Journal of Networks

ISSN 1796-2056

Volume 6, Number 11, November 2011

Contents

Mobility impact on session survivability under the SHIM6 protocol and enhancement of its rehoming procedure

Amine Dhraief Abdelfettah Belghith
HANA Research Group, Manouba University, Tunisia
Email: amine.dhraief@isigk.rnu.tn, abdelfattah.belghith@ensi.rnu.tn

***Abstract*—Multihoming is a solution that enables a fault-tolerant access to the Internet by configuring on each network entity several IP addresses associated with distinct ISPs. IPv6 natively allows end-hosts and end-sites to be multihomed where nodes and routers can have multiple IP addresses. However, a specific support is required to take full advantage of multihoming. The SHIM6 protocol provides such a support.**

We study in this paper to what extent the mobility impacts the SHIM6 protocol component in general and more specifically the context establishment as it is a *sine qua none* condition for session survivability. We focus on possible consequences of mobility before, during, and after the context establishment. We find that in some mobility scenarios, the SHIM6 context is never established and the session survivability cannot be ensured.

***Index Terms*—Multihoming, Mobility, SHIM6, Testbed**

I. Introduction

Providing a redundant and reliable access to the network is a major concern for protocol designers [1]. A solution that enables a fault-tolerant access to the Internet consists in configuring on each network entity several IP addresses associated with distinct Internet Service Providers (ISP). A study conducted by Agrawal et al. [2] revealed that at least 60% of Internet stub autonomous systems (AS) are multihomed to two or more ISP.

Multihoming protocols provide an indispensable support to take full advantage of multihoming and a framework for multiple addresses management [3]. With the forthcoming version of the IP network, IPv6, lots of efforts have been made to enable multihoming benefits, such as reliability, session survivability and load sharing. During the last few years, more than 40 solutions have been proposed for IPv6 multihoming [4]. The majority of these solutions have their own mechanisms to preserve established sessions after a failure.

On the other hand mobility protocols, a family of protocols which provides a support for host mobility, also aim at preserving nodes sessions while moving. Mobility and multihoming are usually studied separately. Mobility protocols do not consider the case of multihomed mobile node, while multihomed protocols do not take into account mobility. Nonetheless, in the Internet, nodes are at the same time mobile and multihomed. Nodes are mobile as they are able to change the access network while having running session; they are multihomed as they are equipped with several interfaces (such as Wi-Fi or Wimax)

In this paper, we study the impact of the mobility on multihoming protocols. For this purpose, we focus on a particular multihoming protocol - the SHIM6 protocol. One of the most important aspects of the SHIM6 protocol is its context establishment, as it is a *sine qua none* condition for session survivability. Without an established SHIM6 context, communicating peers cannot rehome their communications in case of failures. Hence, we focus more precisely on possible consequences of mobility before, during and after the context establishment. We find that in some mobility scenarios, the SHIM6 context is never established and the session survivability cannot be ensured. Furthermore, the rehoming procedure is a key feature of any multihoming protocol. Rehoming a communication implies a change in the used IP address and may result in the change of the upstream ISP. As SHIM6 is designed to be deployed on static nodes, its default rehoming decision strategy does not meet the requirement of mobility. We present in this paper some optimizations in order to improve the SHIM6 rehoming latency in a mobile environment. We demonstrate, through measurements on a real testbed, that these optimizations improves significantly the rehoming latency.

The remainder of this paper is structured as follows. Section II highlights multihoming motivations, functionalities and constraints. Section III starts with an overview of the IPv6 multihoming approaches and then focuses on the SHIM6 protocol. Section IV analyzes the impact of node mobility on the SHIM6 context and evaluates it in an experimental testbed. Section V concludes this paper.

II. Background

A. Definitions

A node identifier refers to a constant that uniquely identifies a node in a given network [5]. A locator is the topological reference of an interface in a network. The later is used by routing protocols to locate any entity in a network.

Mobility is defined as a change in the node locator. In a mobile environment, a node that changes its current locator usually preforms successively two steps: a layer 2 (L2) handover then a layer 3 (L3) handover. The

doi:10.4304/jnw.6.11.1521-1532

L2 handover consists in a change from the link-layer connectivity; whereas, the L3 handover refers to a change from the access network and the acquisition of a new locator [6].

A node is said multihomed when it simultaneously has several available locators. A node may acquire several locators in various configurations: it can be connected to a link where multiple prefixes are advertised or it can be equipped with several interfaces attached to different access networks.

B. Multihoming motivations

Multihoming presents several motivations such as sustained redundancy against failure, improved performance, allowing load sharing and permitting policing [7].

1) Redundancy: The most important motivation of getting attached to several upstream providers is to protect end-sites as well as end-hosts from failures. Abley et al. [8] detailed the potential causes of failures in networks. The most common and important ones are : physical failures, routing protocol failures and ISP failures. Physical failures refer to outages that may affect network components (e.g., routers) or network connections (e.g., fiber cuts). Routing protocol failures are due to misbehaving routing protocols (e.g., withdraw valid routes or announce unstable routes). ISP failures are outages that affect the Internet providers leading to the interruption of the Internet connectivity. By providing redundant paths, multihoming alleviates these failures. If an outage affects one of the available paths, multihoming protocols detect this failure and rehome running sessions onto another working path.

2) Performance: Akella et al. [9] quantified to what extent multihoming improves network performance in terms of delay, available bandwidth and reliability. They showed that a multihomed site connected to two ISPs acquires a 25% improvement in its average performance, and that getting connected to more than 4 providers yields a little further improvement. In [10], Launois et al. showed that multihoming increases the number of concurrent available paths and that lower delays are found among new paths. Hence, multihoming may improve network performance in terms of delay, available bandwidth and resilience against failure.

3) Load sharing: Load balancing refers to the situation where a site spreads its traffic among its available links towards a given destination; whereas, load sharing refers to the case where no destination is specified. Load balancing is then a particular case of load sharing. Load sharing allows end-sites to increase their aggregate throughput and thus improve their performance. For example, as the broadband access price is constantly dropping, small and mid-sized corporate can emulate a T1 link by subscribing to several broadband accesses and distribute their traffic among the different connections. Hence, multihomed sites can use their available connections simultaneously by distributing both incoming and outgoing traffics among their available paths and thus performing load sharing [11].

4) Policy: By being multihomed, a site would like to distribute its traffic according to some policies. Policies are the rules that define the traffic to be forwarded to a given provider. For example, a corporate may subscribe to two providers, one for its e-commerce transactions and the other for its personal Internet usage.

C. Multihoming functionalities

In order to satisfy the aforementioned incentives and motivations, some fundamental functionalities should be provided by multihoming protocols.

1) Decoupling node identification from its localization: TCP/IP has been formally designed to allocate a single IP address per device. The role of an IP address was twofold: it locates an end-host in the network and it identifies the end-host running sessions [12], [13]. In the current Internet, nodes tend to be mobile and multihomed, most of the time they are equipped with multiple interfaces. They have several addresses so they require a more flexible interaction with their address sets. From a session point of view, their identity needs to be independent from their physical administrative domain [14]. Multihomed nodes have several IP addresses and consequently are located in several networks; whereas, they should have a single identity. Therefore, the multihoming paradigm require the decoupling of node identity from its location.

2) Maintaining the set of addresses up-to-date: The major motivation of multihoming is to have redundant accesses to the Internet in order survive failures. Failures that might affect Internet paths and ISP renumbering operations are events that alter the multihomed node address sets. Multihoming solutions must provide mechanisms for failure detection and failure recovery.

Abley et al. [8] detailed the potential causes of failures in networks. The most common and important ones are : physical failures, routing protocol failures and ISP failures. Physical failures refer to outages that may affect network components (e.g., routers) or network connections (e.g., fiber cuts). Routing protocol failures are failures due to misbehaving routing protocols (e.g., withdrawing valid routes or announcing unstable routes) [15]. ISP failures are outages that affect Internet providers leading to the interruption of the Internet connectivity. Similarly to the movement for mobile nodes, failure detection can be used for multihomed nodes as a clue to verify the reachability of the currently used locator. Hence, multihomed nodes should detect such events in order to maintain their address set up-to-date.

3) Traffic engineering: On one hand, multihoming entities (end-sites or end-hosts) are connected to the Internet through several paths characterized by a set of quality of service (QoS) parameters (e.g., delay, bandwidth, jitter). On the other hand, multihoming entities aim at performing load sharing according to some policies. Hence, multihoming entities need to efficiently select the appropriate path satisfying their performance or policy requirement [16]. Traffic Engineering (TE) functionalities allow multihomed entities to optimize the use of available

paths by adapting the route selection mechanism to some requirements [17]. TE functionalities are necessary to achieve suitable performances, load sharing and the policy requirements presented in section II-B.

D. Multihoming constraints

The deployment of multihoming in the Internet faces, however, several constraints. These constraints are thoroughly discussed in [8].

1) Scalability: The scalability issue is a major concern in deploying multihoming. In fact, a multihoming solution that maintains states in the inter provider routing systems is inherently not scalable. For example, one of the most used techniques to achieve multihoming relies on the BGP. BGP is an inter-domain routing protocol, i.e. BGP is responsible of routes announcements in the Internet. The Internet contains today more than 25000 IPv4 autonomous Systems (AS) and the BGP Routing Information Base (RIB) contains approximatively 300000 entries [18]. In addition, at least 60% of stub domains are multihomed [2]. The growth of the BGP table impacts the packet forwarding speed and requires a large memory space.

2) Compatibility with IP routing: A multihoming solution should be compatible with the IP routing [19]. The locators used by such protocol should comply with the Internet topology. The Rekhter Law [20] stipulates that: "Addressing can follow topology or topology can follow addressing; choose one.". Multihoming protocols should follow this law in order to prevent routes disaggregation and allow the routing system to scale. Thus, as end host routes can not be propagated in the whole Internet, the used locator should be topologically correct [21].

3) Compatibility with existing sockets: The POSIX standard [22] defines a generic socket API which must be used by any protocol for interoperability purpose. This socket API uses IP addresses for identification purposes Consequently, a protocol that supports multihoming should be compatible with the existing socket API.

4) Independence: The Independence refers to the absence of cooperation between a multihomed entity and its upstream providers, and among the upstream providers themselves. In order to ensure independence between the multihomied entity and its upstream providers, a multihoming solution should not be dependent on specific configurations enabled on the provider side. This means that an ISP should not provide a specific support to an end-site because this end-site is multihomed so that small corporate can also benefit from multihoming. The independence between providers means that ISPs are not supposed to cooperate between each other because an end site is multihomed with them.

III. The SHIM6 Protocol

We present in this section an overview of the IPv6 multihoming approaches and then we focuss on a host-centric multihoming protocol, namely the SHIM6 protocol.

We distinguish three categories of multihoming protocols: the routing approach, the edge-centric approach and the host-centric approach. The routing approach is based on path diversity inferring and re-routing algorithms in order to provide multihoming functionalities. The edge-centric approach enables the multihoming support at the edge of an IPv6 network. Host-centric approach enables the multihoming support at the end hosts.

A. Multihoming host-centric approaches

In such approaches, end hosts are more aware of their networks and available connections and this requires extra complexity in network stacks. There are mainly two ways for enabling the multihoming in a network stack. The first one consists of modifying an existing layer; whereas the second way consists of adding a new thin sub-layer to handle the multihoming.

Most of the host-centric solutions which are based on the first way of enabling multihoming support modify the transport layer since multihoming aims at providing transport layer survivability. The most used transport protocols (TCP and UDP) use IP addresses to identify communication. If a failure occurs in the used address, the transport session is automatically broken. A possible solution to this problem is to use multiple addresses per end-host in the transport layer to switch from one address to another in case of a failure. Several proposals in the literature enabled multihoming in the transport layer such as Multihomed TCP, TCP-MH, SCTP and DCCP. Multihomed TCP [23] and TCP-MH [24] modify the TCP protocol while SCTP and DCCP are new transport protocols. Multihomed TCP uses a context identifier instead of IP addresses and ports to identify a connection. TCP-MH modifies the SYN segment to contain all the available addresses and implements primitives (MH-Add,MH-Delete) to modify the address currently in use. The SCTP [25] provides a native multihoming support by associating one session with multiple IP addresses and thus with multiple paths. One path is considered as primary and the others are backups. The Datagram Congestion Control Protocol (DCCP) did not originally support multihoming. In fact, multihoming is added as an extension [26] that provides an extension adds primitive to transfer the established connection from one address to another.

Host-centric solutions based on the second way of enabling multihoming at the end-host add a new shim layer in the network stack which decouples a node's identification from its localization. In fact, TCP/IP has been formerly designed to allocate a single IP address per device. The role of the IP address was two-fold: First locate the end-hosts in the network, second identify end-hosts running sessions [12], [13]. Multihomed end-hosts have several addresses and thus they require a more flexible interaction with their address sets. From a session's point of view, identity needs to be independent from their physical administrative domain [14]. Therefore, a solution that manages multiple IP addresses per node and thus

handles multihoming, consists of adding a shim layer which decouples locators from identifiers: the application handles identifiers while IP routing layer handles locators. We can decouple node identity from its localization by creating new namespaces either for node identification or node localization. We can also achieve this by choosing one address as a permanent node identifier and consider the remaining addresses as potential locators. In this case, we should perform an IP address rewriting in order to translate the node identifier into a locator and vice versa.

B. SHIM6

The SHIM6 protocol is an IPv6 multihoming protocol [27], [28]. It introduces a new shim sublayer within the IP layer.

In order to preserve session survivability, SHIM6 uses one of the available addresses as a permanent identifier. This address -called upper layer identifier (ULID)- is a location independent identifier. The remaining addresses are considered as locators. The shim layer performs an address rewriting from ULID to locator and vice versa. Each SHIM6 node stores the information related to its locators, ULID and its correspondent peer addresses in a structure called the SHIM6 context. The SHIM6 context is established after a four-way handshake and can be modified while having an ongoing communication through specific SHIM6 update messages (see Fig. 1).

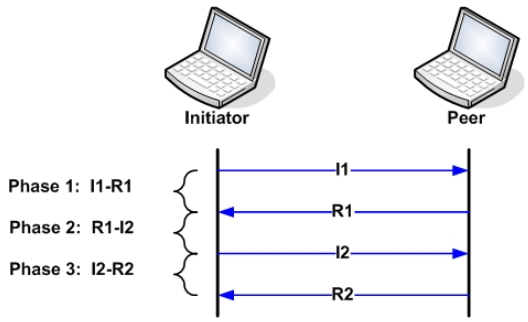


Fig. 1. SHIM6 context establishment

C. Establishing a SHIM6 context

SHIM6 maintains a context per ULID pair which holds information about the established session between these two ULID. A SHIM6 context is identified by two context tags which are 47-bit numbers randomly allocated by each one of the communicating peer. The tag is included in each SHIM6 control message in order to prevent possible attacker to spoof SHIM6 control messages. In order to discover SHIM6 context tags, attackers need to be along the path to sniff the context tag.

The SHIM6 context is established after a four-way handshake control messages I1,R1,I2,R2. We detail in the following each of these messages.

1) I1 Message: The I1 message is the first SHIM6 control message in the context establishment handshake. When one of the communicating peer decides to set up a SHIM6 context, it starts by creating a context state. It allocates a tag to the SHIM6 context, sends an I1 message to the other node and sets the state of its context to I1_SENT. The I1 message embeds the context tag allocated by the initiator and a nonce. The nonce is a 32-bit randomly generated number by the initiator which must be included in the R1 message (the response message to I1) to be used to identify the response. The initiator starts a timer (I1_TIMEOUT) upon sending the I1 message, and if it does not receive an R1 message or an I2 message (in case of a simultaneous context establishment) after the expiration of the timer, it retransmits the I1 message. The retransmission of the I1 message is controlled by an exponential back-off timer. The maximum number of allowed retransmissions is MAX_I1_RETRANSMISSION upon which it is inferred thateither the corespondent peer may have not implemented SHIM6 or a firewall is blocking the I1 message. If the initiator peer receives an ICMP error "Unrecognized Next Header" in response to its I1 message, it is a more reliable indication that the correspondent peer does not implement SHIM6.

2) R1 Message: R1 message is a response message to I1. When a host needs to send an R1 message (the Peer in Fig. 1), it copies the initiator nonce field from the I1 message into the R1 message, generates a responder nonce and a hash of the information contained in the I1 message (context tag, ULID pair, initiator nonce, and a secret S maintained by the peer) called responder validator. Both responder validator and the responder nonce are used by the correspondent peer in order to verify that an I2 message is sent in response to its R1 message. At this stage, the correspondent peer does not allocate any SHIM6 state, it stays in the idle state.

3) I2 Message: When a host receives an R1 message (the Initiator in Fig. 1), it first checks whether it has allocated a context corresponding to the nonce included in the R1 message. If no context is found, the host discards the R1 message, otherwise it sends an I2 message. In this latter case, the host copies the Responder Validator field, and the responder nonce from the R1 message and includes them in the I2 message in addition to its locator list. Finally, the hosts starts an I2_TIMEOUT timer and sends the I2 message changing its state from I1_SENT to I2_SENT. If the host does not receive an R2 message in response to the I2 message before I2_TIMEOUT, it may retransmit the I2 message according to a procedure similar to the one used with I1 messages.

4) R2 Message: Upon receiving an I2 message, the host extracts the Responder Validator value and verifies whether this value correspond to the value that it would have computed. If this verification fails, it discards the I2 message, otherwise the host extracts the locator list from the I2 message. It creates then a context and generates a context tag. It finally sends an R2 message and changes its state from IDLE to ESTABLISHED. In the R2 message,

the host includes its context tag, its responder nonce and its locator list. When the peer receives the R2 message, it verifies whether there is context that matches the nonce included in the R2 message. It extracts then the locator list, the context tag and records these information in its context. Finally it changes its state from I2_SENT to ESTABLISHED.

D. Updating a SHIM6 Context

The set of the available locators of a node supporting SHIM6 may change in time. A locator may become unavailable after an outage or after a renumbering operation performed by the corresponding upstream provider. This node can also acquire new locator(s) when a new router boots on the links it is attached to. As this node shares with its correspondent peers its list of locators (recorded in their SHIM6 contexts), it should inform them about any change that may affect its list of locators. For this purpose, SHIM6 uses a control message, called Update Request (UR), which is used by SHIM6 nodes to inform their correspondent peers about any change that affects their locator set. The UR message should be acknowledged by an Update Acknowledgment UA message. If after ending an UR message the node does not receive any UA message before UPDATE_TIMEOUT time, then it retransmits the UR message. The retransmission of the UR is controlled by a back-off timer and the maximum number of retransmission is MAX_UPDATE_TIMEOUT. After reaching this limit, the node discards its SHIM6 context. The UR message includes a request nonce, the destination context tag and the node new locator(s) list. The UA message includes the destination context tag and the request nonce copied from the UR message.

E. SHIM6 context recovery

When a node receives a payload message containing a SHIM6 extension header or a SHIM6 control message but it has no SHIM6 context already established with the sender, it assumes that it has discarded this context while the sender has not. In such situation, the receiver starts a SHIM6 context recovery procedure. It replies with an R1bis SHIM6 message in order to fast-reestablish the lost SHIM6 context.The R1bis message includes the context tag copied from the packet which has triggered the sending of the R1bis, a responder nonce and a responder validator. The responder validator is a hash of the context tag, the pair of the locator, the responder nonce and a secret maintained by the sender of the R1bis. The responder validator together with the responder nonce are used to identify the I2bis message received as a response to the R1bis message. If a node receives an R1bis message, it first extracts the context tag and the source and destination addresses of the message. In order to conclude that the sender of the R1bis message lost its SHIM6 context, the node must verify two conditions. The first condition is that the context tag included in the R1bis message is bound to local SHIM6 Context in the ESTABLISHED state. The second condition is that the source and destination addresses of the R1bis message match respectively the local preferred locator and the peer preferred locator of this context. If the two conditions are fulfilled, the receiver of the R1bis message, replies with an I2bis message. It includes in this message the responder validator and the responder nonce copied from the R1bis message and an initiator nonce in addition to its locator list. Finally it changes the state of its SHIM6 context from ESTABLISHED to I2BIS_SENT. Upon receiving an I2bis message, the host verifies that the responder validator is the equal to a responder validator that it would has computed. Then it allocates a SHIM6 context, changes its state to ESTABLISHED and sends an R2 message.

As the multihomed node has several addresses, SHIM6 uses a combination of Hash Based Addresses (HBAs) [29] and Cryptographically Generated Addresses (CGAs) [30] to bind a set of addresses with a multihomed node and to verify whether a claimed address belongs to a node [16], [31]. SHIM6 uses the REAchability Protocol (REAP) in order to detect possible failures and recover from them [32], [33]. REAP allows SHIM6 to detect failures either through the absence of keepalives sent by the corresponding peer or through information provided by the upper layer protocol. The recovery mechanism is based on the exploration of the available addresse set. The goal of this exploration process is to find a working address pair.

IV. SHIM6 IN A MOBILE ENVIRONMENT

In this section, we study the behavior of the SHIM6 protocol in a mobile environment. As the SHIM6 context is a key feature of the SHIM6 protocol, we study the mobility impact before, during and after the context establishment. In the following we assume that a node called *Initiator* initiates a SHIM6 context with a *Peer* in the Internet. We assume that any of them can move at anytime. In order to study the impact of movement on the context establishment, we divide the context establishment handshake into three phases: the first phase lasts from the sending of the first I1 message until the reception of an R1 message. The second phase lasts from the reception of an R1 message until the sending of the first I2. The third phase lasts from the sending of the first I2 message until the reception of an R2 message (see Fig. 1).

A. Mobility before the context establishment

If an Initiator executes an L3 handover before establishing a SHIM6 context with its Peer, its ongoing session will be broken. Before establishing a SHIM6 context, the Peer does not know the possible locators of the Initiator. Therefore, if the currently used locator of the Initiator becomes unreachable, the whole session is broken. Consequently, the establishment of a SHIM6 context is a *sine qua non* condition for session survivability as it holds information necessary for rehoming.

B. Mobility during the context establishment

1) Preliminary study: In the following, we aim to show that node mobility impacts the context establishment and leads to the loss of SHIM6 control messages and their retransmission. For this purpose, we set up the testbed presented by Fig. 2.

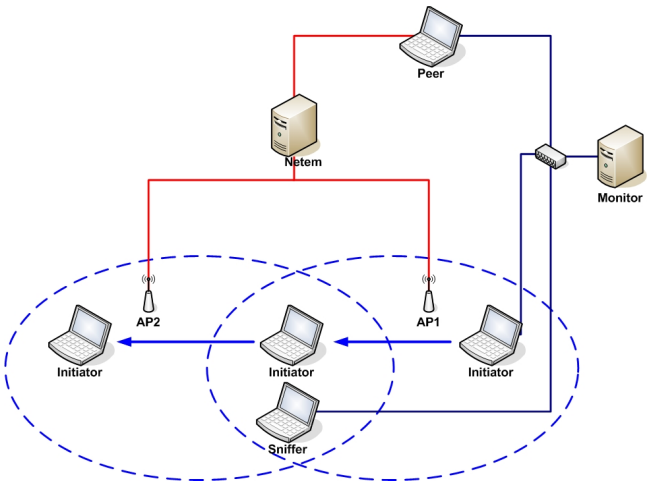


Fig. 2. Testbed 1: impact of nodes mobility on the SHIM6 context establishment latency

Our testbed involves an Initiator that moves between two access points (APs): AP1 and AP2. The APs are connected to the NetEM Node which uses a special feature of the Linux kernel named the network emulator (NetEM) module. This module is used to emulate a large network where we can vary the end-to-end delay and configure the packet loss rate. We configure a one way delay equal to 50 ms +/- 5 ms. The NetEM node is also connected to the SHIM6 nodes. Two other nodes are used to monitor and capture traffic in the experiment. In this experiment, we assume that the Initiator executes an L2 handover. The L2 handover is randomly triggered while the SHIM6 context is being established. In our testbed, the Initiator implements the SHIM6 protocol developed by the UCL University [34].

Table I shows the retransmission frequency of the SHIM6 control messages I1 and I2. We note that in all the cases, either I1 or I2 is retransmitted. A retransmission of a SHIM6 control message indicates that this message was lost during node movement. We can empirically see that node movement, while a SHIM6 context is being established, leads to the loss of a SHIM6 control message and a retransmission either of the I1 or the I2 message.

2) Theoretical study: Node mobility while a SHIM6 context is being established may defer the set up of the context or makes it impossible. The consequences of mobility during the establishment of the SHIM6 context depend on which entity is moving (Initiator, Peer), the handover type (L2, L3) and the phase in which the movement occurs (I1-R1, R1-I2, I2-R2). We use the following notation to capture these parameters: *(moving entity, handover type, context establishment phase)*. In the following, the symbol (*) denotes any possible eventuality. For example, an Initiator executing an L3 handover after sending an I2 message is noted (Initiator, L3, I2-R2). Fig. 3 describes all the possible cases of the mobility during the context establishment. In the following, we assume that all the retransmission timeouts last 4s as suggested by the SHIM6 specifications [28].

	I1	I2
Retransmission frequency	51.2%	48.8%

TABLE I
RETRANSMISSION FREQUENCY OF I1 AND I2 MESSAGES

We first study the consequences of an L2 handover during the context establishment then we focus on the L3 handover case.

