$;

If $T_{B-A} > T_1$ then $V_{B-A} + 1, N_{A-B} - N_1$;

If $T_{B-A} = -1$, then $V_{B-A} - 1, N_{A-B} - N_2$; If $N_{A-B} < 0$, then $N_{A-B} = 0$;

If the value of T_{B-A} always equals -1 in continuous three request cycles, then B can be considered as having left the system and A disconnect the connection with B.

In this way, the source peer can adaptively adjust the number of requested data blocks according to the data return time when the network changes. As a result it can help to alleviate the network congestion and improve the system's performance.

Definition 5, set the online time of peer A as T_A , the out-degree of A is D_OUT_A , the in-degree of A is D_IN_A , and the max out-degree of A is MAX_OUT_A ,

the max in-degree of A is MAX_IN_A . Out-degree is the number of peers which A simultaneously requests data from. In-degree is the number of peers which A simultaneously sends data to. Initially, $D_OUT_A = 1$, $D_IN_A = 5$, and $1 \leq D_IN_A \leq MAX_IN_A$.

If each data block in send queue is sent in less than T_2 time in continuous three send cycles, then $MAX_OUT_A + 1$.

If there is a data block is not sent out in more than T_3 time in continuous three send cycles in the sending queue, then $MAX_OUT_A = D_OUT_A - 1$. And A will reject the data request from its neighbor whose service ability is the lowest in A's neighbor lists and A disconnect the connection between them, then $D_OUT_A - 1$.

When $D_OUT_A = MAX_OUT_A$, if peer B request data from A and the average service ability of B is V_B , then A will choose the lowest service ability peer from the connected neighbor peers, whose service ability is V_C . Compare V_B with V_C , If $V_B > V_C$, A disconnects the connection with C and establish a new connection to B. Otherwise; A rejects the data request from B.

If the required number of data blocks of A can be sent by $D_IN_A - 1$ peers, then A will disconnect a connection, and $D_IN_A = D_IN_A - 1$. Otherwise, A request data from other peers, and $D_IN_A + 1$. Since some peers cannot get enough data blocks due to their low download bandwidth, they may connect to other peers to request data endlessly, result in wasting the system's resources and increasing the control message overhead. So, limit $D_IN_A \leq MAX_IN_A$.

Definition 6, if the service ability of A saved in Tracker Server is V_A , assume A sends data blocks to n peers $B_1, B_2, B_3, \dots, B_n$, and the service abilities of A to each peer are $V_1, V_2, V_3, \dots, V_n$, then

$$V_A = \sum_{i=1}^n \frac{V_i}{n} \tag{1}$$

The sorting rules of the peer-list in the Tracker Server are as follows:

The peer whose out-degree has not reached the maximum is sorted in the front of the peer whose out-degree is full.

Under the rule 1), the peer whose service ability is higher sorts more front.

Under the rules 1) and 2), the peer whose online time is longer sorts more front.

When a peer starts to play a live channel, it first chooses N highest service ability peers whose out-degree are less than max out-degree and M highest service ability peers whose out-degree equal max out-degree from the peer-list to request data. Then it selects peers in accordance with the above definitions. A peer also requests data from the Streaming Distribution Server when it is necessary.

In short, in this algorithm, the source peer always chooses the highest service ability peer to establish a new connection and requests more number of data blocks from the higher service ability peer. And this algorithm adapts to the network changes by dynamically exchanging the number of requested data blocks but not by disconnecting the connection to alleviate network congestion. And it will help to construct a load balance and effective system by set the in/out degree and dynamic change the value to match the peer's ability.

B. A Push-Pull Combination Method for Data Distribution Mechanism

NTLive uses pull model to distribute data. This model can avoid distributing duplicate data, but it will increase the control message overhead and consume bandwidth. In addition, pull model is slower than push model, leading to a longer start-up latency and playback lag. This paper proposes a push-pull combination method to distribute data and combine the advantages of both models. This data distribution mechanism is different from the other pull-push strategies [17-21]. It adaptively chooses the push or pull model to distribute data according to the source peer's data requirements and the destination peer's service abilities. In this way, the mechanism can not only avoid distributing duplicate data, but also decrease the control message overhead and reduce the start-up latency. This push-pull combination mode is described as follows:

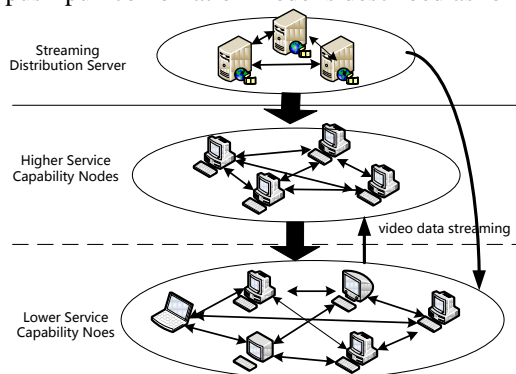


Figure 2. Mesh topology architecture with hierarchy

Rule 1, when a peer A registers to Tracker to start playing a live channel, Tracker not only returns response message to A, but also sends message to another n highest service peers, and let them to push the required data to A.

Rule 2, if there is a peer B whose service ability to peer A (N_{B-A}) is high, and the number of data blocks requested from A to B is the maximum N_3 in more than continuous three request cycles. Then peer A sends request message to peer B to get data blocks at a fixed interval n in the next request cycles. For example, if A requests the data blocks whose identifier are in the interval [a, b] in current request cycle, then B will initiative push the data blocks whose identifiers are in the interval [a + i*n, b + i*n] to A, in the following i (i=1, 2, 3, 4...) request cycles. And B will initiative and

continually push data to A until $T_{B-A} > T_1$. A will requests the other data blocks from the other peers.

Rule 3, Streaming Distribution Server (SDS) requests the top 20 peers from Tracker every 5 minutes. These peers are sorted by online time in the Tracker's peer-list. Then SDS distributes its newest data blocks to those 20 peers. In this way, the new data can be quickly distributed to the whole P2P system.

What's more, it can help to construct a hierarchical mesh topology by using the optimized strategies of peer selection algorithm and data distribution mechanism. As shown in figure 2, the SDS is at the highest layer, and the higher service peers are closer to SDS than the lower service peers. But there is not a strict boundary between high-layer and low-layer in the system.

IV. MEASUREMENT METHODOLOGY AND RESULTS

We applied the optimization strategies to the NTLive live streaming system and compared the performance of start-up latency, playback continuity, data distribution rate and control message overhead before and after the optimization. The measurement scheme is as follows.

A. Measurement Scheme

In our experiment, the Tracker Server was deployed on a server with eight 2.80GHz Xeon(TM) CPUs and 8GB memory. The Streaming Distributions Server was deployed on four servers with eight Xeon(R) CPUs and 8GB memory at Sichuan University of China with 1.0 Gbps Ethernet network access. The users of the NTLive system mainly come from Sichuan University.

NTLive is average used about 400 times per day, which have reached the basic requirements of the experiment. We chose the live channel CCTV-5 as the measure object, which had been totally played more than 30, 000 times before the experiment. We selected 2000 data sets measured in November 23, 2012 to November 27. Every set contains data of start-up latency, control message overhead, playback continuity and the data distribution rate. This paper analyzed the statistical data of the experiment results, and didn't analyze the optimization effect on a single peer.

TABLE I. PARAMETER VALUE

Parameter	T_0	T_1	T_2	T_3	MAX_IN	N
Value	100ms	500ms	500ms	2s	15	3
Parameter	M	N_1	N_2	N_3	MAX_OUT	
Value	2	2	5	20	5	

The value of the parameters mentioned in section III. A are shown in table I and they were obtained by analyzing the related data in the database of NTLive.

B. Measurement Results

1) Start-up Latency

The definition of peer start-up latency in this paper is the time from double click a live channel to it starts play.

The start-up latency is shown in table II. We show the comparison of before and after optimization in Fig. 3 and Fig. 4.

The vertical axis of Fig. 3 shows the percentage of the number of sets whose start-up latency is in the interval time of the horizontal axis. Fig. 4 shows the cumulative distribution function of start-up latency before and after optimization. The number of sets whose start-up latency was less than 2s increased from 405 to 460, namely 14.43%. And the number of sets whose start-up latency is more than 4s reduced from 222 to 190, namely 14.41%. From the data above, we can see that the optimization strategies made more peers start up in a shorter time and reduced the number of peers with long start-up latency.