- (*, L2, I1-R1): if the Initiator or the Peer changes AP during the I1-R1 phase, the SHIM6 protocol retransmits the I1 message after a timeout as it has not received an R1 message in response to its I1 (case 1 and 7).
- (*, L2, I2-R2): if the Initiator or the Peer changes AP during the I2-R2 phase, the SHIM6 protocol retransmits the I2 message after a timeout as it has not received an R2 message in response to its I2 (case 3 and 9).
- (Peer, L2, R1-I2): if the Peer moves during the R1-I2 phase, it may not receive the I2 sent by the Initiator. Thus, the Initiator retransmits the I2 message after a timeout (case 8).
- (Initiator, L2, R1-I2): if the Initiator moves in the R1-I2 phase, it will send the I2 message as soon as it finishes the L2 handover and therefore, there will be no retransmissions (case 2).

As a conclusion, if an L2 handover occurs during the context establishment, the context will be delayed, but always established.

In the following, we examine the L3 handover consequences on the context establishment.

- (Initiator, L3, I1-R1): if the Initiator moves during the I1-R1 phase and acquires a new address, it will not receive the R1 message - as it was sent to its old address. The Initiator will conclude that its I1 message was lost and therefore, it will send it again from the new address after a timeout and the context will be established (case 4).
- (Initiator, L3, R1-I2): The Initiator moves after receiving the R1 message and before sending I2, it will send the I2 message with the newly acquired address. Before the Initiator movement, the Peer sends an R1 message to the Initiator with a responder validator field calculated using the ULID address pair present in the I1 message. After executing an L3 handover, the Initiator copies this responder validator from the R1 message in its I2 message. Upon receiving the I2 message, the Peer finds that the responder validator does not match the one that it would have computed with the previous address. Thus, the Peer silently

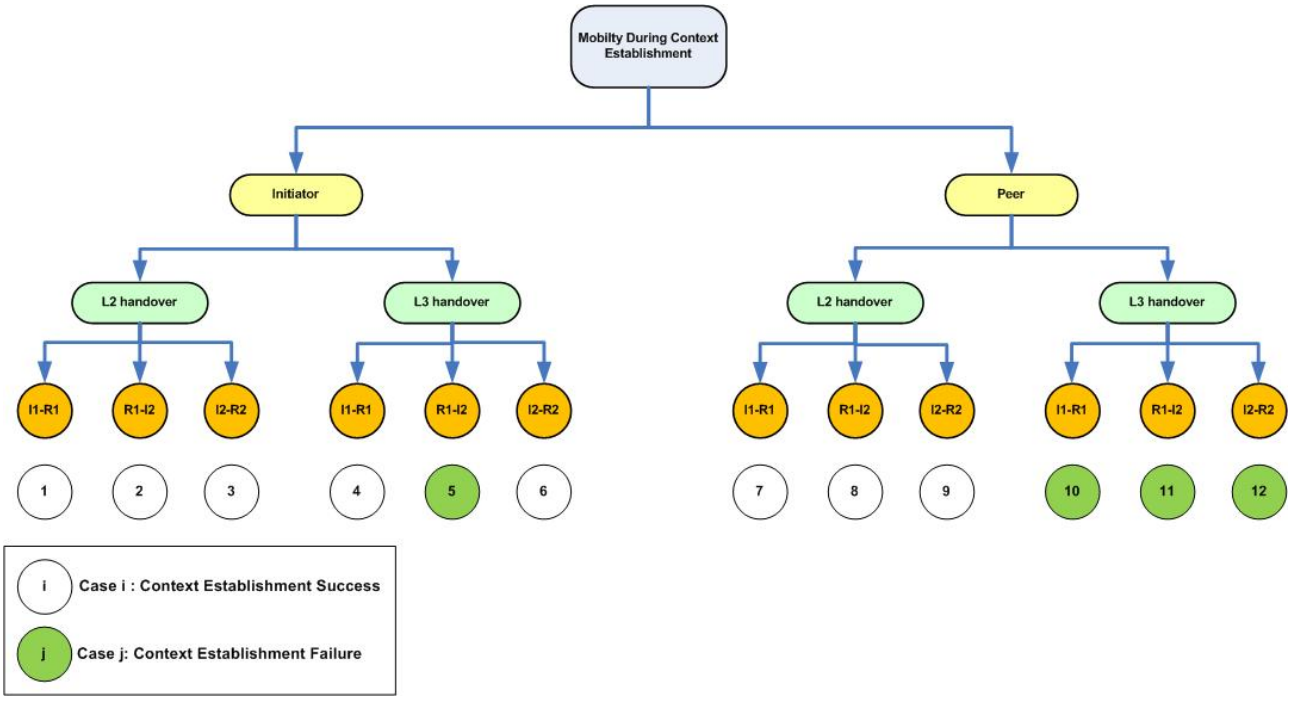


Fig. 3. Possible cases of mobility during the SHIM6 context establishment

discards the I2 even if it retransmitted and the context is not established (case 5).

- (Initiator, L3, I2-R2): if the Initiator moves after sending the I2 message and before receiving R2, it will not receive the R2 message. Before executing an L3 handover, the Initiator sends an I2 message to the Peer containing its old address. The Peer sends the R2 message to the Initiator 's old address and sets its SHIM6 context state to *established*. The Initiator executes an L3 handover and acquires a new address. Meanwhile, it does not receive R2. After a timeout, it sends the I2 message again with the newly acquired address. Upon reception of the I2 message, the Peer verifies that it has already an established context with the Initiator having the same context tag and overlapping locator sets. Thus, the Peer concludes that the Initiator has lost the original context (which is wrong). It discards the old context and sends an R2 message again. Finally the context is successfully established (case 6). This a typical context confusion situation predicted by the SHIM6 protocol specifications. The peer detecting such a situation must not keep two contexts in an established state having the same context tag and having overlapping locator sets.
- (Peer, L3,*): if the Peer executes an L3 handover during a context establishment, the SHIM6 context will never be established (case 10, 11, 12). In fact, if the Peer moves during the I1-R1 phase, it will not receive the I1 message as the Initiator does not know its new address. Similarly, if the Peer moves during the R1-I2 phase or the I2-R2 phase, it will not receive the I2 message.

As a conclusion, if the Peer executes an L3 handover during the context establishment, the context cannot be established. Moreover, if the Initiator executes an L3 handover during the R1-I2 phase the context is not established. In all the other cases, the SHIM6 context can be established with additional delays. In the following, we aim to reduce the SHIM6 establishment delays by using movement detection triggers.

3) Movement detection optimization: The study in the previous section shows that in the majority of the cases, mobility during context establishment leads to retransmitting SHIM6 control messages (cases 1, 3, 4, 6, 7, 8 and 9 of Fig. 3).

The retransmission mechanism in SHIM6 is controlled by a backoff timer having an initial value of 4s. As the SHIM6 messages are lost during the execution of the L2 handover, we propose to improve the SHIM6 retransmission timer by coupling SHIM6 with a movement detection trigger. We use link-layer hints in order to retransmit lost messages quickly. When a new link-layer association is established, we stop the retransmission timer and send again the last SHIM6 control message. If the retransmitted SHIM6 control message was really lost then our proposal significantly reduces the context establishment latency. Otherwise, if the retransmitted SHIM6 message was not lost, then we are in a case of a received duplicated SHIM6 message. In such a case, the receiver will silently discard it as it has an old nonce.

The movement detection optimization will undoubtedly improve the context establishment time in a mobile environment. In the next section, we quantify this improvement through experiments.

C. Evaluation

In this section, we quantify analytically and through measurements on a real testbed the impact of node mobility on the SHIM6 context latency. Our evaluation covers both the L2 handover execution case and a L3 handover execution case during the SHIM6 context establishment. We aim to prove that our proposed movement detection optimization, presented in section IV-B3, significantly reduces the SHIM6 context latency during node movement.

We use in our evaluation the same testbed presented in Fig. 2. We integrated to the SHIM6 implementation developed by the UCL University [34] a movement detection trigger which helps us to quickly retransmit SHIM6 control messages when needed.

1) L2 handover case: As explained in section IV-B, if the Initiator executes an L2 handover during the I1-R1 (case 1 in Fig. 3, (Initiator, L2 ,I1-R1)), it will retransmit the I1 message; whereas, if the Initiator executes an L2 handover during the I2-R2 (case 3 in Fig. 3, (Initiator, L2, I2-R2)), it will retransmit the I2 message. If the movement occurred during the R1-I2 phase, no retransmission is needed (case 2 in Fig. 3, (Initiator, L2, I2-R2)). Hence, in the following, we focus on case 1 and case 3.

We define the context establishment latency as the elapsed time from the sending of the first I1 message to the reception of the R2 message. Let Δ_{ce} be the context establishment latency.

In a first case, we assume that the Initiator waits for a timeout set to 4s before retransmitting its pending SHIM6 control message. Therefore, after a timeout expiration and the exchange of 4 messages, Δ_{ce} will be equal to :

$$\Delta_{ce} = T_{\text{Timeout}} + 4 * T_{\text{OneWayDelay}} \quad (1)$$

In this first case (where we use a timeout), Δ_{ce} is theoretically equal to 4.2s (see Eq. 1), where T_{Timeout} is equal to 4s and $T_{\text{OneWayDelay}}$ is equal to 50ms. In Fig. 4, we plot Δ_{ce} for the case presented by Eq. 1. In our experiment, if the L2 handover occurred during the I1-R1, Δ_{ce} is equal to 4,209s; whereas if the L2 handover occurred during I2-R2 Δ_{ce} is equal to 4,189s. We observed by experimentation approximately the same result as we found by theory.

In a second case, we assume that the Initiator implements a movement detection trigger to quickly retransmit the pending SHIM6 message. When the Initiator detects that a new link-layer association is established, it stops its retransmission timer and sends again the last SHIM6 control message. Thus, if we take into account the movement detection optimization, Δ_{ce} will be equal to :

$$\Delta_{ce} = T_{\text{StartL2handover}} + T_{\text{L2handover}} + 4 * T_{\text{OneWayDelay}} \quad (2)$$

The term $T_{\text{StartL2handover}}$ refers to the time between the sending of a SHIM6 message and the execution of the L2 handover.

In this second case (where we use a movement detection trigger), Δ_{ce} is theoretically between 0.5s and 0.6s (see Eq. 2). $T_{\text{StartL2handover}}$ is between 0 and $2*T_{\text{OneWayDelay}}$ as the L2 handover is executed either just after sending a SHIM6 message or just before receiving a response. The L2 handover latency is evaluated at 0.3s and $T_{\text{OneWayDelay}}$ is equal to 50ms .

In Fig. 4, we plot Δ_{ce} for the case presented by Eq. 2. We observed by experimentation approximately the same result as we found by theory (0.583s for an L2 handover during the I1-R1 phase and 0.565s for an L2 handover during the I2-R2 phase).

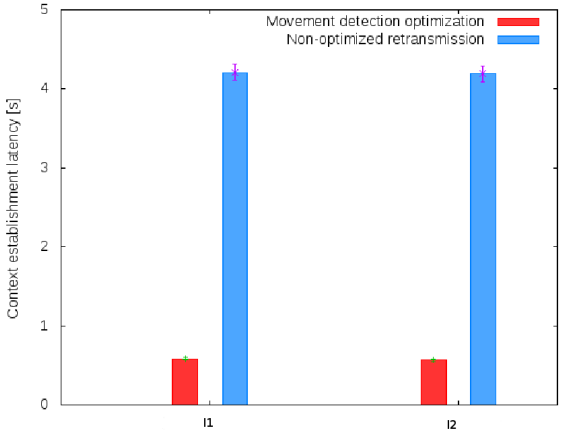


Fig. 4. Context establishment latency in case of Initiator movement

2) L3 handover case: When the mobile node executes an L3 handover during the context establishment, the context is successfully established only in cases 4 and 6 (see Fig. 3). The only difference between case 4 and 6 is the retransmitted SHIM6 message (I1 or I2). Thus, in both cases we obtain the same SHIM6 context latency.

Assuming we use our movement detection optimization, Δ_{ce} is equal to:

$$\Delta_{ce} = T_{\text{StartL2handover}} + T_{\text{L2handover}} + T_{\text{Discovery}} + T_{\text{DAD}} + 4 * T_{\text{OneWayDelay}} \quad (3)$$

T_{DAD} refers to the time of execution of the Duplicated Address Detection algorithm which verifies the uniqueness of the new address [35]. In [36], we evaluate T_{DAD} at 1s. $T_{\text{Discovery}}$ refers to the required time to discover a new network, this value is correlated to the delay between two successive Router Advertisement (RA) messages .

We consider in the following that the RA are randomly sent between 30ms and 70ms. $T_{\text{StartL2handover}}$ + $T_{\text{L2handover}}$ is equal to 0.583s , $T_{\text{Discovery}}$ is equal to 0.16s and T_{DAD} is equal to 1s. Therefore, Δ_{ce} is equal to 1.743s.

3) Conclusion: We note that without a movement detection mechanism, moving while a SHIM6 context is being established results into a consequent delay of the context establishment (4.2s in the case of the L2 handover). While the context is being established, ongoing communications are not protected against possible failures. Thus, it is important to reduce Δ_{ce} when one of the communicating SHIM6 nodes moves.

D. Mobility after context establishment

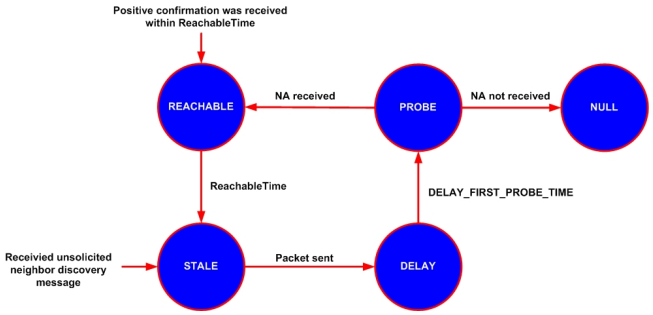


Fig. 5. Neighbor unreachability protocol state machine

We study in this section the mobility consequences on an already established SHIM6 context. We only focus on L3 handover as it affects the set of available locators and thus it might affect the established SHIM6 context.

The L3 handover can be divided into three steps: the first step is the discovery of a new network; the second step concerns the configuration of a new address and the verification of its uniqueness and finally the third step deals with the verification of the reachability of the old address and the update of the neighbor cache entry [37]. In the following we clarify the protocols involved in each step of the L3 handover execution.

A mobile node discovers a new network through the reception of an RA message containing a new IPv6 prefix. After the configuration of a new IPv6 address, the mobile node starts the Duplicated Address Detection (DAD) protocol by sending Neighbor Solicitation (NS) messages [35].

After acquiring a new address, the mobile node should first check whether its old access router (AR) is still reachable. If the old AR is still reachable, the mobile node can keep using it as a default AR, otherwise it must select a new AR. In IPv6, nodes use the Neighbor Unreachability Detection algorithm (NUD) to confirm the unreachability of their old AR after configuring a new address [35]. The node maintains a list of its neighbors in a neighbor cache with an associated state (see Fig. 5). If the entry associated to a neighbor reaches the *PROBE* state, an active reachability confirmation is launched. The node sends three NS messages to confirm or infirm neighbor reachability. If the neighbor is reachable, it responds with Neighbor Advertisement (NA) message to the received NS message. NS and NA are defined in the Neighbor Discovery Protocol [35].

After performing NUD, the mobile node should inform its correspondent node that its current address/location has changed and the traffic needs to be re-routed to its new location. In SHIM6, the mobile node sends to the correspondent node an Update Request message. The correspondent node replies with an Update Acknowledgment if the source address of the Update Request message belongs to its context. If a correspondent node receives an Update Request message from an unknown address but having a context tag bound to an already established context, it concludes that it has a stale context and re-initiates it by sending an R1Bis message. In this case, the mobile node responds with an I2Bis message. Update Request, Update Acknowledgment, R1Bis and I2Bis are SHIM6 control messages defined in [28].

As after executing an L3 handover, the mobile node preferred address may become unreachable, the mobile node must then rehome its communication to another working address pair. The default rehoming decision strategy based on the unreachability detection mechanism of the NUD protocol does not meet the requirement of the mobile environment, as the unreachability of the current AR is detected at a later time. In fact, The DAD phase lasts 1s, and the NUD probing phase lasts 3s. Hence, the whole rehoming procedure lasts at least 4s.

E. Optimizations

The SHIM6 rehoming latency is an obstacle to using SHIM6 with an application that has real time requirements in a mobile environment because data packets are lost this time.

Either during the L2 handover, the network discovery and the IPv6 address configuration, or during the exchange of the Update Request/Update Acknowledgment, data packets are lost. Indeed, before the complete update of the Peer SHIM6 context, data packets cannot reach the Initiator new location.

Data packets lost during the update of the SHIM6 context may have an impact on the operation of TCP which may degrade the quality of the application perceived by the user. Packet losses are interpreted by TCP as a congestion indication leading TCP to reduce its congestion window and retransmit the lost segments [33].

We propose in the following some optimizations in order to improve the rehoming latency in a mobile environment.

1) Fast Router Discovery: Upon the execution of an L2 handover, the Initiator waits for the reception of an RA message which is periodically sent by the AR. The time between the end of the L2 handover and the reception of the new prefix corresponds to the discovery phase. We can improve the latency of this phase by using an L2 trigger to rapidly discover new routers [38]. As soon as the Initiator gets attached to a new AR, it immediately sends a Router Solicitation message (RS). The reception of the RS message by the AR triggers the sending of the RA.

2) Optimized DAD: In IPv6, the address configuration mechanism is controlled by an algorithm called Duplicate Address Detection and lasts 1s [35]. While this latency is tolerable for a node which has just booted and configured its addresses, it does not meet the requirements of real-time applications embedded on an Initiator executing an L3 handover. Before configuring its own address, the node verifies its uniqueness by sending NS message. As the probability that two nodes configure the same address is low, the Optimized DAD procedure suggests immediately

using the newly configured address and concurrently verifies its uniqueness [39].

3) Fast NUD: As explained in section IV-D, the NUD protocol is controlled by a state machine. The verification of the reachability of an address is executed if its entry in the neighbor cache is in the *PROBE* state. In order to improve the NUD execution, we add an L2 trigger to indicate whether an Initiator has just finished an L2 handover. After attaching a new AR, the Initiator changes the state associated with its old AR to *PROBE*. Moreover, in the original specification of NUD, the probing is performed through the sending of a NS message each second. After three unsuccessful trials, the entry is deleted. Hence, the probing phase last 3s when the neighbor is unreachable. We propose to reduce the time between two successive NS to 0.2s, which is a more realistic period for the mobile environment.

F. Evaluation

In this section, we empirically evaluate the consequences of the SHIM6 node mobility on their established SHIM6 context. We estimate the required time to update a SHIM6 context and to rehome a session in a mobile environment. The measurements are conducted on a real testbed presented in Fig. 6.

Our testbed involves three ARs. We use the same access technologies in the testbed (802.11). Nonetheless, this does not prevent to have distinct access technologies, for example one is provided by AR3, the other is provided by both AR1 and AR2. Our testbed also involves an Initiator having two wireless interfaces Eth1 and Eth2. Eth1 is always connected to AR3 while Eth2 changes its point of attachment from AR1 to AR2.

The mobile SHIM6 nodes involved in our testbed embed a modified Linux IPv6 network stack. In order to update the SHIM6 context, we add to the SHIM6 implementation developed by the UCL University [34] the following control messages and their interactions with the context: Update Request, Update Acknowledgment, R1Bis and I2Bis. We modify the Linux kernel IPv6 stack (2.6.17.11 version) with a new implementation of the NUD protocol, we add the Optimistic DAD support to the address configuration, and finally we implement the link-layer trigger to ensure a fast network discovery and a fast retransmission of the pending messages.

1) Metric: We measure in this scenario the context update latency. Let Δ_{cu} be the context update latency. In our scenario, Δ_{cu} is equal to:

$$\Delta_{cu} = T_{\text{L2handover}} + T_{\text{DiscoveryStart}} + T_{\text{Discovery}} + T_{\text{UR-UA}} \quad (4)$$

The $T_{\text{DiscoveryStart}}$ latency refers to the time between the end of the L2 handover execution and the sending of the RS message. The $T_{\text{Discovery}}$ latency corresponds to the exchange of RS and RA messages. Finally the $T_{\text{UR-UA}}$ latency corresponds to the update of the SHIM6 context of the Peer with the newly acquired address.

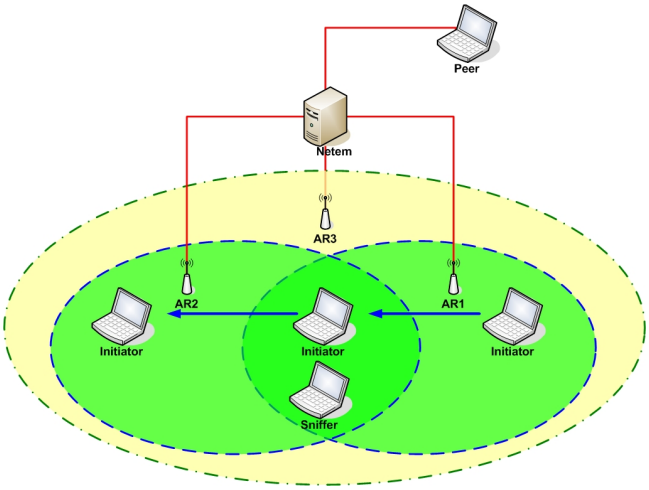


Fig. 6. Testbed 2: Moving after the SHIM6 context establishment

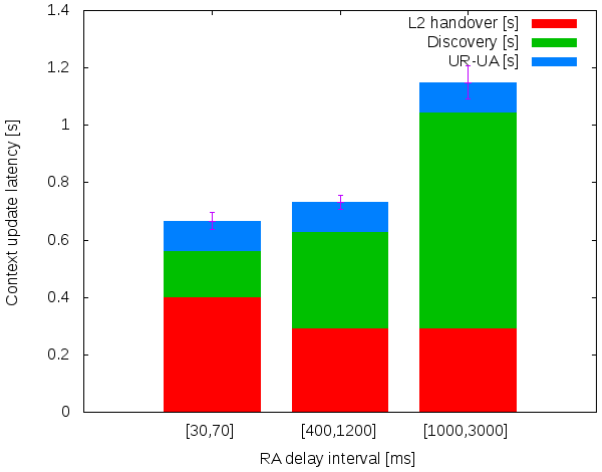


Fig. 7. SHIM6 context update latency vs. RA delay

Fig. 7 presents an evaluation of Δ_{cu} for three RA delay intervals. We observe that while the RA delay increases, Δ_{cu} increases (Δ_{cu} is respectively equal to 0.7s, 0.74s and 1.16s). The increase of the Δ_{cu} is due the increase of the latency of the discovery phase ($T_{\text{Discovery}}$).

Upon receiving an RS message, a router computes a random time between 0 and MAX_RA_DELAY_TIME to generate an RA in response to the solicitation. If this computed value corresponds to a time later than the next scheduled (multicast) RA time, then the RA is sent at its scheduled time (and the random value is ignored). If an RA was sent within the last MIN_DELAY_BETWEEN_RAS, then the next RA should be sent at a time equal MIN_DELAY_BETWEEN_RAS plus the random value between 0 and MAX_RA_DELAY_TIME, in order to rate limit the RAs. In Neighbor Discovery for IPv6 [35], MAX_RA_DELAY_TIME is set to 0.5s and MIN_DELAY_BETWEEN_RAS is set to 3s; whereas, in Mobility Support in IPv6 [40], if the lower limit of the RAs interval (MinRtrAdvInterval) is less than 3s, MIN_DELAY_BETWEEN_RAS is set to this limit and

its minimum value is 30ms.

Therefore, whenever we increase MinRtrAdvInterval, the time to send the scheduled RA in response to the RS increases and thus, $T_{\mathrm{Discovery}}$ increases.

We obtain almost the same value of Δ_{cu} in the two first experiments because the decrease of the $T_{\mathrm{L2handover}}$ almost compensates the increase of the $T_{\mathrm{Discovery}}$. However, for the third result, $T_{\mathrm{L2handover}}$ does not vary while $T_{\mathrm{Discovery}}$ decreases. The L2 handover latency ($T_{\mathrm{L2handover}}$) increases when we send RAs at the highest possible rate because of the saturation of the wireless medium induced by the sending of several RA messages. When RAs are sent at a high rate, the wireless medium is saturated and therefore, the L2 handover execution lasts longer (400ms instead of 300ms). Finally, we notice that $T_{\mathrm{DiscoveryStart}}$ latency does not appear in Fig. 7 as it lasts between 2 and 3ms.

This experiment shows that in a mobile environment the context update latency amounts to 0.7s in the best case. The currently obtained time is an obstacle to use SHIM6 with applications that have real time requirements.

The major contributors to the context update latency are the L2 handover and the network discovery latency. If we consider the case where RAs are sent between 30ms and 70ms, the L2 handover latency represents 57%, and the network discovery represents 28% of the whole context update latency.

Several works (such as [41]) proposed to optimize the L2 handover which can significantly reduce this latency. Depending on the wireless card vendor, the reduction can reach 90% of the overall L2 handover latency (from 300ms to approximately 20ms).

V. Conclusion

We have studied in this paper the impact of node mobility on the SHIM6 multihoming protocol which was originally intended to handle multihoming for static nodes. we focused on the mobility impact on the SHIM6 context establishment as it constitutes its key component. We showed that performing either an L2 or an L3 handover while establishing the SHIM6 context may delay the context establishment. We exacerbated the cases where performing an L3 handover may lead to a context establishment failure. We evaluated the mobility impact during the context establishment on developed testbeds and measured the context establishment latency.