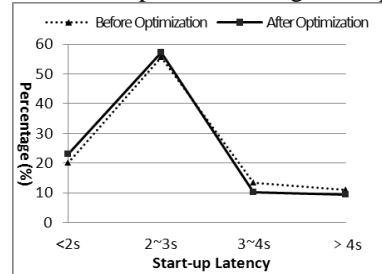


Figure 3. The line chart of start-up latency before and after optimization

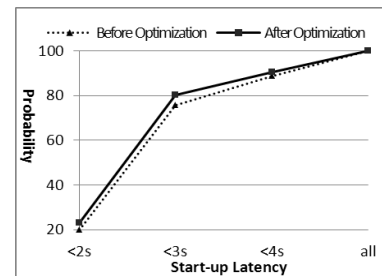


Figure 4. The cumulative distribution function of start-up latency before and after optimization

2) Playback Continuity

We use the total buffer time during a living channel playback to measure the playback continuity. The buffer time is calculated as follows,

$$T = \sum_{i=1}^n T_i \tag{2}$$

In formula (2), n represents the total times that the buffering percentage (B) is less than 100%, T_i represents the i-th buffer time. The calculation of buffering percentage B is as follows,

$$B = N_{got} / 20 * 100\% \tag{3}$$

In formula (3), N_{got} represents the number of the data blocks which are the continuous 20 blocks after the current playing data block and have been received.

The buffer time is shown in table III and the comparison of before and after optimization is shown in Fig. 5 and Fig. 6.

The vertical axis of Fig. 5 shows the percentage of the number of sets whose buffer time is in the interval buffer time of the horizontal axis. Fig. 6 shows the cumulative distribution function of buffer time before and after optimization. The number of sets whose buffer time was

TABLE II. THE PEER START-UP LATENCY BEFORE AND AFTER OPTIMIZATION

Start-up Latency (s)	Number Of Sets		Percentage (%)		Growth Rate (%)
	Before Optimization	After Optimization	Before Optimization	After Optimization	
≤2	402	460	20.10	23.00	14.43
≤3	1512	1606	75.60	80.30	6.22
≤4	1778	1810	88.90	90.50	1.80
>4	222	190	11.10	9.50	-14.41

TABLE III. BUFFER TIME BEFORE AND AFTER OPTIMIZATION

Buffer Time (s)	Number Of Sets		Percentage (%)		Growth Rate (%)
	Before Optimization	After Optimization	Before Optimization	After Optimization	
≤1	1471	1586	73.55	79.30	7.82
≤5	1805	1837	90.25	91.85	1.78
≤10	1874	1881	93.70	94.05	0.37
>10	126	119	6.30	5.95	-5.56

TABLE IV. DISTRIBUTION RATE BEFORE AND AFTER OPTIMIZATION

Data Distribution Rate (%)	Number Of Sets		Percentage (%)		Growth Rate (%)
	Before Optimization	After Optimization	Before Optimization	After Optimization	
>100	563	648	5.40	5.30	15.10
>50	1118	1334	12.15	11.70	19.32
>20	1649	1660	26.55	16.30	0.67
>0	1892	1894	27.75	34.30	0.11
0	108	106	28.15	32.40	-1.85

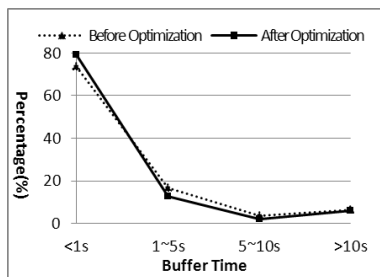


Figure 5. The line chart of buffer time before and after optimization

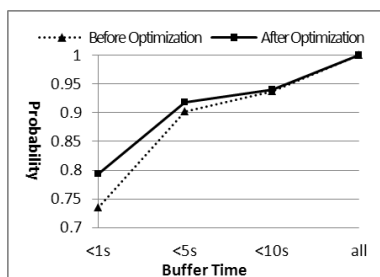


Figure 6. The cumulative distribution function of buffer time before and after optimization

shouter than 1s increased from 1471 to 1586, namely 7.82%. And the number of sets whose buffer time was longer than 10s reduced from 126 to 119, namely 5.56%. In general, the optimization strategies made more peers buffer in a shorter time and reduced the number of peers whose buffer time were long.

3) Data Distribution Rate

The data distribution rate S is calculated as follows,

$$S = D_u / D_d * 100\% \tag{4}$$

In formula (4), D_u represents the total amount of uploading data; D_d represents the total amount of downloading data.

The distribution rate is shown in table IV and the comparison of before and after optimization is shown in Fig. 7 and Fig. 8.

The vertical axis of Fig. 7 shows the percentage of number of sets whose data distribution rate is in the interval percentage of the horizontal axis.

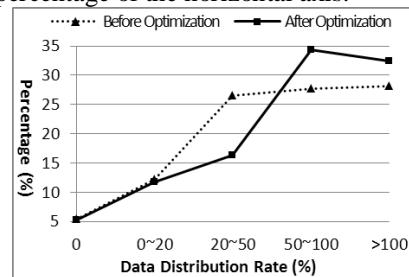


Figure 7. The line chart of distribution rate before and after optimization

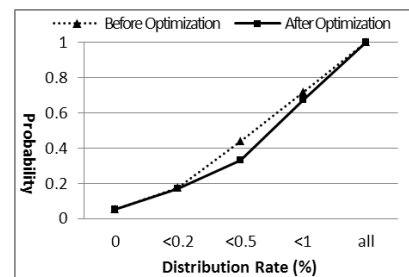


Figure 8. The cumulative distribution function of Distribution rate before and after optimization

Fig. 8 shows the cumulative distribution function of distribution rate before and after optimization. The number of sets whose data distribution rate was more 100% was increased from 563 to 648, namely 15.10% and the number of sets whose data distribution rate was more 50% was increased from 1118 to 1334, namely 19.32%. As we can see in table IV, Fig. 7 and Fig. 8, the

TABLE V. OPTIMIZATION CONTROL MESSAGE OVERHEAD BEFORE AND AFTER OPTIMIZATION

Control Message Overhead (%)	Number Of Sets		Percentage (%)		Growth Rate (%)
	Before Optimization	After Optimization	Before Optimization	After Optimization	
<0.5	581	615	29.05	30.75	5.85
≤1.0	1495	1540	74.75	77	3.01
≤2.0	1928	1943	96.4	97.15	0.78
≤3.0	1975	1976	98.75	98.8	0.05
>3.0	23	24	1.15	1.2	4.35

TABLE VI. THE AVERAGE PERFORMANCE BEFORE AND AFTER OPTIMIZATION

Average Value	Start-up latency	Buffer Time	Data distribution rate	Control message overhead
Before Optimization	2.98s	2.07s	67.92%	0.90%
After Optimization	2.87s	1.83s	74.72%	0.87%
Optimization Rate	3.69%	11.59%	10.01%	3.33%

distribution rate line of the system before optimization was more flat due to the use of randomly peer selection algorithm. After optimization, the system can made full use of the peers' upload ability to distribute data blocks leading to a higher data distribution rate.

4) Control Message Overhead

The control message overhead C is calculated as follows,

$$C = N_control / (N_control + N_data) * 100\% \quad (5)$$

In formula (5), $N_control$ represents the amount of control message; N_data is the amount of video data, they are both of the same program.

The distribution rate is shown in table V and the comparison of before and after optimization is shown in Fig. 9.

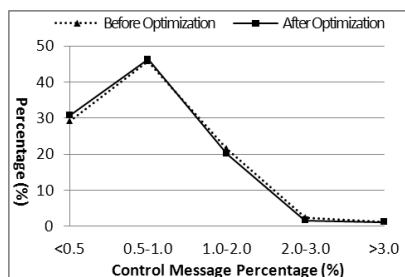


Figure 9. The line chart of control message overhead before and after optimization

The vertical axis of Fig. 9 shows the percentage of the number of sets whose data control message overhead is in the interval percentage of the horizontal axis. After optimization, the system increasing the number of sets whose control message overhead was less than 0.5% from 581 to 615, namely 5.85%. But the number of sets whose control message overhead was more than 2% increased too. The optimization was not very good.

5) Average Performance

Finally, this paper analyzes the average performance of start-up latency, buffer time, data distribution rate, and control message overhead before and after optimization. The result is shown as table VI.