Furthermore, we investigated mobility impacts on an already established SHIM6 context. We showed that the default rehoming decision strategy of the SHIM6 protocol does not meet the mobility requirement. Consequently, we proposed some viable optimizations to enhance the default rehoming decision strategy. The rehoming latency of this enhanced rehoming decision strategy is then evaluated on a developed testbed.

We showed through extensive experiments that SHIM6 can indeed manage mobility by itself. However, the measured rehoming latency stays an obstacle to using SHIM6 for real-time applications. In the quest to reduce the rehoming latency, we are investigating the possibility of executing the NUD in parallel to the ongoing communication in order to proactively detect possible unreachability of the current address. Furthermore, we are currently studying the use of the link identification in the RA message (proposed by the DNA IETF working group) in order to enhance the detection of node mobility. Moreover, as SHIM6 does not take into account a registrar entity in its architecture, it cannot solve the double jump problem (i.e., the two communicating SHIM6 nodes simultaneously move). We are also targeting to propose ways to solve the double-jump problem.

References

[1] J. Day, *Patterns in Network Architecture: A Return to Fundamentals*. Prentice Hall, 2007.

[2] S. Agarwal, C.-N. Chuah, and R. H. Katz, "OPCA: robust interdomain policy routing and traffic control," in *Open Architectures and Network Programming, 2003 IEEE Conference on*, Apr. 2003, pp. 55–64.

[3] C. Launois and M. Bagnulo, "The paths towards IPv6 multihoming," in *IEEE Communications Surveys and Tutorials*, vol. 8, 2006, pp. 38–51.

[4] M. Bagnulo, A. G. Martinez, A. Azcorra, and C. de Launois, "An incremental approach to ipv6 multihoming," *Computer Communications*, vol. 29, no. 5, pp. 582–592, 2006, networks of Excellence.

[5] J. F. Shoch, "Inter-network naming, addressing, and routing." Washington, D.C.: IEEE, Sep. 1978, pp. 72–79.

[6] D. Le, X. Fu, and D. Hogrefe, "A review of mobility support paradigms for the internet," *IEEE Communications Surveys & Tutorials*, vol. 8, no. 1, pp. 38–51, / 2006.

[7] P. Savola and T. Chown, "A survey of IPv6 site multihoming proposals," in *Telecommunications, 2005. ConTEL 2005. Proceedings of the 8th International Conference on*, vol. 1, Jun. 2005, pp. 41–48.

[8] J. Abley, B. Black, and V. Gill, "Goals for IPv6 Site-Multihoming Architectures," RFC 3582 (Informational), Internet Engineering Task Force, Aug. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3582.txt

[9] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *SIGCOMM '03: Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2003, pp. 353–364.

[10] C. de Launois, B. Quoitin, and O. Bonaventure, "Leveraging network performance with IPv6 multihoming and multiple provider-dependent aggregatable prefixes," *Comput. Netw.*, vol. 50, no. 8, pp. 1145–1157, 2006.

[11] D. K. Goldenberg, L. Qiuy, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," in *SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications*. New York, NY, USA: ACM, 2004, pp. 79–92.

[12] J. Saltzer, "On the Naming and Binding of Network Destinations," RFC 1498 (Informational), Internet Engineering Task Force, Aug. 1993. [Online]. Available: http://www.ietf.org/rfc/rfc1498.txt

[13] G. Huston, "Multi-homing and identity in ipv6," Internet Society Publications, June 2004.

[14] S. Herborn, R. Boreli, and A. Seneviratne, "Identity location decoupling in pervasive computing networks," in *Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on*, vol. 2, Mar. 2005, pp. 610–615.

[15] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet routing instability," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 515–528, Oct. 1998.

[16] M. Bagnulo, A. Garcia-Martinez, and A. Azcorra, "Efficient Security for IPv6 Multihoming," *ACM Computer Communications Review*, vol. 35, no. 2, pp. 61–68, April 2005.

[17] S. Uhlig and O. Bonaventure, "Designing bgp-based outbound traffic engineering techniques for stub ases," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 5, pp. 89–106, 2004.

[18] G. Huston, "Bgp reports," 2009, http://bgp.potaroo.net/.

[19] T. R. Henderson, "Host mobility for IP networks: a comparison," *IEEE Network*, vol. 17, no. 6, pp. 18–26, Nov./Dec. 2003.
[20] Y. Rekhter and T. Li, "An Architecture for IP Address Allocation with CIDR," RFC 1518 (Historic), Internet Engineering Task Force, Sep. 1993. [Online]. Available: http://www.ietf.org/rfc/rfc1518.txt
[21] C. Vogt, "Six/one router: a scalable and backwards compatible solution for provider-independent addressing," in *MobiArch '08: Proceedings of the 3rd international workshop on Mobility in the evolving internet architecture*. New York, NY, USA: ACM, 2008, pp. 13–18.
[22] "Standard for information technology - portable operating system interface (posix). shell and utilities," 2004. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1309816
[23] C. Huitema, "Multi-homed tcp," draft-huitema-multi-homed-0 (work in progress), May 1995. [Online]. Available: http://tools.ietf.org/html/draft-huitema-multi-homed-01
[24] A. Matsumoto, M. Kozuka, and K. Fujikawa, "Tcp multi-home options," draft-arifumi-tcp-mh-00 (work in progress), October 2003. [Online]. Available: http://tools.ietf.org/html/draft-arifumi-tcp-mh-00
[25] R. Stewart, "Stream Control Transmission Protocol," RFC 4960 (Proposed Standard), Internet Engineering Task Force, Sep. 2007. [Online]. Available: http://www.ietf.org/rfc/rfc4960.txt
[26] E. Kohler, "Datagram congestion control protocol mobility and multihoming," draft-kohler-dccp-mobility-01 (work in progress), January 2006. [Online]. Available: http://tools.ietf.org/html/draft-kohler-dccp-mobility-01
[27] P. Savola, "Site Multihoming: A Microscopic Analysis of Finnish Networks," in *Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, 2006. ICN/ICONS/MCL 2006. International Conference on*, Apr. 2006, pp. 25–25.
[28] E. Nordmark and M. Bagnulo, "Shim6: Level 3 Multihoming Shim Protocol for IPv6," RFC 5533 (Proposed Standard), Internet Engineering Task Force, 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5533.txt
[29] M. Bagnulo, "Hash-Based Addresses (HBA)," RFC 5535 (Proposed Standard), Internet Engineering Task Force, Jun. 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5535.txt
[30] T. Aura, "Cryptographically Generated Addresses (CGA)," RFC 3972 (Proposed Standard), Internet Engineering Task Force, Mar. 2005, updated by RFCs 4581, 4982. [Online]. Available: http://www.ietf.org/rfc/rfc3972.txt
[31] M. Bagnulo, A. García-Martínez, and A. Azcorra, "Fault tolerant scalable support for network portability and traffic engineering," in *WWIC '07: Proceedings of the 5th international conference on Wired/Wireless Internet Communications*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 129–140.
[32] J. Arkko and I. van Beijnum, "Failure Detection and Locator Pair Exploration Protocol for IPv6 Multihoming," RFC 5534 (Proposed Standard), Internet Engineering Task Force, Jun. 2009. [Online]. Available: http://www.ietf.org/rfc/rfc5534.txt
[33] A. de la Oliva, M. Bagnulo, A. Garcia-Martinez, and I. Soto, "Performance Analysis of the REAchability Protocol for IPv6 Multihoming," in *Next Generation Teletraffic and Wired/Wireless Advanced Networking 7th International Conference, NEW2AN 2007*, September 2007, pp. 443–454.
[34] S. Barré, "Linshim6 - implementation of the shim6 protocol," Université catholique de Louvain, Tech. Rep., Feb 2008.
[35] T. Narten, E. Nordmark, W. Simpson, and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)," RFC 4861 (Draft Standard), Internet Engineering Task Force, Sep. 2007, updated by RFC 5942. [Online]. Available: http://www.ietf.org/rfc/rfc4861.txt
[36] A. Dhraief and N. Montavont, "Toward Mobility and Multihoming Unification- The SHIM6 Protocol: A Case Study," in *Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE*, Las Vegas, Nevada/USA, Mar./Apr. 2008, pp. 2840–2845.
[37] N. Montavont and T. Noel, "Handover management for mobile nodes in IPv6 networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 38–43, Aug. 2002.
[38] G. Daley, B. Pentland, and R. Nelson, "Movement detection optimizations in mobile IPv6," in *Networks, 2003. ICON2003. The 11th IEEE International Conference on*, Sep. 28–Oct.1, 2003, pp. 687–692.
[39] N. Moore, "Optimistic Duplicate Address Detection (DAD) for IPv6," RFC 4429 (Proposed Standard), Internet Engineering Task Force, Apr. 2006. [Online]. Available: http://www.ietf.org/rfc/rfc4429.txt
[40] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775 (Proposed Standard), Internet Engineering Task Force, Jun. 2004. [Online]. Available: http://www.ietf.org/rfc/rfc3775.txt
[41] M. Shin, A. Mishra, and W. A. Arbaugh, "Improving the latency of 802.11 hand-offs using neighbor graphs," in *MobiSys '04: Proceedings of the 2nd international conference on Mobile systems, applications, and services*. New York, NY, USA: ACM, 2004, pp. 70–83.

Dr. Amine Dhraief received his Ph.D. degree in computer science from Telecom Bretagne University of Rennes I in 2009, his Master of Science degree and his Engineering degree both computer science from the National School of Computer Sciences (ENSI), University of Manouba respectively in 2006 and 2005. He is currently an Assistant Professor of Computer Science at ISIGK University of Kairouan, Tunisia. He is also member of the HANA research group at the National School of Computer Sciences, University of Manouba, Tunisia. His current research interests include pervasive and ubiquitous computing and general-purpose computation on graphics processing units.

Dr. Abdelfettah Belghith received his Master of Science and his PhD degrees in computer science from the University of California at Los Angeles (UCLA) respectively in 1982 and 1987. He is since 1992 a full Professor at the National School of Computer Sciences (ENSI), University of Manouba, Tunisia. His research interests include computer networks, wireless networks, multimedia Internet, mobile computing, distributed algorithms, simulation and performance evaluation. He runs several projects in cooperation with other universities, research laboratories and research institutions. He is currently the Director of the Doctoral School (Graduate School) STICODE of the University of Manouba, the responsible of the Network and Distributed Systems Master degree and the head of the HANA Research Group at the National School of Computer Sciences.

Development of Anonymous Networks Based on Symmetric Key Encryptions

Hazim Haddad
University of Fukui, Fukui, Japan
Email: hazimhaddad@yahoo.com

Shinsuke Tamura, Shuji Taniguchi and Tatsuro Yanase
University of Fukui, Fukui, Japan
Email: {tamura, stamigut, yanase}@u-fukui.ac.jp

***Abstract*—Anonymous networks enable entities to send messages without disclosing their identities. Many anonymous networks had been proposed already, such as Mixnet, DC-net, Crowds, etc., however, they still have serious drawbacks. Namely, they require tremendous computation overheads to transmit messages over networks. That is because asymmetric key encryption algorithms are used. This paper proposes ESEBM (Enhanced Symmetric Key Encryption based Mixnet), a new mechanism for anonymous communication that removes drawbacks of existing anonymous networks while exploiting symmetric key encryption algorithms. According to experimentations, throughput of ESEBM is about 1/4.4 of usual non-anonymous networks, and it achieves more than 36 times higher throughput compared with Mixnet. In addition, different from existing anonymous networks, ESEBM can handle reply messages without any additional mechanism, and it can protect itself from various threats, e.g. DOS attacks and message forgeries.**

***Index Terms*—anonymous communication, mixnet, privacy protection, symmetric key encryption algorithm**

I. Introduction

Identities of message senders are sometimes as sensitive as messages themselves. For example, a company may acquire highly confidential information about its rival companies from identities of their customers and suppliers. Therefore, the importance of anonymous communication is increasing as more people are being involved in network based communication. Anonymous networks are ones that enable message senders to send their messages without disclosing their identities, and various anonymous networks had been proposed already, e.g. Mix net [1, 5, 9], DC-net [2], Crowds [4], etc., to protect secrets of entities that communicate through networks. However, they still have serious drawbacks. For example, although Mix net is one of the most promising mechanisms, it requires the tremendous amount of computations to encrypt/decrypt messages that are forwarded from senders to their receivers. That is because asymmetric key encryption/decryption functions are adopted. In this paper, a new anonymous network ESEBM (Enhanced Symmetric Key Encryption based Mix net) is proposed that removes drawbacks of existing anonymous networks by using symmetric key encryption functions.

ESEBM consists of two parts, they are the CP generator (offline) and the anonymous channel (online) each of which is configured as a sequence of servers, and senders obtain secret keys of individual servers in the anonymous channel for encrypting their messages from the CP generator as off-line processes. Then, once encryption keys are shared between senders and servers, servers in the anonymous channel can efficiently transfer messages of senders to their receivers while exploiting symmetric key encryption functions.

According to experimentations, the capacity of ESEBM is more than 36 times higher than that of decryption type Mix net. Different from asymmetric key encryption functions, symmetric key encryption functions also enable message receivers to send reply messages to the anonymous senders in totally the same way as the senders send original messages, and consequently, anyone except the receivers cannot identify even whether messages are replies or not. Also, the CP generator configuration disables unauthorized entities to send messages because only authorized entities that had obtained secret keys from the CP generator can send messages. Therefore, ESEBM is secure against various kinds of attacks including DOS attacks and message forgeries (or modifications) that are difficult to prevent in existing anonymous networks.

II. Requirements for Anonymous Networks

Anonymous networks should satisfy the following requirements, i.e.,

1. no one except senders of messages can know identities of the senders,
2. message senders can confirm their message arrivals at their receivers without disclosing their identities,
3. receivers can send reply messages back to the senders without knowing the senders' identities,

* Graduate School of Engineering, University of Fukui
3-9-1, Bunkyo, Fukui 910-8507, Japan

doi:10.4304/jnw.6.11.1533-1542

4. anonymous networks must be able to protect themselves from accesses from unauthorized entities, and
5. anonymous networks must maintain their performances as same as usual ones.

The 1st requirement is the most important one, and senders of messages must be concealed not only from the receivers but also from network managers, eavesdroppers and any other entities. The 2nd and the 3rd requirements are also important, and especially the 3rd one is essential because information exchanges between entities in many kinds of applications are carried out as conversations between them. To satisfy the 2nd requirement is not so difficult, e. g. senders can confirm deliveries of their messages without disclosing their identities when the receivers put receive signals in public bulletin boards. However, development of practical mechanisms that satisfy the 3rd requirement is not easy as it looks. For example, a receiver, which sends reply message M_R, can identify the sender of the original message by eavesdropping on the communication channel to find out the entity that receives M_R, because it knows M_R. About the 4th requirement, because of anonymity, entities can behave dishonestly much easier than in usual communication systems, therefore, anonymous communication mechanisms must be endowed with the ability to protect them from dishonest events. The important thing here is that dishonest events must be prevented while maintaining anonymities of honest entities. Finally, to use anonymous networks in large scale applications where large volumes of messages are exchanged frequently, they must be efficient enough as usual non-anonymous networks.

III. Related Works

This section summarizes currently available anonymous networks. Although many various kinds of anonymous networks had been proposed already, still they cannot satisfy the requirements in the previous section effectively. Mixnet is an example. It consists of a sequence of mixservers T_1, T_2, ---, T_N, that relay messages from senders to their receivers. Where, senders send their messages while encrypting them repeatedly by public keys of multiple mixservers T_1, T_2, ---, T_N in the sequence. Then, individual mixservers relay their receiving messages to their neighboring servers while decrypting them by their secret decryption keys finally to be sent to their receivers. Namely, sender S encrypts its message M to $E(k_N, E(k_{N-1}, ---, E(k_1, M) ---))$ and each T_j that recieves $E(k_j, E(k_{j-1}, ---, E(k_1, M)---))$ from T_{j+1} decrypts it to $E(k_{j-1}, ---, E(k_1, M)---)$ by its secret decryption key k_j^{-1} to forward it to T_{j-1}, where $E(k_j, M)$ is the encrypted form of M. In this message relaying process, each mixserver stores its incoming messages until pre-defined number of message arrivals, and shuffles decrypted messages before forwarding them to its neighbor. Therefore, each mixserver cannot identify the links between incoming and outgoing messages of other mixservers, and as a consequence, no one except the senders themselves can identify the senders of messages unless all mixservers conspire.

However, Mixnet uses asymmetric key encryption functions, such as RSA or ElGamal, and does not work efficiently in large scale systems where number of senders send large volume of messages. A lot of computation overheads are needed to encrypt and decrypt messages. Asymmetric key encryption functions also make Mixnet require additional mechanisms for sending reply messages to senders of the original messages, therefore, servers can know whether the messages are replies or not [1, 7]. Although Mixnet can protect itself from traffic analysis and replay attacks that are discussed in Sec. VI. A, it cannot prevent DOS attacks or message forgeries (or modifications). Encryption keys are publicly disclosed and servers cannot identify spam or forged messages because they receive messages in their encrypted forms, therefore, anyone can send spam and forged messages.

Crowds [4] also consists of multiple relay servers as same as Mixnet, however, senders send their messages without encrypting them. Instead of encrypting messages, servers randomly decide whether to relay their receiving messages to their receivers or to the other servers in the network. Namely, when a server receives a message from a sender, it forwards it to other server with probability 1-p, and with probability p it sends it to the receiver. Then, it becomes difficult for entities other than the sender to identify the sender, and because no encryption or decryption process is included, Crowds can transfer messages efficiently. However, apparently it cannot disable entities to identify senders by tracing messages from their receivers to their senders. Namely, Crowds cannot satisfy the most important requirement of anonymous networks.

Onion routing [3, 8] uses the same principle as Mixnet, i.e. messages travel from senders to receivers through sequences of servers (onion routers) while being encrypted by public keys of multiple onion routers. The difference from Mixnet is that senders in onion routing encrypt not only their messages but also their routes, i.e. servers in onion routing reroute their receiving messages in unpredictable ways. Therefore, onion routers need not wait for large number of messages to shuffle them and can reduce message travelling times. However, onion routing uses asymmetric key encryption functions and shares the same drawbacks with Mixnet. An additional problem of onion routing is that it is vulnerable to timing attacks, i.e. an adversary can embed messages to know the flow times of different paths. Then, while using these message flow times, entities can know senders of messages by observing message sending and receiving times of individual senders and receivers.

Other anonymous networks such as Tor [8], buses for anonymous message delivery [6], Peer to Peer anonymous mechanisms [12], etc. have the same drawbacks as Mixnet or Onion routing.

In DC-net [2], sender S_q constitutes a group $\{S_1, S_2, ---, S_Q\}$ that includes itself, and entities in the group generate their secret numbers $\{N_1, N_2, ---, N_Q\}$ so that the sum of

them becomes 0 in advance. While using its generating secret number, S_q encrypts its message M to $M + N_q$ to send it to its receiver R. At the same time, each S_j in the group also sends its secret number N_j to R. Therefore, R can extract M from messages of $\{S_1, S_2, ---, S_Q\}$, i.e. $N_1 + N_2 + --- + (M + N_q) + N_{q+1} + --- + N_Q = M + 0 = M$. However, no one except S_q can know the sender of M, because each S_j does not know secret numbers of other senders.

As shown above, DC-net provides almost perfect anonymity, however it has fatal drawbacks about its performance, i.e. multiple senders must behave synchronously. Multiple senders must agree with each other about random numbers to encrypt messages, also only one sender can send a message at a time. Therefore, it is applicable only to small and closed networks. Here, it must be noted that each S_j must change random secret number N_j at every message sending. If every S_j uses same random secret number for different messages sent from senders in the group, an entity X that eavesdrops on the communication can easily identify senders of the messages. Namely, when S_j sends same number N_j as its 1st and 2nd messages, X can know that S_j's random secret number is N_j. Also, when S_j sends $(M_j + N_j)$ and N_j as its 1st and 2nd messages, it is easy for X to extract M_j and to identify the sender.

To decrease computation volumes of encryptions and decryptions, SEBM [13] exploits symmetric key encryption functions. SEBM consists of 2 parts, the encryption part and the decryption part, and messages are forwarded to their receivers while being encrypted by servers in the encryption part and decrypted by servers in the decryption part. Here different from other anonymous networks, senders themselves are included as relay servers in both parts to enable the use of symmetric key encryption functions. Therefore, although SEBM can satisfactory reduce the computation overheads caused by asymmetric key encryptions, senders included in the encryption and decryption parts reduce the stability of the communication. For example, when senders, i.e. volunteer servers, stop operations, messages cannot be forwarded. As another drawback, because messages in SEBM must be encrypted and decrypted by servers both in the encryption and the decryption parts, their travelling times increase. Also, it cannot efficiently handle reply messages or prevent accesses from unauthorized entities either.

IV. ESEBM (ENHANCED SYMMETRIC KEY ENCRYPTION BASED MIXNET)

This section proposes ESEBM, a scheme for anonymous networks that efficiently satisfies all the requirements listed in the previous section. ESEBM removes most drawbacks that exist in other anonymous networks, i.e. it can transfer messages without large overheads, it does not require any additional mechanism for forwarding reply messages, and it can protect itself from various attacks.

A. ESEBM Configuration

ESEBM can be considered as a kind of decryption type Mixnet, in which asymmetric key encryption functions are replaced by symmetric ones, where the encryption keys used for sending messages are distributed to senders in advance. At the same time, it is considered as SEBM in which volunteer servers are replaced by permanent ones in order to make the network stable enough [15].

As shown in Fig. 1, ESEBM consists of 2 parts, i.e. the anonymous channel and the concealing pattern generator (CP generator). The anonymous channel is configured as a sequence of N servers as same as Mixnet, and the CP generator consists of Z-groups, where the g-th group is configured by N_g servers, and each server in the anonymous channel is corresponded to a single server in the CP generator and vice versa, therefore $N = N_1 + N_2 + --- + N_Z$. In the remainder, notation $T_g(k)$ that represents the k-th server in the g-th group of the CP generator is used also for representing the p-th server T_p in the anonymous channel that corresponds to $T_g(k)$, and vice versa.

ESEBM adopts onetime pad as the base algorithm to encrypt and decrypt messages, and sender S of message M_S requests servers in the CP generator to issue a bit string called concealing pattern (CP), a pad for encrypting M_S, in advance as an off-line process.

Provided that servers generate their h-th CP at the request of S, each server T_j in the CP generator generates its h-th CP constructor $x_j(h)$, and the h-th concealing pattern X(h) is constructed as XOR of them, i.e. $X(h) = x_1(h) \oplus x_2(h) \oplus --- \oplus x_N(h)$. Then, S sends M_S to the first server T_1 in the anonymous channel while encrypting it to $M_S \oplus X(h)$. Therefore, the length of CPs and CP constructors are defined as L_M, which is the length of messages. When S sends a long message M_S, M_S is divided into multiple frames of length L_M. Here, S uses different CPs for encrypting different messages including different frames of the same message. Also, although notations X(h) and $x_j(h)$ are accompanied by h they do not include any information about h.

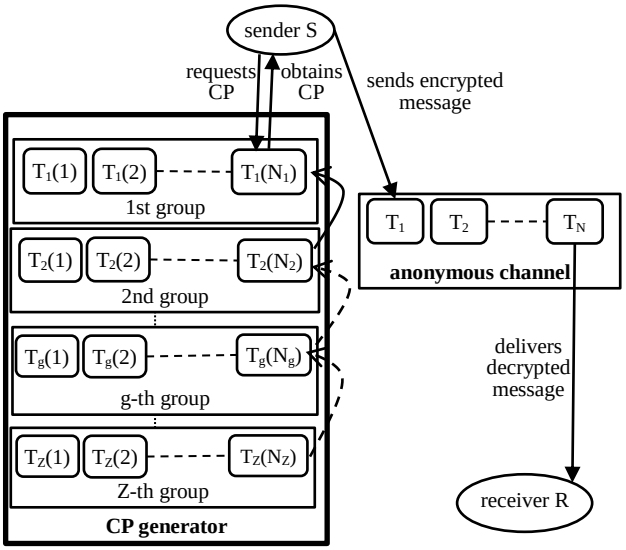


Figure 1. ESEBM configuration

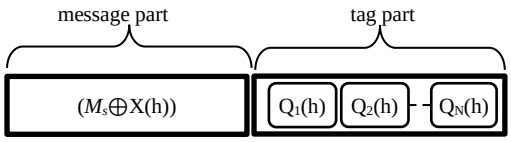


Figure 2. Message structure

As same as usual Mixnet, each server in the anonymous channel stores its receiving messages until it receives the predefined number of messages, and decrypts, shuffles and forwards them to its neighboring server finally to be sent to their receivers. Here, each T_j decrypts its receiving encrypted M_S by simply XORing it by its CP constructor $x_j(h)$ that constitutes X(h), the CP that S had used to encrypt M_S, then, it is apparent that $M_S \oplus X(h)$ is transformed to M_S when all servers decrypt it. On the other hand, because each server knows only its CP constructor $x_j(h)$ in X(h), no one can know the sender of M_S unless all servers conspire with each other as same as in usual Mixnet.

However, different from usual Mixnet where all senders encrypt their messages by using the same single public encryption key of each mixserver, in ESEBM, senders encrypt different messages by using different CPs, Therefore to enable T_j to identify its CP constructor $x_j(h)$ that constitutes X(h) for encrypting M_S, message M_S consists of the message part and the tag part as shown in Fig. 2. The message part maintains encrypted message M_S, i.e. $M_S \oplus X(h)$, and the tag part maintains a sequence of tags, i.e. vector $Q(h) = \{Q_1(h), Q_2(h), ---, Q_N(h)\}$, where server T_j that had generated the CP constructor $x_j(h)$ to construct X(h) can know $x_j(h)$ from $Q_j(h)$. Here, $Q_j(h)$ is constructed so that no one can trace the message by it and no one except T_j can identify $x_j(h)$ from it.