As can be seen in table VI, the optimization is very good, especially in increasing the playback continuity and data distribution rate. After optimization, the high service ability peers were fully utilized due to the adaptive peer

selection algorithm and the data is distributed faster. And the control message is decreased due to the push-pull combination method.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper first introduces the P2P live streaming system named NTLive. Then, proposes two optimization strategies to improve the peer selection algorithm and the data distribution mechanism for the system. One is an adaptive peer selection algorithm based on the peer real-time service ability. In this algorithm, we propose some simple but effective methods to calculate the peers' service ability. And source peer automatically adjust the number of the destination data blocks and the priority of destination peer according to the destination peer real-time service ability. Thus, the system will dynamically adapt to the change of network and fully take use of peers' resources. The other one is a push-pull combination data distribution mechanism. This data distribution mechanism adaptively chooses push or pull model to distribute data according to the requirements of source peer and the service abilities of destination peer. Finally, apply the optimization strategies to the NTLive system and compare the performance before and after optimization. The results show that after optimization more peers start up in a shorter time and the buffer time is shorter. And the average optimization rate of start-up latency and buffer time were 3.69% and 11.59%. What's more, the optimized system makes full use of the peers' upload ability to distribute data blocks and increases much more peers' data distribution rate higher. The average optimization rate of data distribution rate was 10.01%. And the push-pull combination mode also can decrease the control message overload and the average optimization rate of control message overload is 3.33%.

B. Future Work

The proposed optimization strategies have a good effect on the most of peers, but do not good on peers with low service ability. So, we must do more work to solve this problem. In addition, this paper doesn't measure the playback lag since we set a fixed value 30s as the playback lag in the experiment. We are preparing to research the problem how to decrease playback lag.

What's more, we will do experiment to compare the optimized system with other P2P live streaming systems, such as PPStream, PPLive and UUSee.

ACKNOWLEDGMENT

This paper is supported by the National Key Technology R&D Program of China (Grant No. 2012BAH18B05).

REFERENCES

- [1] China Internet Network Information Center. China Internet Development Statistics Report, (2012). http://www.cnnic.cn/research/bgxz/tjbg/201201/t20120116_23668.html.
 - [2] Jiewen Luo. Peer to Peer (P2P) Summary, (modified version), 2005. Beijing: Institute of computing technology, the Chinese academy of sciences.
 - [3] Zhijie Shen. LAN-Awareness: Improved P2P Live Streaming. In *Proceedings of the 21st International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2011.
 - [4] Sijia Huang, Zhihui Lv, Jie Wu. Novel Double Hybrid Architecture of Live Video Streaming System. *Computer Engineering*, Vol. 37, No. 9, pp. 284-287, 2011.
 - [5] Chao Hu, Ming Chen, Changyou Xing, Bo Xu. EUE principle of resource scheduling for live streaming systems underlying CDN-P2P hybrid architecture. *Peer-to-peer Networking and Applications*, Vol. 5, No. 4, pp. 312-322, 2010.
 - [6] Zhixin Sun, Yadang Chen, Zhiguang Ren. P2P live streaming system-based data transmission strategy. *Journal of Communication*, Vol. 32, No. 6, pp. 1-9, 2011.
 - [7] Jing Yang, Runzhi Li, Zongmin Wang. Interval-based cache algorithm in P2P streaming live system *Computer Engineering and Design*, Vol. 31, No. 1, pp. 90-93, 2010.
 - [8] Yifeng Lu, Jinlin Wang, Hang Su. P2P live streaming systems combining IP Multicast. *Micro Computer Information*, Vol. 26, No. 5-3, pp. 14-16, 2010.
 - [9] Junfeng Yuan. The research on churn problem of P2P streaming live system. *Zhengzhou, China: Information Engineering College of Zhengzhou University*, 2010.
 - [10] Kumar R, Liu Y, Ross K. Stochastic fluid Theory for P2P streaming systems. In: *INFOCOM 2007, Anchorage, Alaska*, 2007.
 - [11] XieHai-yong, Richard Yang Y, et al. P4P: provider portal for applications. *SIGCOMM Compute Common, Rev*, Vol. 39, No. 4, pp. 351-362, 2008.
 - [12] Chan K. -H. Kelvin, Chan S. -H. Gary, Begem Ali C. SPANC: Optimizing Scheduling Delay for Peer-to-Peer Live Streaming. *IEEE Transactions on Multimedia*, Vol. 12, No. 7, pp. 743-753, 2010.
 - [13] Eugenio Alessandria, Massimo Gallo, Emilio Leopardi, Marco Melilla, Michaela Moe. Impact of adverse network conditions on P2P-TV systems: *Experimental evidence*. *Computer Networks*, Vol. 55, No. 9, pp. 2035-2050, 2011.
 - [14] RVVSV Prasad, Vegi Srinivas, V. Valli Kumari, KVSVN Raju. An Effective Calculation of Reputation in P2P Networks. *Journal of Networks*, VOL. 4, NO. 5, pp. 332-342, July 2009
 - [15] Takayama, K., Fujimoto, T., Endo, R., Shigeno, H. Neighbor selection based on transmission bandwidth on P2P live streaming service. *IEEE Workshops of International Conference on Advanced Information Networking and Applications*. 2012; 105-110.
 - [16] Junaid Afzal, Thomas Stockhammer, Taigo Gasiba, Wen Xu. Video Streaming over MBMS: A System Design Approach. *Journal of Multimedia*, VOL. 1, NO. 5, August 2006.
 - [17] Liangjin Lu, Jian Wan, Xianghua Xu. Research and Implementation of the push-pull combination P2P live system. *Computer Engineering*, Vol. 34, No. 8, pp. 240-242, 2008.
 - [18] Zi-ao Zhan, Qi-xing Xu. A Stochastic Push Scheme of Streaming Media Partial Content Based on P2P. *Journal of Networks*, Vol. 7, No. 10, pp. 1561-1567, October 2012
 - [19] Hui Wang, Weitao Chen, Yajie Liu. customR2: a hybrid push/pull scheduling method with network coding in P2P live streaming systems. *Journal of Computer Applications*, Vol. 30, No. 2, pp. 285-302, 2010.
 - [20] Laizhong Cui, Yong Jiang, Jianping Wu, Shutao Xia. An Optimal Pull-push Scheduling Algorithm Based on Network Coding for Mesh Peer-to-peer Live Streaming. *IEICE Transactions on Communications*, Vol. E95-B, No. 6, pp. 2022-2033, 2012.
 - [21] Yonghua Xiong, Min Wu, Weijia Jia. Delay Prediction for Real-Time Video Adaptive Transmission over TCP. *Journal of multimedia*, Vol. 5, No. 3, June 2010.
- Xiaosong Wu** was born in 1989. He received his bachelor's degree from College of Computer Science of Sichuan University, Chengdu, China, in 2012. He is currently working toward the master's degree at the same University. His research interests include peer-to-peer IPTV systems, information security, and network measurement.
- Xingshu Chen** was born in 1968. She received her M.S. degree from College of Computer Science of Sichuan University, Chengdu, China, in 1999 and PhD from Mathematical College of Sichuan University, Chengdu, China, in 2004. She is a professor and PhD supervisor of College of Computer Science since 2008. She is currently the director of Network and Trusted Computing Institute (NTCI) and vice director of Information Management Center. She is the deputy secretary general of Sichuan Province Computer Federation since 2002. And she is the members of national information security standards committee since 2012. She awarded Scientific and Technological Progress Second-class Award of Sichuan Province of China in September 2008. Her general research interests include peer-to-peer networks, information security, computer networks and cloud computing.
- Haizhou Wang** was born in 1986. He received his B.E. degrees from College of Computer Science, Sichuan University, China, in 2008. He is currently working toward the Ph.D. degree at the same University. He awarded Outstanding Graduate Student of Sichuan University three times in 2009 to 2010, and 3rd Prize in 2010 NVIDIA CUDA Collegiate Programming Contest of China. His research interests include peer-to-peer IPTV systems, information security, and network measurement.

Instructions for Authors

Manuscript Submission

We invite original, previously unpublished, research papers, review, survey and tutorial papers, application papers, plus case studies, short research notes and letters, on both applied and theoretical aspects. Manuscripts should be written in English. All the papers except survey should ideally not exceed 12,000 words (14 pages) in length. Whenever applicable, submissions must include the following elements: title, authors, affiliations, contacts, abstract, index terms, introduction, main text, conclusions, appendixes, acknowledgement, references, and biographies.

Papers should be formatted into A4-size (8.27" x 11.69") pages, with main text of 10-point Times New Roman, in single-spaced two-column format. Figures and tables must be sized as they are to appear in print. Figures should be placed exactly where they are to appear within the text. There is no strict requirement on the format of the manuscripts. However, authors are strongly recommended to follow the format of the final version.