B. Behavior of the CP Generator

To disable entities to trace messages forwarded through the anonymous channel, not only correspondences between the message parts of input and output messages of individual servers but also those between their tag parts must be concealed. To achieve this, the CP generator generates 2 kinds of secret encryption keys shared between senders and individual servers, the one is CPs and the other is tag vectors (TVs). The CP generator is a set of server groups, each of which consists of at least 3 servers that generate their secret CP constructors and TV constructors independently of others to construct CPs and TVs jointly with other servers. Here, senders communicate only with servers in the 1st group, i.e. with $T_1(1)$, $T_1(2)$, ---, and $T_1(N_1)$, to disable servers in the other groups to know the senders as shown in Fig. 1.

As discussed already, concealing pattern X(h) is calculated as XOR of CP constructor $x_j(h)$ (j = 1, 2, ---, N) generated by each server T_j, and disables anyone to trace the message parts of a message relayed by the servers. On the other hand, individual elements of N-dimensional tag vector $Q(h) = \{Q_1(h), Q_2(h), ---, Q_N(h)\}$ disable anyone to trace the tag part of a message relayed by the servers, and each $Q_i(h)$ is calculated as XOR of the i-th elements of each N-dimensional TV constructor $q_j(h) = \{0, ---0, q_{j(j+1)}(h), q_{j(j+2)}(h), ---, q_{jN}(h)\}$ generated by T_j (j = 1, ---, N). Here, each $q_{jk}(h)$ in vector $q_j(h)$ is a bit pattern of length L_T as discussed later, 0 represents an all zero bit pattern of length L_T, and a sequence of j-zero patterns precedes before the (N-j)-secret bit patterns $\{q_{j(j+1)}(h), q_{j(j+2)}(h), ---, q_{jN}(h)\}$. By XORing CP constructors and TV constructors of individual serves, concealing pattern X(h) and tag vector Q(h) are calculated as $X(h) = x_1(h) \oplus x_2(h) \oplus --- \oplus x_N(h)$ and $Q(h) = \{0, q_{12}(h), q_{13}(h) \oplus q_{23}(h), ---, q_{1N}(h) \oplus q_{2N}(h) \oplus --- \oplus q_{(N-1)N}(h)\}$. Here, the length of bit pattern $x_j(h)$ is equal to the message frame length L_M as mentioned before, and the last server T_N does not generate its TV constructor.

CPs and TVs above are generated as follows. Provided that $T_1(k)$ in the 1st group of the CP generator corresponds to T_{k*} in the anonymous channel, i.e. $T_1(1) = T_{1*}$, $T_1(2) = T_{2*}$, ---, and $T_1(N_1) = T_{N1*}$, firstly, sender S sends a set of its secret private vectors (PVs) $\{P_1(h), P_2(h), ---, P_{N1}(h)\}$ as a request for a CP to servers T_{1*}, T_{2*}, ---, T_{N1*}, respectively, as shown in Fig. 3 (a). Here, each $P_j(h)$ is vector $\{p_{j0}(h), p_{j1}(h), ---p_{jN}(h)\}$ and except $p_{j0}(h)$, $p_{jk}(h)$ is a bit pattern of the same length as element $q_{jk}(h)$ in TV constructor $q_j(h)$. Bit pattern $p_{j0}(h)$ has the same length as CP constructor $x_j(h)$.

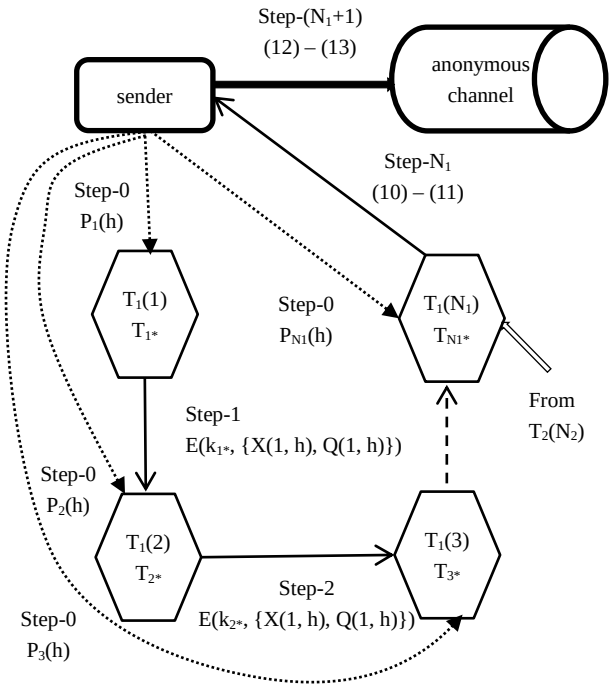
Then, T_{1*} that receives the request with $P_1(h)$, generates its CP constructor $x_{1*}(h)$ and TV constructor $q_{1*}(h) = \{0, ---, 0, q_{1*(1*+1)}(h), q_{1*(1*+2)}(h), ---, q_{1*N}(h)\}$. It also generates $ID_{1*}(x_{1*}(h), q_{1*}(h))$ as an address of CP and TV constructor pair $(x_{1*}(h), q_{1*}(h))$. Here, T_{1*} maintains its CP table, a list of CP and TV constructors that it had generated, and $ID_{1*}(x_{1*}(h), q_{1*}(h))$ represents the address of the constructor pair $\{x_{1*}(h), q_{1*}(h)\}$ in the table. Also, the length of each bit pattern $q_{jk}(h)$ in TV constructor $q_j(h)$ is set as L_T, the length of $ID_j(x_j(h), q_j(h))$.

Then, X(1, h) and Q(1, h), the h-th CP and TV that the 1st group generates, are constructed by 1st server T_1^* as $X(1, h) = p_{10}(h) \oplus x_{1*}(h)$ and $Q(1, h) = \{p_{11}(h), p_{12}(h), ---, p_{11*}(h) \oplus ID_{1*}(x_{1*}(h), q_{1*}(h)), p_{1(1*+1)}(h) \oplus q_{1*(1*+1)}(h), p_{1(1*+2)}(h) \oplus q_{1*(1*+2)}(h), ---, p_{1N}(h) \oplus q_{1*N}(h)\}$, respectively. X(1, h) and Q(1, h) are then forwarded to T_{2*}. However, to protect them from eavesdropping, they are encrypted by the secret key k_{1*} that is shared between T_{1*} and T_{2*}, i.e. X(1, h) and Q(1, h) are sent to T_{2*} in the form $E(k_{1*}, X(1, h), Q(1, h))$, where, $E(k_{1*}, x)$ represents x encrypted by key k_{1*}. It is also possible that T_{1*} encrypts X(1, h) and Q(1, h) by using a public key of T_{2*}, however to decrease encryption overheads, a symmetric key encryption function is adopted here.

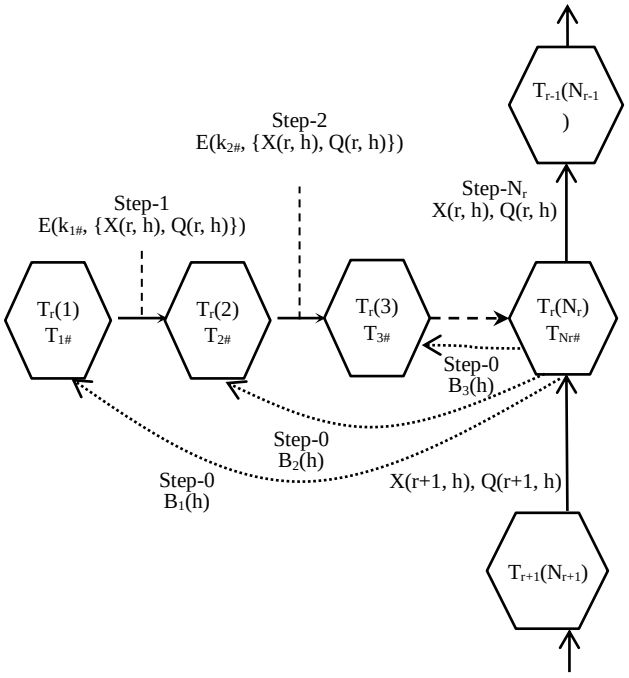
T_{2*} that receives $E(k_{1*}, \{X(1, h), Q(1, h)\})$ decrypts it to $\{X(1, h), Q(1, h)\}$, and generates its CP constructor $x_{2*}(h)$ to modify X(1, h) to $X(1, h) = p_{10}(h) \oplus p_{20}(h) \oplus x_{1*}(h) \oplus x_{2*}(h)$. T_{2*} also generates TV constructor $q_{2*}(h) = (0, ---0, q_{2*(2*+1)}(h), q_{2*(2*+2)}(h), ---, q_{2*N}(h))$ to modify Q(1, h) to $\{p_{11}(h) \oplus p_{21}(h), p_{12}(h) \oplus p_{22}(h), ---, p_{11*}(h) \oplus p_{21*}(h) \oplus ID_{1*}(x_{1*}(h), q_{1*}(h)), p_{1(1*+1)}(h) \oplus p_{2(1*+1)}(h) \oplus q_{1*(1*+1)}(h), ---, p_{12*}(h) \oplus p_{22*}(h) \oplus q_{1*2*}(h) \oplus ID_{2*}(x_{2*}(h), q_{2*}(h)), p_{1(2*+1)}(h) \oplus p_{2(2*+1)}(h) \oplus q_{1*(2*+1)}(h) \oplus q_{2*(2*+1)}(h), ---, p_{1N}(h) \oplus p_{2N}(h) \oplus q_{1*N}(h) \oplus q_{2*N}(h)\}$.

Here as same as T_{1*}, T_{2*} also maintains its CP table, and $ID_{2*}(x_{2*}(h), q_{2*}(h))$ represents the address where

$\{x_{2*}(h), q_{2*}(h)\}$ is located in it. Also, it is not necessary but to simplify the descriptions, it is assumed that servers in the anonymous channel are arranged so that $T_j(g)$ is placed at the earlier position in the anonymous channel than $T_j(h)$ when $g < h$, for every j-th group. Then, X(1, h) and Q(1, h) are sent to T_{3*} while being encrypted by k_{2*}, a secret encryption key shared between T_{2*} and T_{3*}, and this process continues until T_{N1*} calculates X(1, h) and Q(1, h). Therefore, X(1, h) and Q(1, h) = $\{Q_1(1, h), Q_2(1, h), ---, Q_N(1, h)\}$, the CP and the TV pair generated by the 1st group becomes as shown in equations (1) – (3).



(a) 1st group



(b) r-th group

Figure 3. Behaviour of the CP generator

$$X(1, h) = p_{10} \oplus p_{20} \oplus --- \oplus p_{(N1)0} \oplus x_{1*}(h) \oplus x_{2*}(h) \oplus --- \oplus x_{(N1)*}(h) \quad (1)$$

for g* included in the 1st group

$$Q_{g*}(1, h) = p_{1g*}(h) \oplus p_{2g*}(h) \oplus --- \oplus p_{(N1)g*}(h) \oplus q_{1*g*}(h) \oplus q_{2*g*}(h) \oplus --- \oplus q_{(g-1)*g*}(h) \oplus ID_{g*}(x_{g*}(h), q_{g*}(h)), \text{ where } q_{0*g*}(h) = 0 \quad (2)$$

for i not included in the 1st group,

$$Q_i(1, h) = p_{1i}(h) \oplus p_{2i}(h) \oplus --- \oplus p_{(N1)i}(h) \oplus q_{1*i}(h) \oplus q_{2*i}(h) \oplus --- \oplus q_{(gj*)i}(h), \text{ where } g_{j*} < i < g_{(j+1)*} \quad (3)$$

Severs in the r-th group (r > 1) behave in the same way as the 1st group as shown in Fig. 3 (b), where server $T_r(k)$, the k-th server in the r-th group, corresponds to $T_{k\#}$ in the anonymous channel. However, different from the 1st group where senders generate PVs and sends them as a request for a CP to severs T_{1*}, T_{2*}, ---, T_{N1*}, servers $T_{1\#}$, $T_{2\#}$, ---, $T_{Nr\#}$ in the r-th group generate CP and TV pairs spontaneously without requests from senders, also the last server $T_{Nr\#}$ in the r-th group generates group blinding vector B(h) = $\{B_1(h), B_2(h), ---, B_{Nr}(h)\}$. Then, the r-th group calculates X(r, h) and Q(r, h) = $\{Q_1(r, h), Q_2(r, h), ---, Q_N(r, h)\}$ as its h-th CP and TV values as shown in equations (4) – (6). In the equations, the j-th element $B_j(h)$ of B(h) = $\{B_1(h), B_2(h), ---, B_{Nr}(h)\}$ is a vector of patterns $\{b_{j0}(h), b_{j1}(h), ---, b_{jN}(h)\}$, where the length of $b_{j0}(h)$ is L_M and the length of $b_{jk}(h)$ is L_T for each k.

$$X(r, h) = b_{10} \oplus b_{20} \oplus --- \oplus b_{(Nr)0} \oplus x_{1\#}(h) \oplus x_{2\#}(h) \oplus --- \oplus x_{Nr\#}(h) \quad (4)$$

for g# included in the r-th group

$$Q_{g\#}(r, h) = b_{1g\#}(h) \oplus b_{2g\#}(h) \oplus --- \oplus b_{(Nr)g\#}(h) \oplus q_{1\#g\#}(h) \oplus q_{2\#g\#}(h) \oplus --- \oplus q_{(g-1)\#g\#}(h) \oplus ID_{g\#}(x_{g\#}(h), q_{g\#}(h)), \text{ where } q_{0\#g\#}(h) = 0 \quad (5)$$

for i not included in the r-th group,

$$Q_i(r, h) = b_{1i}(h) \oplus b_{2i}(h) \oplus --- \oplus b_{(Nr)i}(h) \oplus q_{1\#i}(h) \oplus q_{2\#i}(h) \oplus --- \oplus q_{(gj\#)i}(h), \text{ where } g_{j\#} < i < g_{(j+1)\#} \quad (6)$$

After calculating X(r, h) and Q(r, h) as equations (4) – (6), $T_{Nr\#}$ removes group blinding vector B(h) by XORing them by B(h). Namely, they are transformed as shown in equations (7) – (9).

$$X(r, h) = x_{1\#}(h) \oplus x_{2\#}(h) \oplus --- \oplus x_{Nr\#}(h) \quad (7)$$

for g# included in the r-th group

$$Q_{g\#}(r, h) = q_{1\#g\#}(h) \oplus q_{2\#g\#}(h) \oplus --- \oplus q_{(g-1)\#g\#}(h) \oplus ID_{g\#}(x_{g\#}(h), q_{g\#}(h)), \text{ where } q_{0\#g\#}(h) = 0 \quad (8)$$

for i not included in the r-th group,

$$Q_i(r, h) = q_{1\#i}(h) \oplus q_{2\#i}(h) \oplus --- \oplus q_{(gj\#)i}(h), \text{ where } g_{j\#} < i < g_{(j+1)\#} \quad (9)$$

The last server $T_r(N_r) = T_{Nr\#}$ in the r-th group also receives X(r+1, h) and Q(r+1, h), the CP and TV values generated by the (r+1)-th group, from $T_{r+1}(N_{r+1})$, the last server in the (r+1)-th group, and it calculates X(r, h)⊕X(r+1, h), and Q(r, h)⊕Q(r+1, h) to combine CPs and

TVs generated by the r-th and the (r+1)-th groups into the single CP and TV, respectively. Then, $T_r(N_r)$ waits for the arrivals of predefined number of CP and TV pairs, and shuffles them to sends the results to the last server $T_{r-1}(N_{r-1})$ of the (r-1)-th group. As the result of the behaviors of all groups, the last server of the 1st group, i.e. $T_1(N_1)$, generates the CP and TV as equations (10) and (11).

$$X(h) = p_{10}(h)\oplus p_{20}(h)\oplus\text{---}\oplus p_{(N1)0}(h)\oplus x_1(h)\oplus x_2(h)\oplus\text{---}\oplus x_N(h) \quad (10)$$

$$Q_g(h) = p_{1g}(h)\oplus\text{---}\oplus p_{(N1)g}(h)\oplus q_{1g}(h)\oplus\text{---}\oplus q_{(g-1)g}(h)\oplus ID_g(x_g(h), q_g(h)), \text{ where } q_{0g}(h) = 0 \quad (11)$$

Then, $T_1(N_1)$ sends X(h) and Q(h) = {$Q_1(h)$, $Q_2(h)$, ---, $Q_N(h)$} to sender S, and S removes private vectors PVs from X(h) and Q(h) by XORing them by PVs. As the result, finally CP and TV values become as (12) and (13).

$$X(h) = x_1(h)\oplus x_2(h)\oplus\text{---}\oplus x_{(N-1)}(h)\oplus x_N(h) \quad (12)$$

$$Q_g(h) = q_{1g}(h)\oplus\text{---}\oplus q_{(g-1)g}(h)\oplus ID_g(x_g(h), q_g(h)), \text{ where } q_{0g}(h) = 0 \quad (13)$$

It must be noted that because PVs and group blinding vectors are secrets of sender S and last server of each group (except the 1st group), respectively, and each server T_j does not disclose $x_j(h)$ or $q_j(h)$ to others, any server cannot know CP or TV constructors of other servers. No server can know X(h) or Q(h) either unless all servers conspire with each other.

C. Behavior of the Anonymous Channel

Fig. 4 shows the behavior of the anonymous channel. Firstly, sender S encrypts its message M_S by XORing it by concealing pattern X(h) that it had acquired from $T_1(N_1)$. S also attaches tag vector Q(h) = {$Q_1(h)$, $Q_2(h)$, ---, $Q_N(h)$} corresponding to X(h), to the message, and sends {$\underline{M_S} = x_1(h)\oplus x_2(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$, $Q_1(h)$, $Q_2(h)$, ---, $Q_N(h)$} to the 1st server T_1 in the anonymous channel. Here, $Q_1(h)$ has the form $ID_1(x_1(h), q_1(h))$.

Then, T_1 that receives {$x_1(h)\oplus x_2(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$, $Q_1(h)$, $Q_2(h)$, ---, $Q_N(h)$} retrieves CP constructor $x_1(h)$ and TV constructor $q_1(h)$ from its CP table based on $ID_1(x_1(h), q_1(h))$ in $Q_1(h)$, calculates XOR of $x_1(h)$ and $\underline{M_S}$, and $q_{1j}(h)$ and $Q_j(h)$ for each j as new values of $\underline{M_S}$ and $Q_j(h)$. Therefore, $\underline{M_S}$ and $Q_j(h)$ become $\underline{M_S} = x_1(h)\oplus(x_1(h)\oplus x_2(h)\oplus\text{---}\oplus x_N(h)\oplus M_S) = x_2(h)\oplus x_3(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$ and $Q_j(h) = q_{1j}(h)\oplus(q_{1j}(h)\oplus q_{2j}(h)\oplus\text{---}\oplus q_{(j-1)j}(h)\oplus ID_j(x_j(h), q_j(h))) = q_{2j}(h)\oplus q_{3j}(h)\oplus\text{---}\oplus q_{(j-1)j}(h)\oplus ID_j(x_j(h), q_j(h))$. After that, T_1 removes $Q_1(h)$ from the tag part, waits for the predefined number of message arrivals, and shuffles them to forward each result to server T_2.

All servers in the anonymous channel perform in the same way, i.e. each T_j converts its incoming message to {$\underline{M_S} = x_{j+1}(h)\oplus x_{j+2}(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$, $Q_{j+1}(h)$, $Q_{j+2}(h)$, ---, $Q_N(h)$}, where $Q_g(h) = q_{(j+1)g}(h)\oplus\text{---}\oplus q_{(g-1)g}(h)\oplus ID_g(x_g(h), q_g(h))$. Consequently, when T_N, the last server in the anonymous channel, completes its operations on the message, the message is converted into M_S, and T_N can deliver M_S to its receiver while extracting the address of the receiver from M_S.

The anonymous channel together with the CP generator protects identities of message senders from various threats as follows. Firstly, each server T_j transforms the message part while XORing it by CP constructor $x_j(h)$ which is not known to other servers and also T_j assigns different values as CP constructors for encrypting different messages. Therefore no one including other server T_i can identify the input and output pair of T_j that corresponding to M_S by comparing message parts of T_j's receiving and forwarding messages. For T_i, 2 input and output pairs of T_j, e. g. {$x_j(h)\oplus x_{j+1}(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$, $x_{j+1}(h_1)\oplus\text{---}\oplus x_N(h_1)\oplus M_1$} and {$x_j(h)\oplus x_{j+1}(h)\oplus\text{---}\oplus x_N(h)\oplus M_S$, $x_{j+1}(h_2)\oplus\text{---}\oplus x_N(h_2)\oplus M_2$}, have equal possibilities that they are encrypted form pairs of M_S. As a consequence, it is impossible for entities including servers to identify the sender of message M_S by tracing the message parts of messages unless all servers conspire.

Any entity cannot trace M_S by examining the tag parts of messages either. Because each T_j generates different secret TV constructors for different messages and assigns different bit patterns to individual elements {$q_{j(j+1)}(h)$, ---, $q_{jN}(h)$} in TV constructor $q_j(h)$, it is impossible for other entities to identify links between incoming messages of T_j and its outgoing messages by examining pattern transitions in individual tags made by T_j. Namely, individual tags change their forms within T_j in different ways, and entities except T_j cannot extract any relation between transitions of different tags in the tag part to identify input and output pairs of same messages.

Also, although, each server T_{j*} in the 1st group in the CP generator can know the senders of encrypted messages from their CP and TV constructors, because T_{j*} generates them at requests of the senders, when T_{j*} is placed at the earlier position of the anonymous channel, its tags disappear in the later positions, i.e. the tag parts of messages that are received by servers at later positions of the anonymous channel do not include tags of any server in the 1st group, therefore even if T_{j*} conspires with servers at the later positions, it is not possible to identify senders.

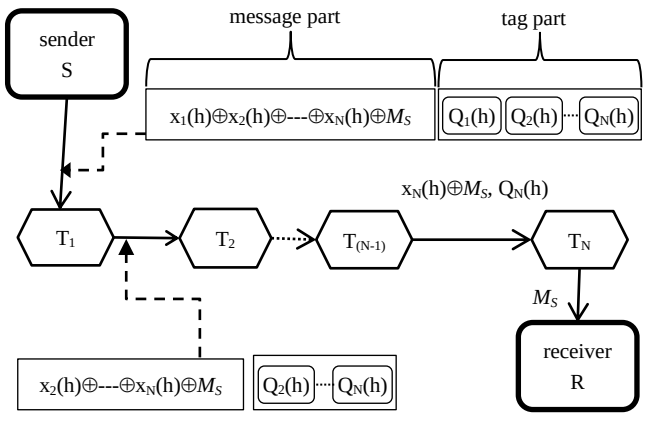


Figure 4. Behavior of the anonymous channel

V. REPLIES TO ANONYMOUS SENDERS

Different from other existing mechanisms [2, 7], in ESEBM, individual servers can handle reply messages to anonymous senders without any additional mechanism. This means that servers cannot decide even whether a message is the reply or not. Sender S can receive reply messages as follows. Firstly, S obtains 2 CP and TV pairs $\{X(h_1), Q(h_1)\}$, $\{X(h_2), Q(h_2)\}$, and constructs its message while attaching tag vector $Q(h_2)$ and its encrypted address A_S to its sending message M_S as shown in Fig. 5 (a). Namely, S constructs $M_S \| Q(h_2) \| (X_U(h_2) \oplus A_S)$, concatenation of M_S, $Q(h_2)$, and $X_U(h_2) \oplus A_S$. Where bit strings $X_U(h_2)$ and $X_L(h_2)$ are upper and lower parts of bit string $X(h_2)$, in other words, $X(h_2) = X_U(h_2) \| X_L(h_2)$. Also it is assumed that message M_S includes its destination address at its left most bit positions.

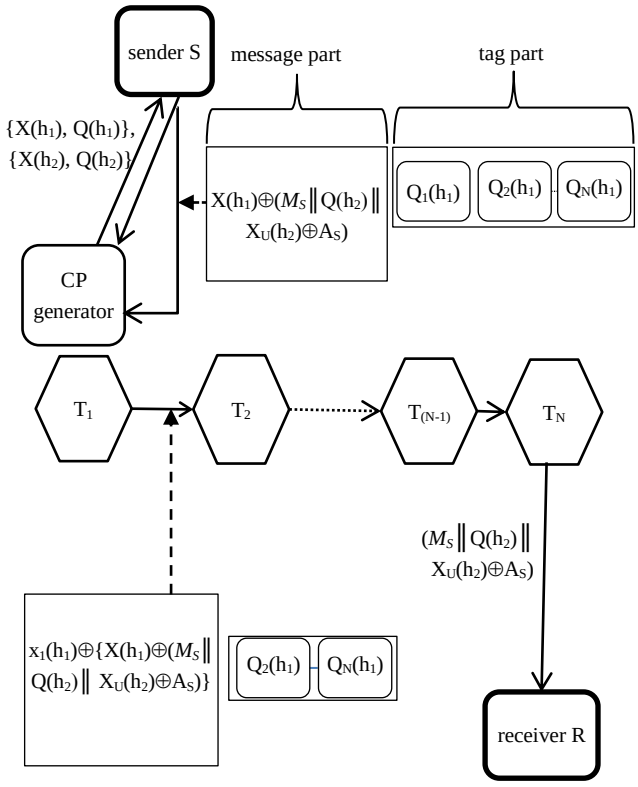
After that, S encrypts $M_S \| Q(h_2) \| X_U(h_2) \oplus A_S$ to $X(h_1) \oplus (M_S \| Q(h_2) \| X_U(h_2) \oplus A_S)$, and sends $\{X(h_1) \oplus (M_S \| Q(h_2) \| X_U(h_2) \oplus A_S)$, $Q_1(h_1)$, $Q_2(h_1)$, ---, $Q_N(h_1)\}$ to the 1st server T_1 in the anonymous channel. Then, T_1 decrypts it by $x_1(h_1)$, CP constructor of T_1. As a result, the message becomes $\{x_1(h_1) \oplus X(h_1) \oplus (M_S \| Q(h_2) \| X_U(h_2) \oplus A_S)$, $Q_2(h_1)$, ---, $Q_N(h_1)\} = \{x_2(h_1) \oplus \text{---} \oplus x_N(h_1) \oplus (M_S \| Q(h_2) \| X_U(h_2) \oplus A_S)$, $Q_2(h_1)$,---, $Q_N(h_1)\}$. Each server T_j in the anonymous channel carries out the same procedure until receiver R receives $M_S \| Q(h_2) \| X_U(h_2) \oplus A_S$. Then R can extract message M_S, encrypted address $X_U(h_2) \oplus A_S$ of S and tag vector $Q(h_2)$ to construct its reply message as $\{(X_U(h_2) \oplus A_S) \| M_R, Q_1(h_2), \text{---}, Q_N(h_2)\}$ to be encrypted to $X(h_2) \oplus \{X_U(h_2) \oplus A_S \| M_R\} = \{A_S \| X_L(h_2) \oplus M_R\}$, by the anonymous channel as shown in Fig. 5 (b). Therefore, T_N can deliver $X_L(h_2) \oplus M_R$ to S and finally S that knows $X_L(h_2)$ decrypts $X_L(h_2) \oplus M_R$ to $X_L(h_2) \oplus X_L(h_2) \oplus M_R = M_R$.

In the above, R receives $M_S \| Q(h_2) \| X_U(h_2) \oplus A_S$, and it cannot know A_S because $X_U(h_2)$ is known only to S. Also, message $X_U(h_2) \oplus A_S \| M_R$ sent by R is transformed to $A_S \| X_L(h_2) \oplus M_R$ in the anonymous channel, therefore, no one except S can know that $X_L(h_2) \oplus M_R$ corresponds to M_R, and consequently even receiver R that knows M_R cannot identify the original sender of M_S. In this way, servers in ESEBM can handle original and reply messages totally in the same way, different from usual Mixnets where each mixserver adds extra operations on reply messages.

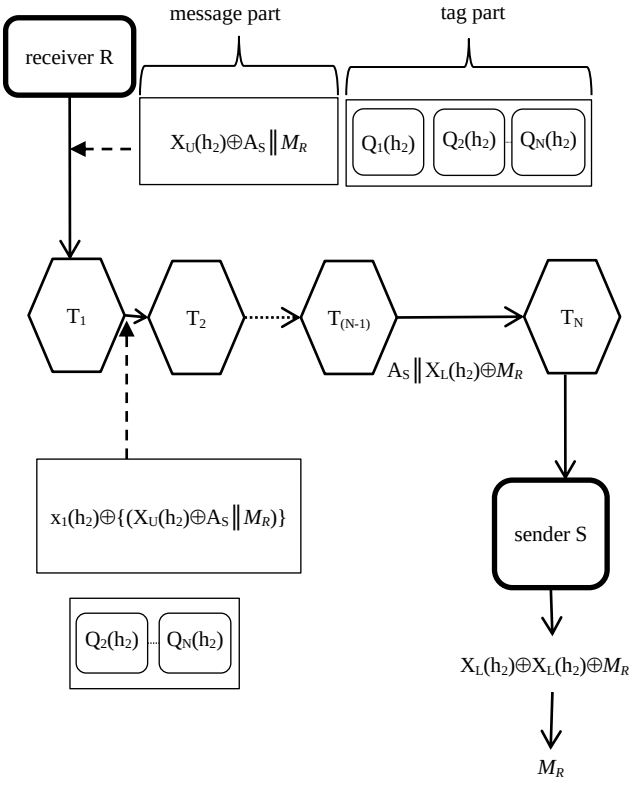
VI. EVALUATION OF ESEBM

A. Analysis of ESEBM Behavior

ESEBM satisfies the requirements for anonymous networks listed in Sec. II as follows. Firstly as discussed in Sec IV. C, no one except senders themselves can trace messages from senders to receivers. Secondly, the message reply mechanism discussed in Sec. V enables receivers to send replies to senders of original messages without knowing identities of the senders. Also by this reply mechanism, senders can confirm the deliveries of their messages. In addition the reply mechanism of ESEBM does not require additional operations on reply messages, therefore different from other existing anonymous networks, servers cannot know even whether their handling messages are replies or not.



(a) From sender to receiver



(b) From receiver to sender

Figure 5. Anonymous reply mechanism

About the efficiency, the configuration of ESEBM where senders must obtain CPs before sending their individual messages is obviously a disadvantage, e. g. message travelling times increase when durations required for obtaining CPs are counted. However because senders can obtain CPs as offline processes, actual message traveling times can be suppressed at values comparable to Mixnet. Also, when each server is configured by 2 independent CPUs, tasks for generating CPs and forwarding messages can be assigned to different CPUs so that the anonymous channel can forward messages without being interrupted by tasks for the CP generator. Then, despite of the disadvantages of the CP obtaining process, ESEBM configuration enables anonymous networks to adopt symmetric key encryption functions that make ESEBM efficient enough as usual non-anonymous networks to handle messages in practical scale applications as demonstrated in the next subsection.

ESEBM configuration brings advantageous features not only about the efficiency but also about security as follows. Among various threats to networks, DOS attacks [10], in which meaningless or spam messages are sent to decrease availabilities of networks, and illegitimate message forgeries (modifications), in which malicious entities forge (modify) messages sent from anonymous senders, are especially serious in anonymous networks. Different from in usual networks where all entities that send messages can be identified if costs and efforts are not considered, in anonymous networks where identities of senders are completely hidden, entities can behave dishonestly more easily. In addition, about message forgeries (modifications), in many cases receivers cannot notice even if their receiving messages are forged (modified) because their senders are anonymous.

The CP generator in ESEBM reduces the occurrence of DOS attacks substantially and makes forged (modified) messages detectable. Namely, senders must attach consistent TVs to their messages to let servers transfer the messages; however, the CP generator gives CPs and TVs only to authorized entities. Therefore, unauthorized entities must send their messages while attaching nonregistered TVs, and servers in ESEBM that cannot find CPs and TVs from their CP tables discard the messages immediately, as the consequence, messages from unauthorized entities do not decrease the availability of the network. About the malicious message forgeries (modification), provided that the malicious entity X does not know the original message *M*, X cannot forge (modify) encrypted *M* consistently because no one except the sender of *M* knows the CP used for encrypting *M*, then the receiver of *M* can notice the forgeries (modification) because its receiving message is meaningless.

In the same way, ESEBM disables entities to carry out traffic analysis attacks and replay attacks. A traffic analysis attack is a way to identify the sender S of a message *M* by sending multiple replies to it [7, 14]. Namely, when receiver R of *M* sends many replies at a time or periodically to S, R can identify S by observing entities that receives many messages at a time or periodically. However, in ESEBM every message must have different CPs and TVs, and this means that every server discards CP and TV constructors in its CP table once they are used. Therefore, provided that at least one of the servers is honest, even when R sends multiple replies only one of them is delivered to S, and R cannot identify S. It must be noted that, it is also possible to enable receivers to send up to predefined number of replies. If each server T_j maintains F(h), the number of messages allowed to send by using tag vector Q(h), in its CP table in addition to $\{x_j(h), q_j(h)\}$, T_j does not invalidate $\{x_j(h), q_j(h)\}$ until it receives F(h)-messages attached by Q(h).

In a replay attack [11], an entity X identifies sender S of message *M* by eavesdropping on the network to pick M_*, encrypted form of *M*, just sent from S, and putting M_* to the network repeatedly. Then, because *M* is delivered to the same receiver R many times, X can easily identify the correspondence between S and *M* received by R. Apparently ESEBM can disable replay attacks in the same way as disabling traffic analysis attacks.

B. Message Processing Performance

Performance of ESEBM has been compared with that of the usual non-anonymous networks and Mixnet each of which consisted of multiple PCs that worked as relay servers. Where individual PCs were equipped with 1.6GHz CPUs and 1GB of RAM and they were connected by 100Mbits/sec Ethernet. Because delays of message arrivals depend on the number of relay servers and the time that individual servers must wait for shuffling messages, only the throughput were compared while changing the sizes of messages. For evaluating ESEBM, 16 tags each of which consisted of 64 bits were attached to individual messages, therefore for ESEBM, the actual length of a 10 Kbits message is 11 Kbits for example. For Mixnet, RSA with 1K bits length key was adopted as the encryption function. In real applications, a sender must combine its message *M* with random secret numbers to make the encryption function probabilistic. Also to maintain strengths of encryption keys, different servers must use different modulo arithmetic. However in this evaluation, random bit strings were not attached to messages, and all servers used the same modulo arithmetic.

Table 1 shows the computation times required by each server in non-anonymous network, ESEBM and Mixnet to transfer different sizes of messages, and Fig. 6 graphically represents them. For example, while ESEBM needs less than 6 seconds to transfer a 20Mbits message, Mixnet needs more than 3 minutes to transfer the same message. Fig. 7 shows the volume of messages that usual non-anonymous networks, ESEBM and Mixnet can send within 1 second. These results show that, although the throughput of ESEBM is 1/4.4 of that of non-anonymous networks, it is more than 36 times higher than that of Mixnet. According to statistics [16], e-mail message size is 59KB on average, therefore, even in the environments used for evaluations, ESEBM can handle 7 clients at a time that send usual e-mail messages while the non-anonymous network can handle 33 clients at a time. On the other hand, Mixnet can handle only 0.2 clients. The beneficial thing is that, when multiple processors are

available, volume of messages can be processed almost in parallel. Therefore, ESEBM can transfer the same volume of messages as usual non-anonymous networks do when each server is constituted by multiple processors and memories with 4.4 times of costs. Here, although it depends on individual applications, value 4.4 can be considered acceptable. On the other hand, to improve the performance of Mixnet as non-anonymous networks, 158 times of costs are necessary. Namely, ESEBM can be used for large scale networks, in which number of clients exchange usual sizes of messages at less extra costs.

TABLE I. COMPUTATION TIME FOR TRANSFERRING DIFFERENT SIZES OF MESSAGES

Message size Mbits	Non-anonymous (msec)	ESEBM (msec)	Mixnet (msec)
10	625	2780	105255
20	1230	5510	207440
30	1924	8556	310986
40	2520	11225	412679
50	3125	14127	528276
60	3745	17342	---
70	4325	19710	---
80	4995	22862	---
90	5643	25595	---
100	6246	28344	---

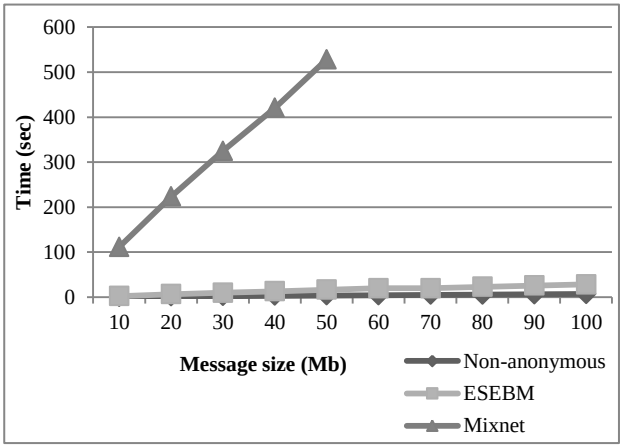


Figure 6. Comparison of computation time for transferring different sizes of messages

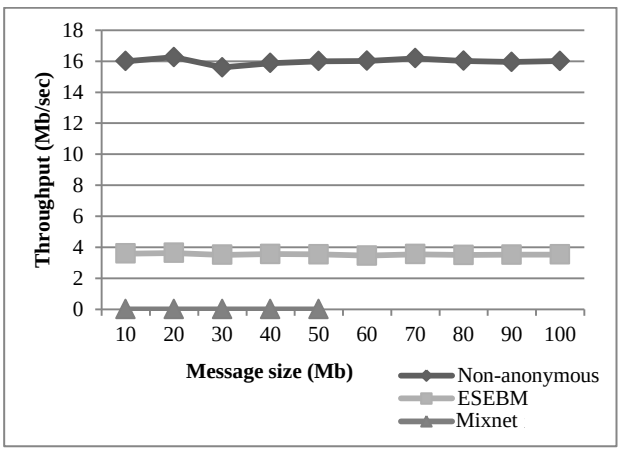


Figure 7. Comparison of throughputs for transferring different sizes of messages

About the breakdown of message processing time of each server in ESEBM, it consists of shuffling (31%), message decryption (26%), and others (43%). On the other hand, message processing time of each server in Mixnet consists of shuffling (0.8%), message decryption (98.6%), and others (0.6%). As shown above, different from Mixnet in which message decryptions require 123 times of message shuffling time, in ESEBM, message decryptions require less than 0.84 times of the shuffling time. When the fact that both ESEBM and Mixnet shuffle same number of messages is considered, this means that message decryption process in Mixnet degrades its overall performance seriously. In other words, symmetric key encryption functions used in ESEBM had successfully reduced decryption times. Namely, while RSA used in Mixnet requires the number of multiplications that is proportional to $\log_2(n)$, onetime pad used in ESEBM requires only a single XOR operation, where n is the size of encryption keys.

SEBM also uses symmetric key encryption functions [13], and as ESEBM, it can achieve the higher throughput than other anonymous networks such as Mixnet. However, when compared with ESEBM, in SEBM, more servers must be involved in forwarding messages, because it consists of encryption and decryption servers. Therefore, message traveling times in SEBM become longer than that of ESEBM, i.e. different from in ESEBM where messages are encrypted by their senders, in SEBM, they are encrypted by a sequence of encryption servers. As other advantages of ESEBM over SEBM, ESEBM works more stably because all servers in ESEBM are permanent servers different from SEBM where senders are included as servers. Also a mechanism for reply messages is not straightforward in SEBM.

VII. CONCLUSION

Enhanced symmetric key encryption based Mixnet has been proposed that removes the drawbacks of many existing anonymous networks such as Mixnet, DC-net, etc. It satisfies all the requirements of anonymous networks. Most importantly, while being supported by concealing patterns, those requirements are satisfied in a simple and efficient way. Unlike complicated Mixnet based systems, the simplified computational requirements of individual entities make the scheme practical and scalable.

As a drawback of ESEBM, a sender must acquire a concealing pattern from the CP generator in advance to send its every message as an offline process. However because of ESEBM configuration, i.e. by dividing the network into the CP generator (off-line) and the anonymous channel (on-line) parts, every time-consuming task is removed from the anonymous channel part and highly efficient communication becomes possible. Moreover, concealing patterns enable receivers not only to send replies to the original anonymous message senders but also to receive messages without disclosing their identities. Namely, when concealing patterns are publicly disclosed with the receivers' interests, the receivers can receive messages from senders without disclosing their identities.

As a future work, mechanisms that enhance the

reliability of ESEBM are necessary. When senders or receivers claim that some server is dishonest, ESEBM must prove all servers are honest or detect dishonest servers if exist. Also, ESEBM must continue its operations even some of servers are out of their services.

REFERENCES

[1] D. Chaum, "Untraceable electronic mail, return address and digital pseudonyms," *Communications of the ACM*, vol. 24, no. 2, pp. 84-88, 1981.

[2] D. Chaum, "The dining cryptographers problem: unconditional sender and recipient untraceability," *Journal of Cryptology*, vol. 1, pp. 65-75, 1988.

[3] M. G. Reed, P. F. Syverson and D. M. Goldschlag, "Anonymous connections and onion routing," *Selected Areas in Communications*, vol. 16, no. 4, pp. 482-494, May 1998.

[4] M. K. Reiter and A. D. Rubin, "Crowds: anonymity for Web transactions," *ACM Transactions on Information and System Security*, vol. 1, no. 1, pp. 66-92, Nov 1998.

[5] R. Ingledine, M. J. Freedman, D. Hopwood and D. Molnar, "A reputation system to increase MIX-Net reliability," *Proc. of the 4th international Workshop on information Hiding. I. S. Moskowitz, Ed. Lecture Notes In Computer Science, Springer-Verlag*, vol. 2137, London, pp. 126-141, April 2001.

[6] A. Beimel and S. Dolev, "Buses for anonymous message delivery," *Proc. of the Second International Conference on FUN with Algorithms*, Elba, Italy, pp. 1-13, May 2001.

[7] P. Golle and M. Jakobsson, "Reusable anonymous return channels," *Proc. of the 2003 ACM Workshop on Privacy in the Electronic Society*, (Washington, DC), WPES '03, ACM, New York, NY, pp. 94-100, 2003.

[8] R. Dingledine and N. Mathewson, "Tor: The second-generation onion router," *Proc. of the 13th USENIX Security Symposium*, San Diego, CA, USA, pp. 303-320, August 2004.

[9] P. Golle, M. Jakobsson, A. Juels and P. Syverson, "Universal re-encryption for Mixnets," *RSA Conference Cryptographers' Track '04, Springer-Verlag*, pp. 163-178, 2004.

[10] T. Znati, J. Amadei, D. R. Pazehoski and S. Sweeny, "On the design and performance of an adaptive, global Strategy for detecting and mitigating distributed DOS attacks in GRID and collaborative workflow environments," *Simulation*, vol. 83, pp. 291-303, March 2007.

[11] S. Y. Kang, J. S. Park and I. Y. Lee, "A study on authentication protocol in offering identification synchronization and position detection in RFID system," *Proc. of The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*, pp. 150-154, 2007.

[12] X. Wang and J. Luo, "A collaboration scheme for making peer-to-peer anonymous routing resilient," *Computer Supported Cooperative Work in Design, 2008, CSCWD 2008*, pp. 70-75, April 2008.

[13] S. Tamura, K. Kouro, M. Sasatani, K. M. Alam and H. A. Haddad, "An information system platform for anonymous product recycling," *Journal of Software*, vol. 3, no. 6, pp. 46-56, 2008.

[14] L. Li, S. Fu and X. Che, "Active attacks on reputable Mix Networks," *ispa, 2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pp. 447-450, 2009.

[15] H. Haddad, H. Tsurugi and S. Tamura, "A mechanism for enhanced symmetric key encryption based Mixnet," *SMC 2009 IEEE International Conference on Systems, Man and Cybernetics*, San Antonio, TX, USA, pp. 4541–4546, 11-14 Oct 2009, doi: 10.1109/ ICSMC.2009.5346788.

[16] http://www.idc.com/

Hazim A. Haddad received the B.E. degree in Computer science and Engineering from Ittihad University, from UAE (United Arab Emirates) in 2003, M.S. degree in nuclear and safety engineering, from the University of Fukui in 2008. He is currently a doctor course student of University of Fukui.

Shinsuke Tamura was born in Hyogo, Japan on Jan. 16, 1948, and received the B.S., M.S. and Dr. (Eng.) degrees in Control Engineering from Osaka University in 1970, 1972 and 1991, respectively. During 1972 to 2001, he worked for Toshiba Corporation. He is currently a professor of Graduate School of Engineering, University of Fukui, Japan. Prof. Tamura is a member of IEEJ, SICE and JSST.

Shuji Taniguchi received the B.E. and Ph.D. degrees in electronics engineering from University of Fukui, Fukui, Japan, in 1973, 1996, respectively. In 1973-1978, he was with the Hitachi co. Ltd. He is currently an associate professor of Graduate School of Engineering in University of Fukui.

Tatsuro Yanase received the Dr. (Eng.) degrees in Electric & Electronic Engineering from Nagoya University in 1977. During 1967 to 1969 he worked for Nippon Calculating Machine Corporation. He is now an associate professor of Graduate School of Engineering, University of Fukui.

Development of a Ubiquitous Industrial Data Acquisition System for Rotogravure Printing Press

Yuhuang Zheng
Department of Physics, Guangdong University of Education, Guangzhou, China
Email: zhyhaa@126.com

***Abstract*—This paper describes a data acquisition system developed to solute the problem of different data acquisition modules communication in ubiquitous industrial environment. The system can allow faster reconfiguration of plant-floor networks as applications change. It can achieve higher throughput, lower average message delay and less average message dropping rate in wireless communication. The development of a data acquisition system for rotogravure printing press in ubiquitous industrial environment also is reported. It illustrates that the system can perform well in industrial application.**

***Index Terms*—Wireless networks, Ubiquitous Computing, Data Acquisition Module (DAM)**

I. INTRODUCTION

Wireless networks have been under rapid development during recent years. Types of technologies being developed to wireless personal area network for short range, point-to multi-point communications, such as Bluetooth and ZigBee [1]. The application of wireless technology for industrial communication and control systems has the potential to provide major benefits in terms of flexible installation and maintenance of field devices and reduction in costs and problems due to wire cabling [2].

Wireless communications from machine to machine greatly enhance automation of an industrial system. Ubiquitous industrial environment is coming and allows the engineers to acquire and control the real-time data of wireless networks of the factory at anytime anywhere [3].

A key issue currently limiting ubiquitous industrial environment development involves compatibility among components in industrial environment from different suppliers, generally referred to as interoperability. Full compatibility among components would also provide end users with the flexibility to connect highly specialized, high-end sensors with best-in-class wireless interface devices [4].

Interoperability in ubiquitous industrial environment means wireless communication protocol and the protocol of monitoring and controlling industrial equipments are interoperable. Interoperable wireless protocols are making or have appeared by some international organizations and alliances, such as ISO, WINA, ZigBee, etc. Most industrial equipments have their special monitoring and controlling protocols. Data Acquisition Module (DAM) is the most important equipment in industrial application, but different brands almost have different inherent monitoring protocols. For example, Advantech ADAM 4000 series support both ADAM ASCII and MODBUS protocols. But different ADAM 4000 Modules have different command sets. It is common for a factory to using different kinds of DAMs in their product lines. How to make these different DAMs can communicate with each other is a key problem.

This paper addresses the MPCS protocol to solute this problem. MPCS is the abbreviation of MODBUS for Producer/Consumer Services. MPCS protocol applies ZigBee as wireless protocol among wireless nodes. The core of MPCS is to use the MODBUS protocol without polling to carry industrial equipments protocol. MODBUS protocol is applied to an electronic controller on the lingua franca. The most important is the protocol also must be supported by typical DAMs. But most of the industrial monitoring systems adopt fixed period polling with less consideration about dynamic period in using MODBUS. So in ubiquitous industrial environment, the MODBUS protocol cannot satisfy the latency requirement of wireless protocol and it cannot guarantee the real-time monitoring of industrial environment conditions. And the polling method of MODBUS adds extra loads and burdens the wireless channel. The MPCS protocol changes the polling mechanism of MODBUS and the slave equipments can send the messages by themselves periodically without receiving query command from the master equipment.

MPCS protocol applies ZigBee as wireless protocol among wireless nodes and MODBUS with sending message periodically as industrial monitoring protocol. Experiment shows that the combination of MPCS and ZigBee is a good way to solute the interoperability in ubiquitous industrial environment. MPCS has the advantages in saving bandwidth and lightening servers'

Manuscript received January 1, 2011; revised June 1, 2011; accepted July 1, 2011.

doi:10.4304/jnw.6.11.1543-1548

load and enhances the real-time performance of industrial wireless sensor networks.

This paper first introduces the data acquisition system for rotogravure printing press in ubiquitous industrial environment. Secondly ZigBee and ZigBee gateway design are discussed. Thirdly it introduces MPCS protocol in different DAMs. There is an experiment to test MPCS protocol on rotogravure printing press monitoring system in this section. Finally it is a conclusion.

II. System Overview

Wireless sensors provide the network with the ability to reconfigure on the fly without being tied down by signal cables. The goal of the system is to implement such a network using DAMs connected by Zigbee transceivers to a central computer that interfaces with a database accessible. The three major components consist of different kinds of DAMs, CC2530 Low Power Transceivers for ZigBee, and the SCADA software hosted on the central computer. A block diagram of the high level design is provided in Fig. 1 below.