All paper submissions will be handled electronically in EDAS via the JNW Submission Page (URL: <http://edas.info/N10935>). After login EDAS, you will first register the paper. Afterwards, you will be able to add authors and submit the manuscript (file). If you do not have an EDAS account, you can obtain one. If for some technical reason submission through EDAS is not possible, the author can contact jnw.editorial@gmail.com for support.

Authors may suggest 2-4 reviewers when submitting their works, by providing us with the reviewers' title, full name and contact information. The editor will decide whether the recommendations will be used or not.

Conference Version

Submissions previously published in conference proceedings are eligible for consideration provided that the author informs the Editors at the time of submission and that the submission has undergone substantial revision. In the new submission, authors are required to cite the previous publication and very clearly indicate how the new submission offers substantively novel or different contributions beyond those of the previously published work. The appropriate way to indicate that your paper has been revised substantially is for the new paper to have a new title. Author should supply a copy of the previous version to the Editor, and provide a brief description of the differences between the submitted manuscript and the previous version.

If the authors provide a previously published conference submission, Editors will check the submission to determine whether there has been sufficient new material added to warrant publication in the Journal. The Academy Publisher's guidelines are that the submission should contain a significant amount of new material, that is, material that has not been published elsewhere. New results are not required; however, the submission should contain expansions of key ideas, examples, elaborations, and so on, of the conference submission. The paper submitting to the journal should differ from the previously published material by at least 30 percent.

Review Process

Submissions are accepted for review with the understanding that the same work has been neither submitted to, nor published in, another publication. Concurrent submission to other publications will result in immediate rejection of the submission.

All manuscripts will be subject to a well established, fair, unbiased peer review and refereeing procedure, and are considered on the basis of their significance, novelty and usefulness to the Journals readership. The reviewing structure will always ensure the anonymity of the referees. The review output will be one of the following decisions: Accept, Accept with minor changes, Accept with major changes, or Reject.

The review process may take approximately three months to be completed. Should authors be requested by the editor to revise the text, the revised version should be submitted within three months for a major revision or one month for a minor revision. Authors who need more time are kindly requested to contact the Editor. The Editor reserves the right to reject a paper if it does not meet the aims and scope of the journal, it is not technically sound, it is not revised satisfactorily, or if it is inadequate in presentation.

Revised and Final Version Submission

Revised version should follow the same requirements as for the final version to format the paper, plus a short summary about the modifications authors have made and author's response to reviewer's comments.

Authors are requested to use the Academy Publisher Journal Style for preparing the final camera-ready version. A template in PDF and an MS word template can be downloaded from the web site. Authors are requested to strictly follow the guidelines specified in the templates. Only PDF format is acceptable. The PDF document should be sent as an open file, i.e. without any data protection. Authors should submit their paper electronically through email to the Journal's submission address. Please always refer to the paper ID in the submissions and any further enquiries.

Please do not use the Adobe Acrobat PDFWriter to generate the PDF file. Use the Adobe Acrobat Distiller instead, which is contained in the same package as the Acrobat PDFWriter. Make sure that you have used Type 1 or True Type Fonts (check with the Acrobat Reader or Acrobat Writer by clicking on File>Document Properties>Fonts to see the list of fonts and their type used in the PDF document).

Copyright

Submission of your paper to this journal implies that the paper is not under submission for publication elsewhere. Material which has been previously copyrighted, published, or accepted for publication will not be considered for publication in this journal. Submission of a manuscript is interpreted as a statement of certification that no part of the manuscript is copyrighted by any other publisher nor is under review by any other formal publication.

Submitted papers are assumed to contain no proprietary material unprotected by patent or patent application; responsibility for technical content and for protection of proprietary material rests solely with the author(s) and their organizations and is not the responsibility of the Academy Publisher or its editorial staff. The main author is responsible for ensuring that the article has been seen and approved by all the other authors. It is the responsibility of the author to obtain all necessary copyright release permissions for the use of any copyrighted materials in the manuscript prior to the submission. More information about permission request can be found at the web site.

Authors are asked to sign a warranty and copyright agreement upon acceptance of their manuscript, before the manuscript can be published. The Copyright Transfer Agreement can be downloaded from the web site.

Publication Charges and Re-print

The author's company or institution will be requested to pay a flat publication fee of EUR 360 for an accepted manuscript regardless of the length of the paper. The page charges are mandatory. Authors are entitled to a 30% discount on the journal, which is EUR 100 per copy. Reprints of the paper can be ordered with a price of EUR 100 per 20 copies. An allowance of 50% discount may be granted for individuals without a host institution and from less developed countries, upon application. Such application however will be handled case by case.

More information is available on the web site at <http://www.academypublisher.com/jnw/authorguide.html>.

Reliable Transmission Protocol based on Network Coding in Delay Tolerant Mobile Sensor Network <i>Luo Kan, Wang Hua, and Shyi-Ching Liang</i>	1027
Routing Optimization Based on Taboo Search Algorithm for Logistic Distribution <i>Yang Hongxue and Xuan Lingling</i>	1033
Opportunistic Cooperative Reliable Transmission Protocol for Wireless Sensor Networks <i>Hua Guo, Yu Sheng-Wen, and Douglas Leith</i>	1040
An Improved Channel Estimation Method based on Jointly Preprocessing of Time-frequency Domain in TD-LTE System <i>Yang Jianning, Lin Kun, and Zhao Xie</i>	1047
Dynamic Routing Algorithm Based on the Channel Quality Control for Farmland Sensor Networks <i>Dongfeng Xu</i>	1055
DCSK Multi-Access Scheme for UHF RFID System <i>Keqiang Yue, Lingling Sun, Bin You, and Shengzhou Zhang</i>	1061
An Effective Scheme for Performance Improvement of P2P Live Streaming Systems <i>Xiaosong Wu, Xingshu Chen, and Haizhou Wang</i>	1067

A Multicast Routing Algorithm for Datagram Service in Delta LEO Satellite Constellation Networks <i>Yanpeng Ma, Xiaofeng Wang, Jinshu Su, Chunqing Wu, Wanrong Yu, and Baokang Zhao</i>	896
Bandwidth Consumption Efficiency Using Collective Rejoin in Hierarchical Peer-To-Peer <i>Sri Wahjuni, A.A.Putri Ratna, and Kalamullah Ramli</i>	908
A Method of Case Retrieval for Web-based Remote Customization Platform <i>Yuhuai Wang, Hong Jia, and Xiaojing Zhu</i>	914
Symbol Timing Estimation with Multi-h CPM Signals <i>Sheng Zhong, Chun Yang, and Jian Zhang</i>	921
Vector-Based Sensitive Information Protecting Scheme in Automatic Trust Negotiation <i>Jianyun Lei and Yanhong Li</i>	927
An Improved Byzantine Fault-tolerant Program for WSNs <i>Yi Tian</i>	932
EESA Algorithm in Wireless Sensor Networks <i>Zhang Pei and Feng Lu</i>	941
Routing Algorithm Based on Delay Rate in Wireless Cognitive Radio Network <i>Gan Yan, Yuxiang Lv, Qiyin Wang, and Yishuang Geng</i>	948
Energy Hole Solution Algorithm in Wireless Sensor Network <i>Lu Yuting and Wang Weiyang</i>	956
Identification Method of Attack Path Based on Immune Intrusion Detection <i>Huang Wenhua and Yishuang Geng</i>	964
Online Order Priority Evaluation Based on Hybrid Harmony Search Algorithm of Optimized Support Vector Machines <i>Zhao Yuanyuan and Chen Qian</i>	972
Framework and Modeling Method for Heterogeneous Systems Information Integration Base on Semantic Gateway <i>Xianwang Li, Yuchuan Song, Ping Yan, and Xuehai Chen</i>	979
Satellite Formation based on SDDF Method <i>Wang Yu, Wu Zhi-qiang, and Zhu Xin-hua</i>	986
Heterogeneous Web Data Extraction Algorithm Based On Modified Hidden Conditional Random Fields <i>Cheng Cui</i>	993
Nearly Optimal Solution for Restricted Euclidean Bottleneck Steiner Tree Problem <i>Zimao Li and Wenying Xiao</i>	1000
Computer Crime Forensics Based on Improved Decision Tree Algorithm <i>Ying Wang, Xinguang Peng, and Jing Bian</i>	1005
Demand-oriented Traffic Measuring Method for Network Security Situation Assessment <i>Xu Zhenhua</i>	1012
Visual Simulation of Explosion Effects Based on Mathematical Model and Particle System <i>Gong Lin and Hu Dingjun</i>	1020