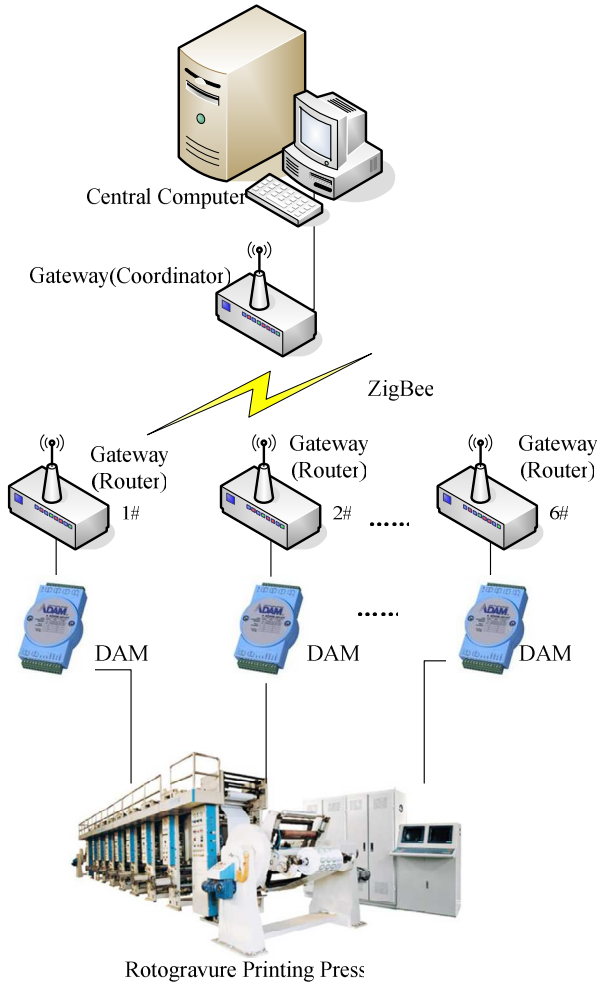
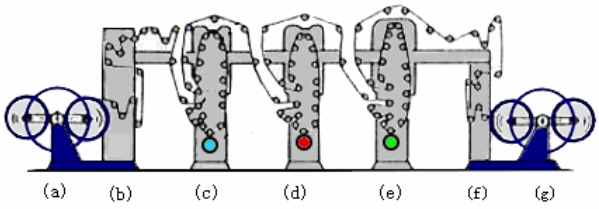


Figure 1. System Architecture of System

There are DAMs in this system, including three modules of ADAM-4011(thermocouple input module) and three modules of ADAM-4012(analog input module). ADAM-4011 and ADAM-4012 use MODBUS protocol. One ADAM-4011 module records dryer air temperature of a color unit. One ADAM-4012 module logs tension of a color unit. All DAMs collect data from the industrial machinery and transmit them in MODBUS format to ZigBee gateway in which data are processed. ZigBee gateway packs data according to the ZigBee protocol, and transmits them to via radio. Finally, data are transmitted to ZigBee gateway of the central computer. At the central computer, incoming data from ZigBee gateway of it are received and processed by the SCADA software which is developed in Kingview [5].

Rotogravure printing press consists of paper web unwinder, infeeding unit, rotogravure printing units, outfeeding unit, and paper web rewinder. In the rotogravure printing process, a web from a continuous roll is passed over the image surface of a revolving gravure cylinder.

The printing images are formed by many tiny recesses engraved into the surface of the gravure cylinder. The cylinder is about one-fourth submerged in a fountain of low- viscosity mixed ink. The mixed ink is picked up by the cells on the revolving cylinder surface and is continuously applied to the paper web. After impression is made, the web travels through an enclosed heated air dryer to evaporate the volatile solvent. The web is then guided along a series of rollers to the next printing unit [8, 9]. Fig. 2 shows the structure of rotogravure printing press in out application.



(a) Paper web Unwinder
(b) Infeeding unit
(c) Blue color rotogravure printing unit
(d) Red color rotogravure printing unit
(e) Green color rotogravure printing unit
(f) Outfeeding unit
(g) Paper web Rewinder

Figure 2. Structure of Rotogravure Printing Press

III. ZigBee

Wireless sensors provide the network with the ability to reconfigure on the fly without being tied down by signal cables. The goal of the system is to implement such a network using DAMs connected by Zigbee transceivers to a central computer that interfaces with a database accessible. The three major components consist of different kinds of DAMs, CC2530 Low Power Transceivers for ZigBee, and the SCADA software hosted on the central computer. A block diagram of the high level design is provided in Fig. 1.

A. ZigBee in Industrial Environment

There are thousands of devices in a factory, such as, DAM, HMI, IPC, smart sensor, and so on. ZigBee is focused on control and automation. ZigBee standards

have a characteristic of "three low" of low electricity consumption (year's cell life), low cost (less than $5) and low data rate (250 Kb/s) [6-7]. ZigBee works with small packet devices and supports a larger number of devices and a longer range between devices than other technologies. ZigBee devices can form mesh networks connecting hundreds to thousands of devices together. Devices use very little power and can operate on a cell battery for many years. In timing critical applications, such as industrial application, ZigBee is designed to respond quickly. ZigBee is a good wireless technology in industrial application.

B. ZigBee Gateway Designation

The ZigBee gateway is based on the CC2530 System-on-Chip, which combines a RF transceiver with an industry-standard enhanced 8051 MCU, in-system programmable flash memory, 8-KB RAM, and other powerful peripherals. The gateway which connects the central computer operates as the coordinator and the

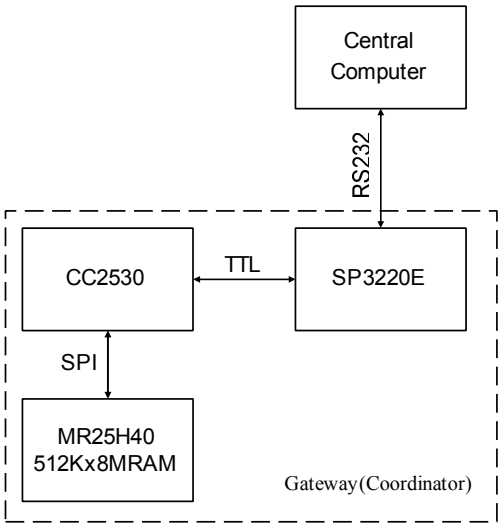


Figure 3. A Coordinator Gateway

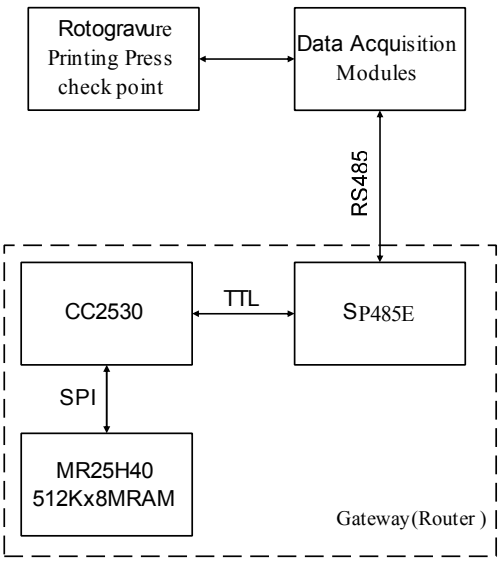


Figure 4. A Router Gateway

gateway that connects a DAM runs as the router in ZigBee network. Because of large acquired data, these gateways includes 512KB RAM as the data buffer. Fig. 3 is the architecture of coordinator gateway. This gateway consists of CC2530, SP3220E devices, external RAM and some interfaces. Fig. 4 is the architecture of router gateway. This gateway consists of CC2530, SP485E devices, external RAM.

Because a computer has a UART interface and the interface usually is RS-232. But UART of CC2530 is TTL, so RS232-TTL conversion is done with a SP3220E chip. And a DAM has a UART interface which is RS-485, so RS485-TTL conversion is done with a SP485E chip.

ZigBee gateway allows device containing UART to communicate via radio with other devices. Each device connects to ZigBee gateway. In this system, six ZigBee gateways provide the radio communication link.

When the ZigBee gateway is used in an application, it is assumed that a permanent power source will be available at both ends of the wireless link. This means the on-chip radio can always be active, eliminating the need to synchronize the transmission/reception of data. The link is designed to operate at up to 19200 baud.

The ZigBee gateway of PC must act as a PAN Coordinator. The PAN Coordinator is responsible for starting the network and allocating an address to the other gateway, which acts as a Router. Fig. 5 is the program flow chart of PAN Coordinator.

The ZigBee gateway of DAM acts as a router device. The router device scans the radio channels, looking for the PAN Coordinator. Once it has found the Coordinator, it associates with it. Data transfer between radio and the on-chip UART is identical to that described above. Data received via the radio is output to the connected device using the on-chip UART, and data received by the on-chip UART from the device is transmitted over the radio. This process is repeated every 20ms. Fig. 6 is the program flow chart of Router.

IV. COMMUNICATION PROTOCOL

The communication protocol between the central computer and DAMs is MPCS. MPCS is important to have a protocol at the application layer that allows DAMs to take advantage of producer/consumer services. Using producer/consumer, the data "producer" which is a DAM, puts the PAN Coordinator ID at the front of each packet. The message is then sent out of the network and the Coordinator screens the ID to determine if the data is for its consumption. If so, the Coordinator becomes the "consumer." As a result, multi-cast communication happens naturally and efficiently in a producer/consumer service.

MODBUS is designed for source/destination communication or master/slave model. MPCS, however, joins forces with producer/consumer technology to offer a superior message-delivery mechanism of MODBUS. MPCS supports all of the major communication relationships of MODBUS. MPCS is a flexible protocol and results in efficient use of bandwidth.

MODBUS-compliant devices are common in industrial application. Users can achieve MODBUS by device's

internal standard, direct interface and external converter. Internal standard means MODBUS is the basic protocol that DAM uses for parsing messages, such as Link-Max and Advantech DAMs. Some DAMs have their internal standard protocols, but these DAMs also provide MODBUS communication interface and MODBUS communication instruction. For example, the Advantech DAM family supports RS-485, direct interface to MODBUS master-slave networks without an external electrical interface converter. These DAMs are internal support MODBUS.

In this data acquisition system for rotogravure printing press, MODBUS is internal standard of Advantech DAM. To verify the performance and interoperability of MPCS protocol, we do a test of the system in this section. The monitoring period of control center is 1 second. The DAMs are distributed among 30 meters distance. Table 1 shows the result of the test. Six DAMs send 2000 packets relatively and the central computer records these packets. Basically, the test results are satisfied, and the MPCS protocol are suitable in non-critical industrial environment [10-13].

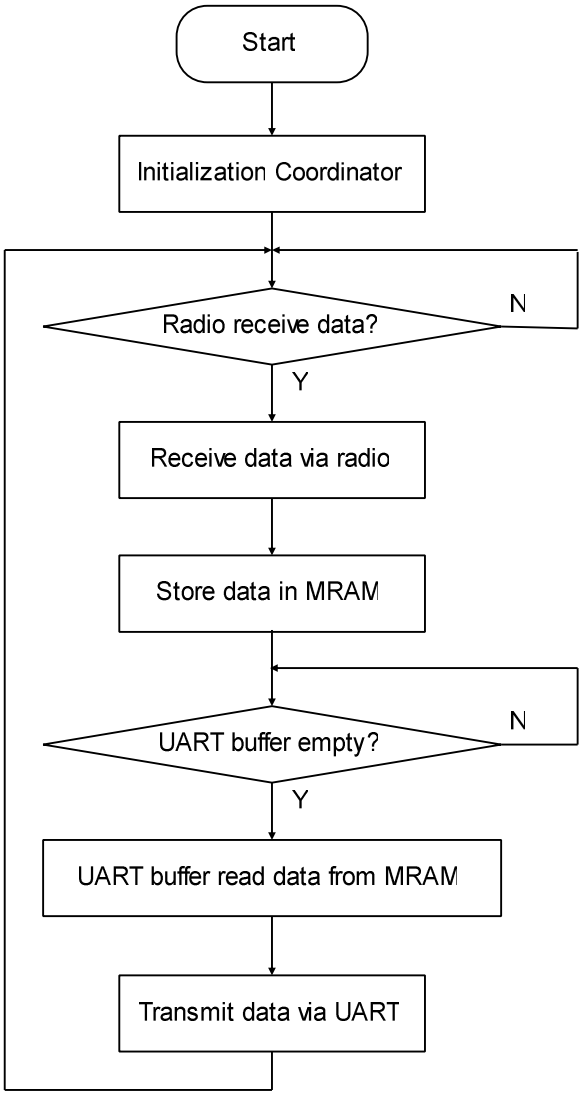


Figure 5. Coordinator Program

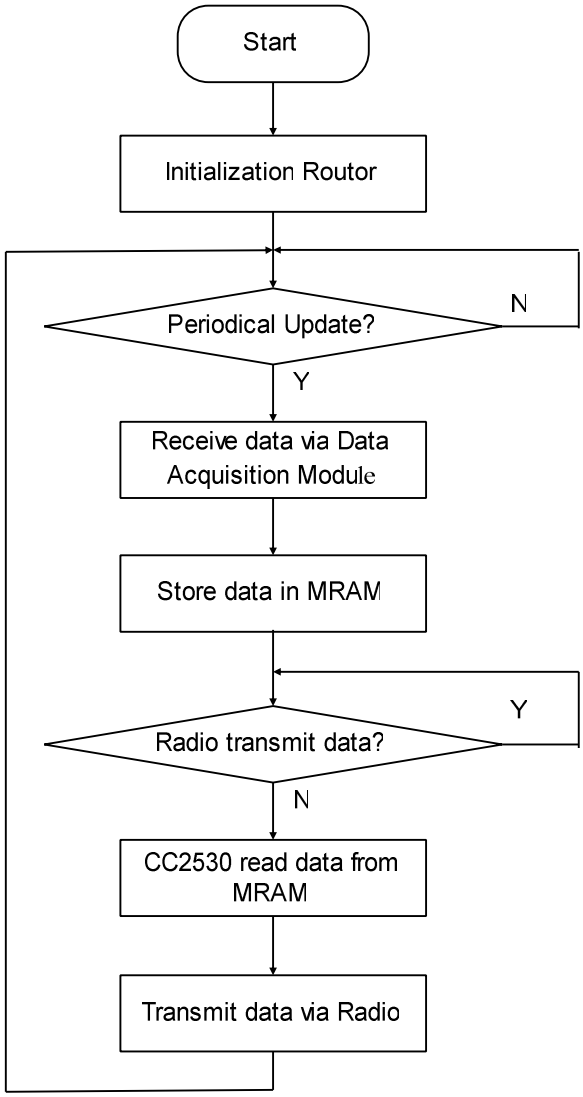


Figure 6. Router Program

TABLE I.
TEST RESULT

Data Acquisition Modules	Transmit Packages	Receive Packages	Successful Rate
1#	2000	1893	94.65%
2#	2000	1972	98.60%
3#	2000	1937	96.85%
4#	2000	1963	98.15%
5#	2000	1952	97.60%
6#	2000	1918	95.90%

V. SUPERVISORY SOFTWARE DESIGN

Supervisory software is built on Kingview. Kingview is a kind of HMI/SCADA software with abundant functions. Kingview provides integrated development

environment. The function of the software consists of system administration module, database and extension module, which are shown in Fig. 7.

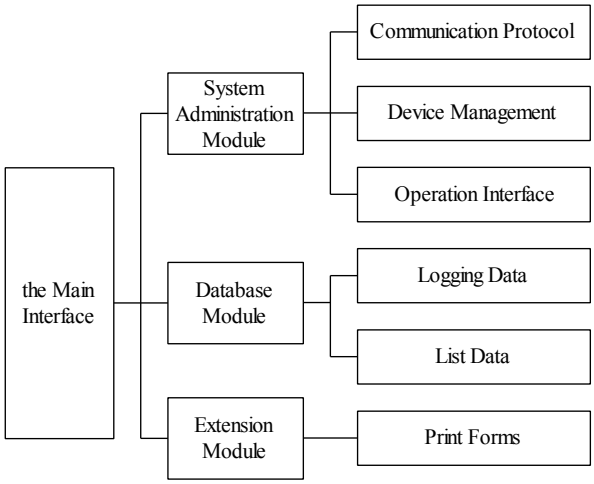


Figure 7. Temperature of Dryer Air

A. System Administration Module

System administration module consists of administrator logging in, password setting, user administrating. The part of administrator logging in is to restrict the operation of users. The part of code setting is to provide all the users to modify their own code. The part of which can only be operated by administrator is to provide the detail information of users and give or repeal the administration right to some users. According to requirement analysis, the system should divide all the users to three parts, and allow them to have different operations. The administrator can administrate all the users; browse all pictures and data in supervisory software. They are the top tier. Engineers have similar power of administrator except user administrating, so they are the middle tier. Operators can only browse given pictures on the industrial computer and they cannot get the running data of rotogravure printing press. Fig. 8 is the main program interface of this data acquisition system on central computer [14].

B. Database Module

The database module is the core of the whole system, in which the rule searching is carried out. Authorized users can read real time data in dynamic report forms. Meanwhile data will be recorded in database. Nobody can modify information in the database. Authorized users can browse all history data in static report forms. All report forms can be printed. Fig. 9 shows some temperature data of dryer air and tension data recorded by this system [15-18].

C. Extension Module

The extension module makes the query result of database can be displayed graphically in curves. User also can print the curves and the query result. Fig. 10 is the curves of some temperature data about dryer air.

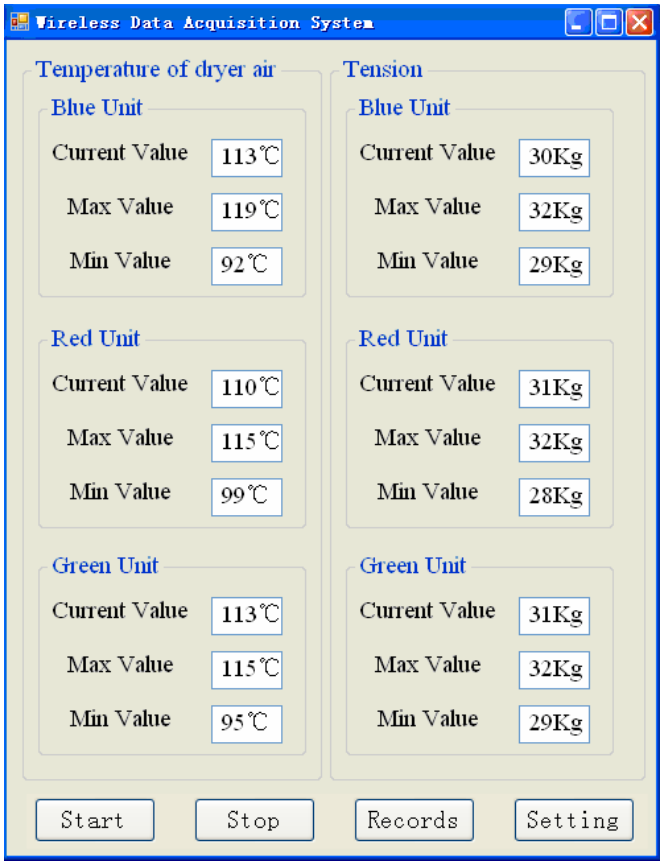


Figure 8. Program interface

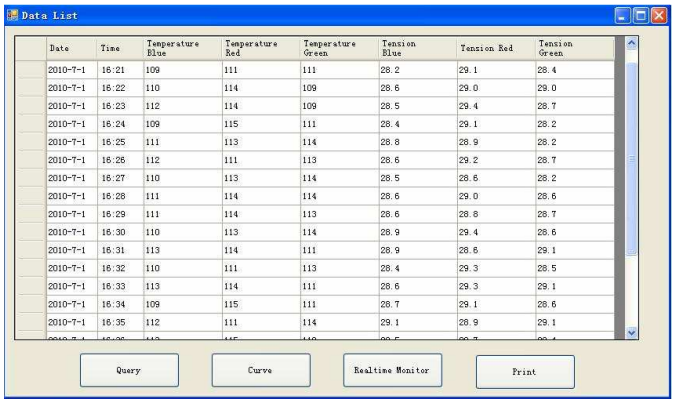


Figure 9. Temperature and Tension Data

VI. Conclusion

The design and implementation of a ubiquitous data acquisition system for rotogravure printing press is discussed and presented for industrial monitoring system. Tests are carried out to determine system performance for both the instrumentation and maintenance applications, and as the results are quite satisfactory. The results show the performance and interoperability for the wireless data acquisition system is good enough for some monitoring and non-critical instrument systems.

Further efforts are necessary to improve reliability of sensor nodes, security, and standardization of interfaces

and interoperability. In addition, further studies are necessary to improve the protocol's functionality by checking the impact of the mobility of sensor nodes.

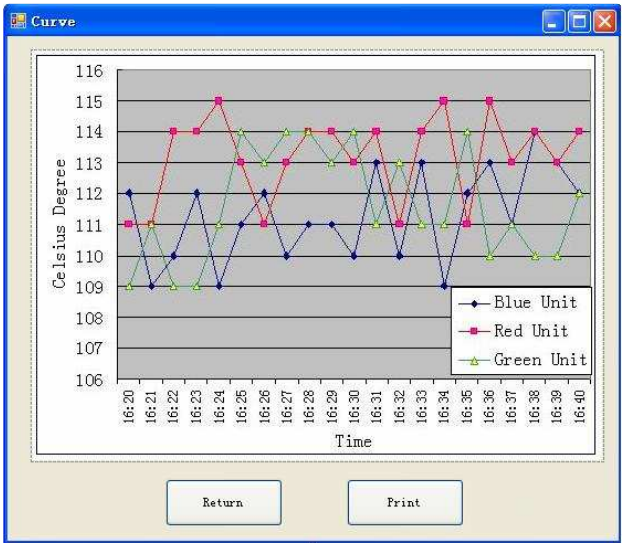


Figure 10. Temperature Curves of Dryer Air

REFERENCES

[1] W. Ning , Z. Naiqian, and W. Maohua, "Wireless sensors in agriculture and food industry - Recent development and future perspective", *Computers and Electronics in Agriculture*, vol. 50, pp. 1-14, January 2006

[2] B. Alvise, C. Luca, and S.V. Alberto, "Platform-based design of wireless sensor networks for industrial applications", *International Conference on Design, Automation and Test in Europe*. Munich, vol. 1, pp.4-10 March 2006

[3] L. K. Soon, W. W. N. Nu and E. M. Joo, "Wireless sensor networks for industrial environments", *International Conference on Computational Intelligence for Modeling, Control and Automation.* Vienna, vol. 2, pp.271-276, November 2005

[4] Industrial Wireless Technology for the 21st Century, www1.eere.energy.gov/industry/sensors_automation/pdfs/wireless_technology.pdf

[5] X. Xueliang, T. Cheng and F. Xingyuan, "A Health Care System Based on PLC and ZigBee", *International Conference on Wireless Communications, Networking and Mobile Computing*. New York, pp.3063-3066, October 2007

[6] L. Zheng, "ZigBee wireless sensor network in industrial applications", *International Joint Conference*. Korea, pp. 1067-1070, October 2006

[7] G. Vehbi and H. Gerhard, "Industrial wireless sensor networks: Challenges, design principles, and technical approaches", *IEEE Trans. on Industrial Electronics*, vol. 56, pp.4258-4265, October 2009

[8] Information on: www.epa.gov/ttn/chief/ap42/ch04/final/c4s09-2.pdf

[9] Information on:bbs.ca800.com/html/UploadFile/bbs/20080628/20080628161526240.doc

[10] S. Sooyeon, K. Taekyoung, J. G. Yong, P. Youngman, and R. Haekyu, "An experimental study of hierarchical intrusion detection for wireless industrial sensor networks", *IEEE Trans. on Industrial Informatics*, vol. 6, pp.744-757, November 2010

[11] G. Sumeet, V. Shekhar and A. R. Kumar. "Intelligent industrial data acquisition and energy monitoring using wireless sensor networks", *International Journal of Grid and High Performance Computing*, vol. 2 pp.44-59, July 2010

[12] C. Jiming, C. Xianghui, C. Peng, X. Yang and S. Youxian. "Distributed collaborative control for industrial automation with wireless sensor and actuator networks", *IEEE Trans. on Industrial Electronics*, vol. 57, pp.4219-4230, December 2010

[13] U. Alphan, G. Ozgur, and O. Ahmet, "Wireless model-based predictive networked control system over cooperative wireless network", *IEEE Trans. on Industrial Informatics*, vol. 7, pp.41-51, February 2011

[14] G. M. Coates, K. M. Hopkinson, S.R. Graham and S. H. Kurkowski, "A trust system architecture for SCADA network security", *IEEE Trans. on Power Delivery*, vol. 25, pp.158-169, January 2010

[15] H.T. Sánchez, P. Sánchez and M. Estrems, "SCADA system improvement in PVC high frequency plastic welding", International *Journal of Advanced Manufacturing Technology*, vol. 40, pp 84-94, January 2009

[16] G. N. Korres and N. M. Manousakis, "State estimation and bad data processing for systems including PMU and SCADA measurements", *Electric Power Systems Research*, vol. 81, pp.1514-1524, July 2011

[17] A.Kusiak and Z. Zhang, "Analysis of wind turbine vibrations based on SCADA data", *Journal of Solar Energy Engineering*, vol. 132, pp.0310081-03100812, August 2010

[18] H. Lee, D.H. Yoon, G. Jang and J.K. Park, "Study of the design of data acquisition and analysis systems for multi-purpose regional energy systems", *Journal of Electrical Engineering and Technology*, vol. 5, pp.16-20, March 2010

Yuhuang Zheng received his B.S. and M.S. degree from the Faculty of Automation, Guangdong University of Technology, Guangzhou, China in 2002 and 2006 respectively. In 2009, he received his Ph.D. from School of Mechanical & Automotive Engineering, South China University of Technology. His main interests are industrial automation, embedded system design, and pervasive computing. He is a lecturer at the Dept. of Physics, Guangdong University of Education.

An Application of the Modification of Slow Start Algorithm in Campus Network

Guo-hong Gao
School of Information Engineer, Henan Institute of Science and Technology, Henan Xinxiang, 453003, CHINA
Sanyuerj03@126.com

Wen-xian Xiao, Zhen Liu
Network Center, Henan Institute of Science and Technology , Henan Xinxiang, 453003, CHINA
Xwenx@yeah.net

Wen-long Wan
Foreign language department,Henan Institute of Science and Technology,Henan Xinxiang, 453003,China
e-mail:wanwenlong@hist.edu.cn

Abstract—**For the problems existing in TCP/IP congestion control method, a modified slow start algorithm is introduced to the internal services of campus network to make congestion control method more effective. Network simulation results show that: the modified congestion control algorithm can effectively improve the network transmission efficiency in a specific network environment.**

***Index Terms*—congestion control;slow start;campus network; simulation**

I. INTRODUCTION

Congestion control is one of the key factors for ensuring the stability and robustness of computer network. With the expansion of network size, the continuous increase of network bandwidth and the increasing diversification of networking forms, congestion control has encountered some new problems needing solving. When the packet number which reaches the network is greater than the processing capacity of network, network performance would drop dramatically, and then the congestion will inevitably happen. In order to avoid congestion, people use congestion control algorithm in the network to make it work properly. TCP congestion control algorithms include such four basic algorithms as slow start, congestion avoidance, fast retransmit and fast recovery. Slow start and congestion avoidance algorithm are the methods that the TCP data sender must follow while sending windows[1,2].

In RFC2581 and RFC2001, the slow start algorithm of TCP was described, and its pseudo code is described as follows (in which, swin is the sender window, awin is notification window for the recipient) :

```
swin = min (cwnd, awin)
cwnd = 1;
ssthresh = 65535bytes;
When the new data packet confirms that the ACK arrives:
If (cwnd> = ssthresh)
/ Congestion avoidance /
cwnd = cwnd +1 / cwnd;
else
/ Slow start /
cwnd = cwnd +1;
Timeout:
ssthresh = max (2, min (cwnd / 2, awin));
  cwnd = 1;
Re-enter the phase of slow start.
```

From the above description of congestion control, algorithm has trouble in efficiency .Its way of using the progressive increase to find out the right sending bandwidth makes TCP connections unable to fully use the available network bandwidth at the beginning of connections[2,3,4]. For the connections which have small transmission data and strong paroxysm(such as Web streaming), the whole connection may have always been in the slow-start state with a small sending window, thus making the network transmission efficiency lower. In addition ,in the process of slow start, whether the initial window and slow start threshold selected is appropriate also directly affects the transmission performance of the network. A continual loss in the connection process will cause that the slow-start threshold value decreases rapidly; the system is under the control of small congestion window with slow growth for a long time, especially when the loss occurred in the initial window, the slow start threshold value will be reduced to the size of two data segments, the connection with the small amount of data sent may have always been unable to get the right bandwidth[5].

The packet loss is taken as a basis to judge congestion in prevailing TCP congestion control .This approach is successful in cable transmission network. The reliability of network transmission is relatively low; the transmission signals are susceptible to be interfered by external factors. Packet loss is not necessarily caused by network congestion. The network congestion detection mechanism has also been challenged. In this case, the use of traditional TCP algorithm with progressive increase and reducing in times of multiplication may result in the decrease in the utilization of network resources, while, in the end to end TCP congestion control strategies, the judge of congestion is determined by the feedback

doi:10.4304/jnw.6.11.1549-1556

information at the receiving end, the data sender doesn't get information about network congestion until congestion occurs for at least 1/2 RTT, the program that the sender slows down the sending speed to avoid network congestion is implemented after congestion occurs for a period of time , and most of the data packets sent in the period from the occur of network congestion to the relieve of the network congestion may be discarded. If the data sender can detect congestion in a timely manner, network congestion control mechanism would be more effective, the utilization of network resource will be higher, and the costs for data transmission will be smaller. Timely detection of network congestion, is also a problem that TCP congestion control strategies should solve[6,7,8].

From the description of the slow start, the way using the progressive increase to find out the right sending bandwidth makes TCP connections unable to fully use the available network bandwidth at the beginning of connections. For the connections which have small transmission data and strong paroxysm, the whole connection may have always been in the slow-start state with a small sending window, thus making the network transmission efficiency lower. If continual packet losses appear in the process of connection, the value of the threshold of slow start will sharply reduce[9]. The system will be under the control of small congestion window with slow growth for a long time, so it may have always been unable to get the right bandwidth. In connection with the deficiency above, the modified strategies have been put forward in the paper[10].

II. ALGORITHM MODIFICATION

The purpose of congestion control is to ensure the security, stability and efficiency of network operation, but because of the complexity of network congestion control, there are some difficulties to ensure all the above objectives simultaneously. For the network using TCP/IP protocol, the congestion control is achieved by TCP congestion control and IP layer congestion control; TCP's congestion control algorithm runs on the nodes, whose algorithm complexity has little influence on the overall network efficiency; for the IP layer congestion control algorithm, its congestion control is implemented by core router. In order to ensure network efficiency, IP layer congestion control algorithm must be efficient, simple. In the IP layer congestion control algorithm, the time to determine whether a packet should be discarded must be much smaller than the time that the packet is sent, otherwise, the algorithm has no practical application value[11,12].

LOW etc proposed TCP/AQM dual model based on the Theory of Optimization. It regards the existing TCP congestion control and AQM algorithms in the network layer as the solution of the distributed algorithm with the appropriate utility function and the optimal rate allocation problem; hence it can theoretically analyze network performance in equilibrium, such as packet loss rate, throughput, efficiency, delay and queue length. In the whole congestion control mechanism, the rate of the sending end and the congestion extent of router influence each other. According to the congestion extent the router feeds back, the sending end adjusts its sending rate; on the other hand, the transmission rate of every sending end would affect the degree of network congestion, thus creating a closed loop congestion control system. The main idea of the model is to think of the transmission rate as the original variable, the congestion metric as the dual variables[13]. The existing TCP/AQM algorithm can be seen as the largest Lagrangian method which makes the utilization rate of total resource largest, as is the basic method of optimization theory.Assuming the network contains L links, which are shared by S roots.

When the root S sends at the rate of x_s, its utility function value is $U_s(x_s)$. Assuming that U_s is an increasing function which is convex, continuous and differentiable ,Link L adjusts the amount of the degree of congestion according to stimulated roots passing through it, while root S adjusts the sending rate according to the amount of congestion extent of link in its transmission path. It is written as $x(t) = (x_s(t), s \in S)$, $p(t) = (p_l(t), l \in L)$, and then

$$x(t+1) = F(x(t), p(t))$$
$$p(t+1) = G(x(t), p(t))$$

(Equation 2.1)

Function F is sender's congestion control algorithm, such as TCP Vegas, TCP Reno, etc.; function G as a router queue management, such as RED. This way, TCP / AQM strategies can be described with a triple function (F, G, U). For convenience, when the link with the capacity of C is shared by S roots, the goal whose transmission rate (i.e. the original variables) is the congestion control is selected to make

$$\max_{x_s \geq 0} \sum_S U_s(x_s)$$
$$s.t. \sum_S x_s \leq C$$

(Equation 2.2)

Equation (2.2) becomes the original problem, and the corresponding dual problem as the amount of congestion extent selected (i.e., dual variables), so that

$$\min_{p \geq 0} D(p) = \sum_S \max_{x_s \geq 0} U_s(x_s) - x_s p + pc$$

(Equation 2.3)

Based on Kuhn-Tucker theorem in Optimization Theory, the existence of non-negative value (x^*, p^*) makes x^* become the solution of the original problem, p^* as the solution to the dual problem. Different mechanisms should adopt the corresponding amount of congestion extent, such as queuing delay based on TCP Vegas, TCP Reno based on packet loss, RED queue based on the length of queue, REM based on the price. some common TCP/AQM strategies are given in reference [14], such as Reno/RED, Reno/REM, Reno/DropTail, Vegas/DropTail and the specific form of triples of Vegas/REM (F, G, U), and their steady-

nature is analyzed. Recently, PAGANINI, based on above model, further analyzed the robustness and stability of optimized congestion control by means of Feedback Control Theory. KELLY also proposed congestion control framework of another type based on Optimization Theory, and use Lyapunov Stability Theory to analyze the stability of congestion control system.

For each TCP connection, the sender maintains two parameters, namely the congestion window and slow start threshold. The congestion window is used to describe the maximum amount of data the sender can send over the network before receiving the confirmation message; the slow start threshold is what the sending side uses to determine whether slow start algorithm or congestion avoidance algorithm will be adopted to control the data transfer. The minimum value of congestion window (CWND) and the receiver advertised window (RWND) determines a maximum amount of data the sender can transmit[15].

The pseudo codes of improved algorithm : (in which, swin is the sender window, awin is notification window for the recipient)

```
cwnd = IW;
swin = min (wnd, awin);
ssthresh = 65535bytes (default value);
Sender receives a packet ACK which is newly confirmed:
if (cwnd <ssthresh)
{
/ Slow start begins /
cwnd = cwnd +1;
}
else
{
/ Congestion avoidance /
cwnd = cwnd +1 / cwnd;
}
when senderdoesn't receive packet ACK confirmed because of timeout :
if (cwnd = IW)
ssthresh = ssthresh;
else
{
ssthresh = max (2, min (cwnd / 2, awin));
cwnd = IW;
}
```

This algorithm can still further improvements, making the algorithm is more effective, the improved algorithm in network timeout listed in this way processing .

Overtime not received confirmation bag when an ACK:

```
If (cwnd< threshold)
{
   ssthresh=ssthresh;
}
 else
 {
 ssthresh=max(2,min(cwnd/2,awin));
 cwnd= IW;
}
```

The improvement of algorithm is mainly for handling packet loss. In the improved algorithm, when packet loss occurs, the value of the network's congestion window will firstly be checked. If the value of the network's congestion window is the initial size of congestion control window, slow start threshold maintains the same, so as to effectively prevent the premature entry into congestion avoidance phase. Improved slow start algorithm can reduce the impact of packet loss caused by non-congestion factors on the congestion control algorithm, which has a positive significance for the application of the wireless network[16,17].

This algorithm could be further improved to make the algorithm more efficient. In the improved algorithm, network timeouts will be handled as follows:

when sender doesn't receive packet ACK confirmed because of timeout :

```
If (cwnd <threshold)
{
ssthresh = ssthresh;
}
else
{
ssthresh = max (2, min (cwnd / 2, awin));
cwnd = IW;
```

During the initial stage of the modified slow start algorithm, congestion control window is set as the initial window. After the sender sends the data package, the retransmission of packet does not change the slow start threshold in the case of receiving no ACK confirmation information in the RTO timeout set. Packet loss may occur in the initial congestion window or a small congestion window[18]. The modified congestion control algorithms need to address the small window packet loss. Modified congestion control algorithms do not affect the whole network performance, and have some practical value for specific network environment.

In accordance with the modified algorithm, if packet loss occurs in the initial congestion window or small congestion window during the TCP connection, slow start threshold will not change, so congestion control will be prevented from early entering congestion avoidance phase, which has a protective effect on the connection with a small amount of data sent, thereby enhancing the system's transmission efficiency[19].

The above analysis shows that: the modified slow start algorithm mainly deals with packet loss. When the network packet loss occurs, the value of network congestion should be firstly checked, if the value of network congestion window is the initial congestion control window size, slow start threshold should be maintained the same, so as to effectively prevent the premature entry into congestion avoidance phase. The algorithm can reduce the impact of packet loss caused by non-congestion on the congestion control algorithm[20]. Modified congestion control algorithm only changes the sender instead of modifying the entire network or the receiver protocols, so it does not affect network performance, and has a practical value for the application in specific network environment.

III. Application of Modified Algorithm in Campus Network

Campus network is a common local area network applied within schools, with the characteristics of limited connectivity, simple topology and high transmission bandwidth. In campus network, the connection of internal nodes with Education Network or Internet is usually by means of the public exports rented. Compared with high link bandwidth within the campus network, the export bandwidth of the campus network is usually small, and often forms a bottleneck in the export areas. Campus network, in its interior, is usually connected through the switch of the core, aggregation and access layer, and is divided into different segments according to different geographical locations. VLAN approach is generally adopted in the management of the user within the network; the users in the same VLAN are with the same desktop connection bandwidth, typically 100Mbps[21].

The topologies of campus network are roughly similar, basically based on the three-tier exchange technology; Figure 1 is the network topology of Henan Institute of Technology, which is a typical campus network connection topology, within the campus network, the core switch is connected with aggregation switches with Gigabit; aggregation switches are connected with switches with Gigabit; switches access to the users with Fast; the school's total export bandwidth as 100M is uplinked to Henan Normal University. All the campus networks are roughly the same, typical of the tree. According to different objects, there are two campus network services: internal services and external services. Internal services provide services to internal users through the connection of core switch with the backbone network, and provide the campus network users with WEB, FTP, VOD, MAIL, and other management systems services; for example, all the servers of Henan Institute of Technology used HPDL850, which provides dual-gigabit to core switch (DB10808) for the above services. Within the school, in terms of all net users, this bandwidth is adequate, and network congestion will not occur because of bandwidth problems in internal services; but the bandwidth at the exit of Henan Normal University is 100M, easy to form a bottleneck here and easy to cause obstruction here in terms of foreign service of the campus network.

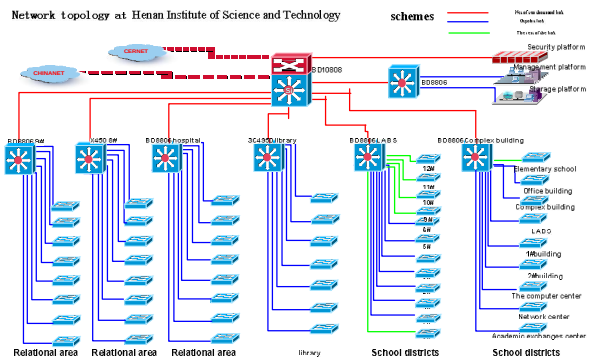


Figure 1. Campus network topology structure

Within the campus network, the terminals in the same VLAN can share network congestion information, so a node can be designed to represent the connection of an entire VLAN. Specific topology is described as follows: node n1 stands for connection node of the internal server; node n2 stands for source node of fixed rate stream; node n3 is used to replace multiple network segments within the campus network. When TCP connections of node n3 are greater than a certain number, it can be assumed that the total TCP connection bandwidth is close to a stable value, and network connection between node n2 and n3 is similar to the fixed flow rate (CBR)[22]. It can also be assumed that this kind of connection represents campus network exit bandwidth. Since the fixed rate of flow has a certain flexibility in the competition of bandwidth with the TCP connection, so as to simulate the actual network situation better, that is why the connection between node n2 and n3 uses fixed-rate flow instead of limiting bandwidth directly; Designing another two network connection nodes n4 and n5, the connection between node n1 and n5 uses the classical TCP congestion control to transmit data while the connection between node n1 and n4 uses the modified TCP (it is called TCPLAN in the later sections of the thesis) congestion control algorithm for data transmission. Network congestion occurs in the link from R0 to R1.

When network congestion occurs, applied control strategies use modified slow start. Within the campus network, small network transfer rate will drop the overall transmission efficiency; using modified slow start algorithm, we do not change the slow start threshold, so that all TCP connections can grow at a smooth rate, thus we can take advantage of the internal resources of network.

IV. Simulation

NS consists of two basic components, one is the extended and object-oriented TCL interpreter, and the other is the NS simulation database. The simulation database contains event schedules, a variety of simulated network entity object and the modules related with the network settings. Ordinary users can general use NS by the script description language, the specific process is as follows: First, set the network topology with NS. In the process of setting the topology, can be achieved using different network objects, and network objects function can be obtained through the object database; Then, set the data source and data receiver object; Finally, tell data sources what time to begin data transmission and what time to end it through event schedules. In NS, the user can add a new network object. Although the user can write themselves a new object, it is more accepted to deserve a new object from other libraries. Network flow simulation is a very complex task, but the use of NS software makes all the work become very simple[23].

Simulation of network flow using NS contains the following sections:(1)Programming: Using OTCL language programming: mainly including the creation of classes and objects, topology design, event design. As to the simulation of the underlying network protocol, the classes and objects existed in library can be used, while as for the design of the new agreement, new classes and objects must be generated by means of designing and

inheriting. (2) Simulation: to simulate using the NS, just enter ns file name in the appropriate path. (3) Analysis of the result: Currently, there are a number of ways to analyse the results of NS simulation: to simulate using animation, generating the image, or analyzing the event trace file. The analysis of trace file can help to thoroughly learn about the situation of the transmission of each packet.

In order to achieve a new TCP congestion control method, new NS components must be developed. In the NS-2, there are two ways to design a component: First, to achieve by writing new code, and second, to build a new component by inheriting existing components[24]. In the experiment, the inherited methods are used to build a new component TCPLAN.The construction of component by means of inheritace includes the following steps:

(1) TCPLAN header files

The header files of improved congestion control method are as follows:

```
New file: tcplan.h
class LANTCPAgent: public TcpAgent {
public:
virtual void set_initial_window ()
{
cwnd_ = var_initial;
}
Private:
Int_cwnd_ = var_initial;
};
```

(2) to create TCPLAN the C + + class files

The C++ class files of improved congestion control methods are as follows:

```
New file: tcplan.cc
static LANTcpClass: public TclClass {
public:
LANTcpClass (): TclClass ("Agent / TCP / Lan") {}
TclObject * create (int, const char * const *) {
return (new LANTcpAgent ());
}
};
LANTcpAgent:: LANTcpAgent () {
bind ("var_inital", var_initial);
}
```

In addition, the following work needs to be done: to define OTCL link, compose OTCL code, compile and so on.

According to the characteristics of intra campus network, the modified congestion control algorithm is simulated with NS-2. The simulation results show that the modified algorithm is suitable for intra-campus network services.

In the simulation experiment, the campus network topology and connectivity are shown in Figure 2. In the actual analysis, the discussion will be made according to agreements. Assuming there is only one network protocol, it will be discussed later in heterogeneous network structure.

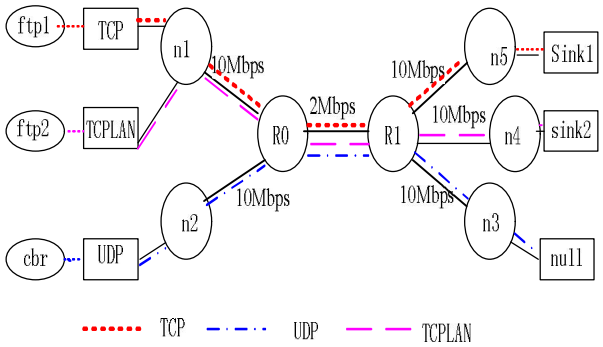


Figure 2. Network topology and connection diagram

Let us check the changes of the TCP's cwnd by means of simulation. In the link from n2 to R0, we use udp services. The using of a stable cbr flow to transmit above it does not cause congestion at the link node R0. In the link from n1 to R0, we use the FTP stream to transmit, with slow start algorithm as its control strategy.

It is described as follows with the TCL language.

```
Set ftp [new Agent/ftp/newreno]
$ns attach –agent $n1 $R0 $ftp
Set tcplan [new Agent/tcplan/newreno]
$ns attach –agent $n1 $R0 $tcplan
Set udp [new Agent/udp/newreno]
$ns attach –agent $n2 $R0 $udp
Set sink [new Agent/tcpsink/newreno]
$ns attach –agent $n5 $R1 $sink1
$ns attach –agent $n4 $R1 $sink2
$ns attach –agent $n3 $R1 $null
$ns connect $tcp $sink
$ns duplex-link $R0 $R1 2Mb 10ms DropTail
$ns duplex-link $R0 $n1 10Mb 2ms DropTail
$ns duplex-link $R0 $n2 10Mb 2ms DropTail
$ns duplex-link $n3 $R1 10Mb 2ms DropTail
$ns duplex-link $n4 $R1 10Mb 2ms DropTail
$ns duplex-link $n5 $R1 10Mb 2ms DropTail
$ns duplex-link – op $n1 $R0 orient left-up
$ns duplex-link – op $n2 $R0 orient left-down
$ns duplex-link – op $n3 $R1 orient right-down
$ns duplex-link – op $n4 $R1 orient right
$ns duplex-link – op $n5 $R1 orient right-up
$ns queue − limit $n1 $R0 10;
Set cbr [new application/traffic/cbr]
$cbr attach –agent $udp
$cbr set type_CBR
$cbr set packet_size 1000
$cbr set rate_1mb
$cbr set random_false
$ns at 0 “$cbr start ”
$ns at 0 “$tcp start”
$ns at 0 “$tcplan start”
$ns at 1.2 “finsh”
$ns run
$ns trace – queue $n1 ,$R0 ,$tracefile
```

Using gnuplot to analyze, cwnd changes are shown in Figure 3.

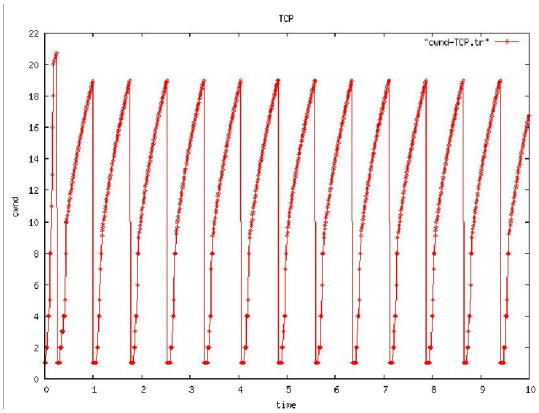


Figure 3. TCP's cwnd changes

The results show that, the value of TCP's Congestion Window will show repeated periodical change[25]. When TCP begins to execute, it first begins from Slow-start, then enters the Congestion Avoidance phase when cwnd is over Ssthresh. When the packets sent to the network continuously increase until the number which can be transmitted on the network is over the tolerance, the router starts using Drop-tail to discard the packets. When packet loss occurs, TCP will set ssthresh as 1/2 of Window value which is the value when packet loss is found, then set the value of the Window as 1. TCP have to re-start from the slow-start when each packet loss occurs.

A. TCPLAN transmission effect

The ftp connection is still used to simulate new agreement; the network topology is shown in Figure 4.

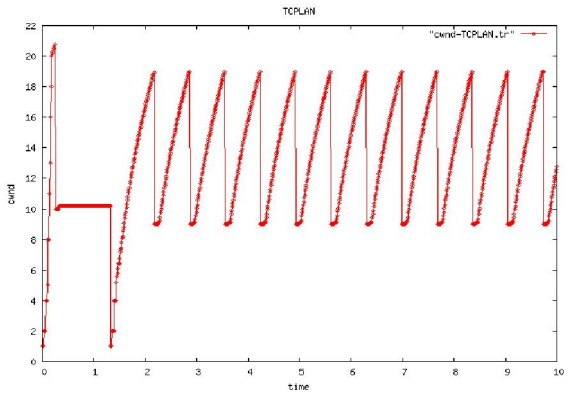


Figure 4. The cwnd change map of TCPLAN

To describe the following with the TCL language:

```
Set ftp [new Agent/ftp/newreno]
$ns attach –agent $n1 $R0 $ftp
Set tcplan [new Agent/tcplan/newreno]
$ns attach –agent $n1 $R0 $tcplan
Set udp [new Agent/udp/newreno]
$ns attach –agent $n2 $R0 $udp
Set sink [new Agent/tcpsink/newreno]
$ns attach –agent $n5 $R1 $sink1
$ns attach –agent $n4 $R1 $sink2
$ns attach –agent $n3 $R1 $null
$ns connect $tcp $sink
$ns duplex-link $R0 $R1 2Mb 10ms DropTail
$ns duplex-link $R0 $n1 10Mb 2ms DropTail
$ns duplex-link $R0 $n2 10Mb 2ms DropTail
$ns duplex-link $n3 $R1 10Mb 2ms DropTail
$ns duplex-link $n4 $R1 10Mb 2ms DropTail
$ns duplex-link $n5 $R1 10Mb 2ms DropTail
$ns duplex-link – op $n1 $R0 orient left-up
$ns duplex-link – op $n2 $R0 orient left-down
$ns duplex-link – op $n3 $R1 orient right-down
$ns duplex-link – op $n4 $R1 orient right
$ns duplex-link – op $n5 $R1 orient right-up
$ns queue – limit $n1 $R0 10;
Set cbr [new application/traffic/cbr]
$cbr attach –agent $udp
$cbr set type_CBR
$cbr set packet_size 1000
$cbr set rate_1mb
$cbr set random_false
$ns at 0 “$cbr start ”
$ns at 0 “$tcp start”
$ns at 0 “$tcplan start”
$ns at 1.2 “finsh”
$ns run
$ns trace – queue $n1 ,$R0 ,$tracefile
```

The results show that in the network environment with a large amount of data transmission, TCPLAN can always improve their sending rate in a very short period of time. With the increase in the amount of data transfer, according to the data 2 seconds later from the start of connection, it is faster for TCPLAN to probe the available bandwidth than TCP with the increase of connection time.Under the same condition, the transmission rate of TCPLAN is still higher than the transmission rate of TCP.

This kind of congestion control method, which is the one that network topology and connectivity features have been known, can be used within the campus network services, but the algorithm is lack of certain versatility.

B. The transmission effect of TCPLAN in the environment of heterogeneity

The above experiment is only based on the operation of only one TCP protocol, but in the actual network, it is bound to consider running in co-existence with other TCP versions. The following is the comparison on the condition of heterogeneous structure between TCPLAN and TCP Vegas.

In simulation experiment, the campus network is simplified as Figure 5:

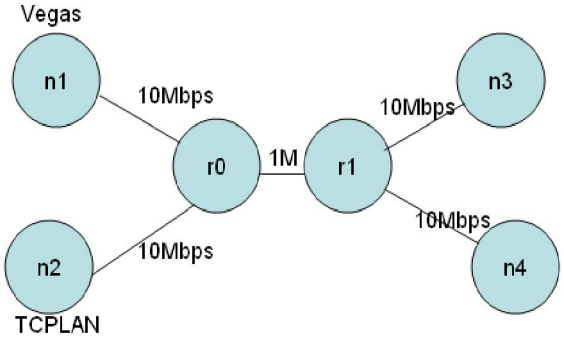


Figure 5. Network topology and connection diagram

In which, the delay between r0 and r1 is assumed to be 20ms, while the one among other links is assumed to be 1ms.

The main code is described as follows with TCL language:

```
$ns duplex-link $n0 $r0 10Mb 1ms DropTail
$ns duplex-link $n2 $r0 10Mb 1ms DropTail
$ns duplex-link $r0 $r1 1Mb  20ms RED
$ns duplex-link $r1 $n1 10Mb 1ms DropTail
$ns duplex-link $r1 $n3 10Mb 1ms DropTail
set buffer_size 15
$ns queue-limit $r0 $r1 $buffer_size
set tcp0 [new Agent/TCP/Vegas];
$tcp0 set v_alpha_ 1
$tcp0 set v_beta_ 3
$tcp0 set debug_ 0
$tcp0 set window_ 24
$tcp0 set fid_ 0
$ns attach-agent $n0 $tcp0
set tcp0sink [new Agent/TCPSink]
$tcp0sink set fid_ 0
$ns attach-agent $n1 $tcp0sink
$ns connect $tcp0 $tcp0sink
set ftp0 [new Application/FTP]
$ftp0 attach-agent $tcp0
Set tcp1 [new Agent / TCP / FTPLAN]
set tcp1 [new Agent/TCP/FTPLAN]
$tcp1 set window_ 24
$tcp1 set fid_ 1
$ns attach-agent $n2 $tcp1
set tcp1sink [new Agent/TCPSink]
$tcp1sink set fid_ 1
$ns attach-agent $n3 $tcp1sink
$ns connect $tcp1 $tcp1sink
set ftp1 [new Application/FTP]
$ftp1 attach-agent $tcp1
```

Figure 6 shows the cwnd change of Vegas and TCPLAN

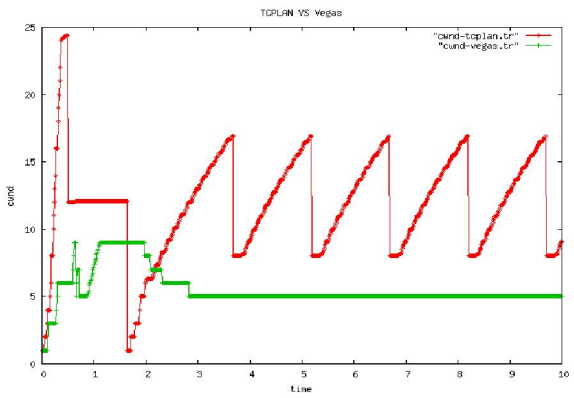


Figure 6 The cwnd change map of TCPLAN and Vegas

It can be seen, TCPLAN always places a higher vibration, while the Window of Vegas is always kept at a lower position. As TCPLAN uses a more aggressive congestion control strategy, the transmitter will continue to send packets on the network, while Vegas uses a more conservative approach. In contrast, TCPLAN has higher bandwidth occupation capabilities.

The results show that: the modified slow start algorithm can improve network transmission efficiency in the particular network environment; Because the algorithm only changes the agreement at the sending end, and there is no requirement on the receiving end, so the using of the algorithm does not affect the internal service performance of network. The algorithm is suitable for the connection within the campus network, and also applies to other network connections which have same connection characteristics.

V. CONCLUSION

Although in network congestion control has been doing a lot of research, but because the network rapid development and the complexity of congestion control, network congestion control will still faced a series of problems need to be solved. With the expanding of network size and networking mode diversification, need more reliable, more timely network congestion detection mechanism, single depend on lost package or rely on the judging repeat and confirm receipt method of network congestion already can not adapt to the needs of the development of the network.

For congestion control problem of research, based on theoretical analysis method and using the simulation software simulation methods have some shortcomings, congestion control complexity is a long-standing problems. As IPv6 technology development and mature, how to play the IP layer in congestion control function, as well as TCP and IP congestion control cooperation and is also a new problems waiting to be solved. TCP/IP congestion control the design and realization of facing hordes of compromise, the existing congestion control method and technology in multi-objective various environments faced with many challenges, there are many areas in need of improvement, congestion control will attract many researchers for congestion control solution of the problems but work hard to struggle.

REFERENCES

[1] Floyd S. A report on some recent developments in TCP congestion control [J]. IEEE Communication Magazine, 2005, 35(4):84-90

[2] Tranenbaum A S，Computer networks，2006,137~152

[3] Low, S.H. Paganini, F. Doyle, J.C. Internet congestion control . 2006, 22 (1): 28-43

[4] J. Nagle ，Congestion control in TCP/IP internetworks，ACM SIGCOMM Computer Communication Review, 2008,127~139
http://www.sprintlabs.com/People/diot/publications.html

[5] Feng, W., Kandlur, D., Saha, D., et al. A self-configuring RED gateway. In: Doshi, B., ed. Proceedings of the IEEE INFOCOM.New York: IEEE Communications Society, 2005, 1320-1328

[6] H. Yousefi'zadeh, and H. Jafarkhani ,A Finite-State Markov Chain Model for Statistical Loss Across a RED Queue , Systems Communications, 2005,213~215

[7] Allman, M., Floyd, S., Partridge, C. Increasing TCP's Initial Window. RFC 2414, 2009,321~325

[8] D Katabi, M Handley, C Rohrs, Internet congestion control for future high bandwidth-delay product environments - ACM SIGCOMM, 2002 pittsburgh,august,2004,45~48

[9] S Floyd, S Ratnasamy, S Shenker Modifying TCP's Congestion Control for High Speeds,- Preliminary Draft. http://www. icir. org/floyd/papers/ hstcp.html,2010.
[10] S Floyd , HighSpeed TCP for Large Congestion Windows, IETF draft, work in progress, 2004,12~14.
[11] S.Floyd, Limited slow-start for tcp with large congestion window, IETF,2006,43~45.
[12] Athuraliya S, LOW S H. Optimization flow control (2): implementation [DB/OL]. http://netlab.caltech.edu,2010.
[13] Low S H. A duality model of TCP and queue management algorithms [DB/OL]. http://netlab.caltech.edu,2009.
[14] MISRA V. Fluid-based analysis of a network of aqm routers supporting TCP flows with an application to RED [DB/OL]. http://www.net.cs.umsaa.Edu/ misra,2008
[15] Soohyun Cho, Riccardo Bettati. Collaborative Congestion Control in Parallel TCP Flows , 2005 ,1026-1031
[16] Internet Engineering Task Force. TCP Friendly Rate Control (TFRC): Protocol Specification. Internet Draft, 2006,63~64
[17] MISRAV. Fluid-based analysis of a network of aqm routers supporting TCP flows with an application to RED [DB/OL]. http://www.net.cs.umsaa.Edu/ misra,2008
[18] Soohyun Cho, Riccardo Bettati. Collaborative Congestion Control in Parallel TCP Flows , 2005 ,1026-1031
[19] Low S H. A duality model of TCP and queue management algorithms [DB/OL]. http://netlab.caltech.edu,2010.
[20] F. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks:Shadow price proportional fairness and stability," J. Oper. Res.Soc, 2004,237–252
[21] Jeonghoon Mo and Jean Walrand,Fair End-to-End Window-Based Congestion Control , IEEE/ACM TRANSACTIONS ON NETWORKING, 2005,94~102
[22] T.Kelly,Scalable TCP:Improving performance in highspeed wide area networks,Submitted for publication,December 2004,211~237
[23] S.Floyd, Limited slow-start for tcp with large congestion window, IETF,2006,43~45
[24] S Floyd, S Ratnasamy, S Shenker Modifying TCP's Congestion Control for High Speeds,- Preliminary Draft. http://www. icir. org/floyd/papers/ hstcp.html,2009.
[25] Allman, M., Floyd, S., Partridge, C. Increasing TCP's Initial Window. RFC 2414, 2004,321~325

Gao Guohong (1975-), was born in Zhengzhou, China. He received his B.S degree in 2000 form Computer and Applications, Henan normal university in Xinxiang, his M.S degree in 2008 form School of Computer Technology, Huazhong University of Science and Technology, and enroll in Wuhan University of Technology in 2009, work hard at D.S degree. Currently he is a professor in the School of Information Engineer, Henan Institute of Science and Technology, Henan Xinxiang,China. The main publications include: Compute Operating System ;network and information security computer software.

Xiao Wenxian (1975-), was born in Nanyang, China. He received his B.S degree in 2000 form Computer and Applications, PLA information engineering university , his M.S degree in 2009 form School of Computer Technology, Huazhong University of Science and Technology, he is a associate professor in Henan Institute of Science and Technology, The main publications include: Compute Operating System (Beijing, China, National Defense University Press, 2010), Asp.Net Web Programming (Beijing, China, Nation Defense University Press, 2008). His research interests include: network and information security computer software. He is a advanced membership of WASE society.

Cryptanalysis and Improvement of Selvi et al.'s Identity-Based Threshold Signcryption Scheme

Wei Yuan
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: yuanwei1@126.com

Liang Hu
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: hul@jlu.edu.cn

Hongtu Li
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: li_hongtu@hotmail.com

Jianfeng Chu
Department of Computer Science and Technology, Jilin University, Changchun, China
Corresponding author, Email: chujf@jlu.edu.cn

Yuyu Sun
College of Computer Science and Technology, Jilin University, Changchun 130012, China,
Software Institue, Changchun University, Changchun 130022, China
E-mail: sunyy@ccu.edu.cn

***Abstract*—Signcryption can realize the function of encryption and signature in a reasonable logic step, which can lower computational costs and communication overheads. In 2008, S. S. D. Selvi et al. proposed an identity-based threshold signcryption scheme. In this paper, we show that the threshold signcryption scheme of S. S. D. Selvi et al. is vulnerable if the attacker can replaces the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.**

***Index Terms*—identity-based, Signcryption, bilinear pairing, cryptanalysis, attack**

I. INTRODUCTION

Encryption and signature are the two basic cryptographic tools offered by public key cryptography for achieving confidentiality and authentication. Signcryption can realize the function of encryption and signature in a reasonable logic step which is proposed by ZHENG [1] in 1997. Comparing to the traditional way of signature then encryption or encryption then signature, signcryption can lower the computational costs and communication overheads. As a result, a number of signcryption schemes [2][3][4][5][6][7][8]were proposed following ZHENG's work. The security notion for signcryption was first formally defined in 2002 by Baek et al. [9] against adaptive chosen ciphertext attack and adaptive chosen message attack. The same as signature and encryption, signcryption meets the attributes of confidentiality and unforgeability as well.

In 1984, A.Shamir [10] introduced identity-based public key cryptosystem, in which a user's public key can be calculated from his identity and defined hash function, while the user's private key can be calculated by a trusted party called Private Key Generator (PKG). The identity can be any binary string, such as an email address and needn't to be authenticated by the certification authentication. As a result, the identity-based public key cryptosystem simplifies the program of key management to the conventional public key infrastructure. In 2001, Boneh and Franklin [11] found bilinear pairings positive in cryptography and proposed the first practical identity-based encryption protocol using bilinear pairings. Soon, many identity-based [12][14][15][16] and other relational [13][17][18] schemes were proposed and the bilinear pairings became important tools in constructing identity-based protocols.

Group-oriented cryptography [19] was introduced by Desmedt in 1987. Elaborating on this concept, Desmedt and Frankel [20] proposed a (t,n) threshold signature scheme based RSA system [21]. In such a (t,n) threshold signature scheme, any to out of n signers in the group can collaboratively sign messages on behalf of the group for sharing the signing capability.

Identity-based signcryption schemes combine the advantages of identity-based public key cryptosystem and Signcryption. The first identity-based threshold signature scheme was proposed by Baek and Zheng [22] in 2004. Then Duan et al. proposed an identity-based threshold signcryption scheme [23] in the same year by combining the concepts of identity based threshold signature and encryption together. However, in Duan et al.'s scheme, the master-key of the PKG is distributed to a number of other PKGs, which creates a bottleneck on the PKGs. In

doi:10.4304/jnw.6.11.1557-1564

2005, Peng and Li proposed an identity-based threshold signcryption scheme [24] based on Libert and Quisquater's identity-based signcryption scheme [25]. However, Peng and Li's scheme dose not provide the forward security. In 2008, another scheme [26] was proposed by Fagen Li et al., which is more efficient comparing to previous scheme. However, S. S. D. Selvi et al. pointed out that Fagen Li et al.'s scheme is not equilibrium between the usual members and a dealer called clerk in Fagen Li et al.'s scheme and proposed an improved scheme [27].

In this paper, we show that the threshold signcryption scheme of S. S. D. Selvi et al. is vulnerable if the attacker can replaces the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

II. PRELIMINARIES

A. Bilinear pairing

Let G_1 be a cyclic additive group generated by P, whose order is a prime q, and G_2 be a cyclic multiplicative group with the same order q. A bilinear pairing is a map $e: G_1 \times G_1 \to G_2$ with the following properties:

1. Bilinearity: $e(aP, bQ) = e(P, Q)^{ab}$ for all $P, Q \in G_1$, $a, b \in Z_q$.

2. Non-degenerative: There exists $P, Q \in G_1$ such that $e(P, Q) \neq 1$.

3. Computable: There is an efficient algorithm to compute $e(P, Q)$ for all $P, Q \in G_1$.

B. Computational problems

Let G_1 and G_2 be two groups of prime order q, let $e: G_1 \times G_1 \to G_2$ be a bilinear pairing and let P be a generator of G_1.

- Discrete Logarithm Problem (DLP)
  Given $P, Q \in G_1$, find $n \in Z_q$ such that $P = nQ$ whenever such n exists.
- Computational Diffie-Hellman Problem (CDHP)
  Given $(P, aP, bP) \in G_1$ for $a, b \in Z_q^*$, find the element abP.
- Bilinear Diffie-Hellman Problem (BDHP)
  Given $(P, aP, bP, cP) \in G_1$ for $a, b, c \in Z_q^*$, compute $e(P, P)^{xyz} \in G_2$
- Bilinear Diffie-Hellman Problem (DBDHP)
  Given $(P, aP, bP, cP, \tau) \in G_1^4 \times G_2$ for $a, b, c \in Z_q^*$, decide whether $\tau = e(P, P)^{abc}$

C. Identity Based Threshold Signcryption

A generic identity-bases threshold signcryption scheme with total n players and t threshold limit consists of the following five algorithms:

Setup: Given a security parameter k, the private key generator (PKG) generates the system's public parameters. Among the parameters produced by Setup is a public key P_{pub}. There is also a corresponding master key s that is kept secret by PKG.

Extract: Given a user's identity ID, the PKG will compute a public key Q_{ID}, generate the private key S_{ID} and transmit the private key to its owner in a secure way.

Keydis: Given a private key S_{ID} associated with an identity ID that stands for a group of users, the number of signcryption members n and a threshold parameter t, this algorithm generates n shares of S_{ID} and provides each one to the signcryption members $M_1, M_2, \ldots, M_n$. It also generates a set of verification keys that can be used to check the validity of each shared private key. We denote the shared private keys and the matching verification keys by $\{S_i\}_{i=1,\ldots,n}$ and $\{y_i\}_{i=1,\ldots,n}$, respectively. Note that each (S_i, y_i) is sent to M_i, then M_i publishes y_i but keeps S_i secret.

Signcrypt: Give a message m, the private keys of t members $\{S_i\}_{i=1,\ldots,t}$ in a sender group U_A, the receiver's public key Q_{ID_B}, the Signcrypt algorithm outputs an identity-based (t,n) threshold signcryption σ on the message m.

Designcrypt: Give a ciphertext σ, the private key of the receiver S_{ID_B}, the public key the sender group Q_{ID_A}, it outputs the plain text m or $\perp$ if σ is an invalid ciphertext between the group U_A and the receiver.

D. Security notions for identity-based Threshold signcryption

The notion of semantic security of public key encryption was extended to identity-based signcryption scheme by Malone-Lee [28]. This was later modified by Sherman et al. [29] which incorporates indistinguishability against adaptive chosen ciphertext and identity attacks (IND-IDTSC-CCA2) and existential unforgeability against adaptive chosen message and identity attacks (EUF-IDTSC). We describe below the security notions for confidentiality and unforgeability given in [30], this is the strongest security notion for this problem.

Confidentiality: A signcryption scheme is semantically secure against chosen ciphertext and identity attacks (IND-IDTSC-CCA2) if no probabilistic

polynomial time adversary Eve has a non-negligible advantage in the following game:

1. The challenger C runs the Setup algorithm and sends the system public parameters to the adversary Eve.

2. In the first phase, Eve makes polynomial bounded number of queries to the following oracles.

Extract Oracle: Eve produces an identity ID_i and queries for the secret key of user i. The Extract Oracle returns S_i to Eve.

Signcrypt Oracle: Eve produces a message m, sender identity ID_A and receiver identity ID_B. C computes the secret key S_A from Extract Oracle and returns to Eve, the signcrypted ciphertext from Signcrypt $m, \{S_i\}_{i=1,\ldots,t}, ID_j$.

Unsigncrypt Oracle: Eve produces a sender identity ID_A and receiver identity ID_B and a signcryption σ. The challenger C computes the secret key S_B from Extract Oracle, returning the result of $Unsigncrypt(\sigma, Q_{ID_A}, S_B)$ to Eve. The result returned is $\perp$ if σ is a valid signcryption from U_A to U_B.

3. A produces two messages m_0 and m_1 of equal length from the message space M and an arbitrary sender identity ID_A. The challenger C flips a coin, sampling a bit $b \in \{0,1\}$ and computes $\sigma^* = Signcrypt(m_b, \{S_i\}_{i=1,\ldots,t}, ID_B)$. σ^* is return to Eve as challenge signcrypted ciphertext.

4. Eve is allowed to make polynomial bounded number of new queries as in step 2 with the restrictions that it should not query the Unsigncryption oracle for the unsigncryption of σ^*, the Signcryption Oracle for the signcryption of m_0 or m_1 under the sender identity ID_A and the Extract Oracle for the secret keys of ID_B.

5. At the end of this game, Eve outputs a bit b'. Eve wins the game if $b' = b$.

Unforgeability: A signcryption scheme is existentially unforgeable under chosen message attack (EUF-IDTSC) if no probabilistic polynomial time adversary Eve has a non-negligible advantage in the following game.

1. The challenger C runs the Setup algorithm to generate the master public and private keys params and msk respectively. C gives system public parameters params to Eve and keeps the master private key msk secret from Eve.

2. The adversary Eve makes polynomial bounded number of queries to the oracles as described in step 2 of the confidentiality game.

3. Eve produces a signcrypted ciphertext σ and wins the game if the private key of sender U_A was not queried in the previous step and $\perp$ is not returned by $Unsigncrypt(\sigma, Q_{ID_A}, S_B)$ and σ is not the output of a previous query to the Signcrypt Oracle with ID_A as sender.

III. Review of S. S. D. Selvi et al.'s Identity-Based Threshold Signcryption Scheme

The scheme involves four roles: the PKG, a trust dealer, a sender group $U_A = \{M_1, M_2, \ldots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B.

Setup: Given a security parameter k, the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1, a bilinear map $e: G_1 \times G_1 \to G_2$, a secure symmetric cipher (E,D) and hash functions $H_1: \{0,1\}^* \to G_1$, $H_2: G_2 \to \{0,1\}^{n_1}$, $H_3: \{0,1\}^* \to Z_q^*$. The PKG chooses a master-key $s \in_R Z_q^*$ and computes $P_{pub} = sP$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$ and keeps the master-key s secret.

Extract: Given an identity ID, the PKG computes $Q_{ID} = H_1(ID)$ and the private key $S_{ID} = sQ_{ID}$. Then PKG sends the private key to its owner in a secure way.

Keydis: Suppose that a threshold t and n satisfy $1 \le t \le n < q$. To share the private key S_{ID_A} among the group U_A, the trusted dealer performs the steps below.

1) Choose $F_1, \ldots, F_{t-1}$ uniformly at random from G_1^*, construct a polynomial $F(x) = S_{ID_A} + xF_1 + \cdots + x^{t-1}F_{t-1}$

2) Compute $S_i = F(i)$ for $i = 0, \ldots, n$. ($S_0 = S_{ID_A}$). Send S_i to member M_i for $i = 1, \ldots, n$ secretly.

3). Broadcast $y_0 = e(S_{ID_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \ldots, t-1$.

4) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{i^j}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt: Let $M_1, \ldots, M_t$ are the t members who want to cooperate to signcrypt a message m on behalf of the group U_A.

1) Each M_i chooses $x_i \in_R Z_q^*$, computes $R_{1i} = x_i P$, $R_{2i} = x_i P_{pub}$, $\tau_i = e(R_{2i}, Q_{ID_B})$ and sends (R_{1i}, τ) to the clerk C.

2) The clerk C (one among the t cooperating players) computes $R_1 = \prod_{i=1}^{t} R_{1i}$, $\tau = \prod_{i=1}^{t} \tau_i$, $k = H_2(\tau)$, $c = E_k(m)$, and $h = H_3(m, R_1, k)$.

3) Then the clerk C sends h to M_i for $i = 0, \ldots, t$.

4) Each M_i computes the partial signature $W_i = x_i P_{pub} + h\eta_i S_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j\neq i}^{t} -j(i-j)^{-1} \bmod q$.

5) Clerk C verifies the correctness of partial signatures by checking if the following equation holds:

$$e(P, W_i) = e(R_{1i}, P_{pub})(\prod_{j=0}^{t-1} y_j^{i^j})^{h\eta_i}$$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^{t} W_i$; otherwise rejects it and requests a valid one.

6) The final threshold signcryption is $\sigma = (c, R_1, W)$.

Unsigncrypt: When receiving σ, Bob follows the steps below.

1) Compute $\tau = e(R_1, S_{ID_B})$ and $k = H_2(\tau)$.

2) Recover $m = D_k(c)$

3) Compute $h = H_3(m, R_1, k)$ and accept σ if and only if the following equation holds:

$$e(P, W) = e(P_{pub}, R_1 + hQ_{ID_A})$$

Ⅳ. CRYPTANALYSIS OF S. S. D. SELVI ET AL.'S SCHEME

The two schemes are both insecure from the view of attack by a malicious attacker who can control the communication channel.

The attacker intercepts the ciphertext $\sigma = (c, R_1, W)$ from sender.

1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message m'

2) Compute $R_1' = xP$, $R_2' = xP_{pub}$, $\tau' = e(R_2', Q_{ID_B})$, $k' = H_2(\tau)$, $c' = E_{k'}(m')$, $h' = H_3(m', R_1', k')$.

3) Compute $W' = \alpha P_{pub}$, set $Q_A' = (\alpha - x)P / h'$ as a public key of U_A

4) The final ciphertext is $\sigma' = (c', R_1', W')$.

5) Attacker sends the forged ciphertext and the replaced public key to the receiver.

After receiving the ciphertext $\sigma' = (c', R_1', W')$, the receiver

1) Compute $\tau = e(R_1', S_{ID_B}) = e(R_2', Q_{ID_B}) = \tau'$, $k = H_2(\tau) = H_2(\tau') = k'$

2) Recover $m = D_k(c') = D_{k'}(c') = m'$, $h = H_3(m', R_1', k') = h'$.

3) Verify $e(P, W') \stackrel{?}{=} e(P_{pub}, R_1' + hQ_{ID_A}')$

$$\because e(P_{pub}, R_1' + hQ_{ID_A}') = e(P_{pub}, xP + h \cdot (\alpha - x)P / h') = e(P_{pub}, \alpha P) = e(P, W')$$

$\therefore$ The equation $e(P, W') = e(P_{pub}, R_1' + hQ_{ID_A}')$ set.

Discussion

In the view of the attacker, [27] can be simulated as following basic Signcryption scheme:

A sender "Alice" with key pairs $\{Q_A = H_1(Alice), S_A = sH_1(Alice)\}$

A receiver "Bob" with key pairs $\{Q_{Bob} = H_1(Bob), S_B = sH_1(Bob)\}$

Alice chooses $x \in Z_q^*$, $R_1 = xP$, $R_2 = xP_{pub}$, $\tau = e(R_2, Q_B)$, $k = H_2(\tau)$, $c = E_k(m)$, $h = H_3(m, R_1, k)$, $W = xP_{pub} + hS_A$ and sends $\sigma = (c, R_1, W)$ to Bob as the ciphertext of his message.

There is a small mistake of the definition $H_3 : \{0,1\}^* \to Z_q^*$. We think the authors' real intention is $H_3 : \{0,1\}^* \times G_1 \times \{0,1\}^* \to Z_q^*$ to meet $h = H_3(m, R_1, k)$. In this hash function, any message about the sender is not contained. If an attacker Eve say "I am Alice" to Bob, Bob can not distinguish only with the hash value h. Our attack just utilizes this attribute of Li's scheme.

Suppose that H_3 is defined as $H_3 : \{0,1\}^* \times G_1 \times \{0,1\}^* \times G_1 \to Z_q^*$, and $h = H_3(m, R_1, k, Q_{ID_A})$. The attacker Eve intercepts the ciphertext $\sigma = (c, R_1, W)$ from sender Alice and she runs the algorithm of forging ciphertext like:

1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message m'

2) Compute $R_1' = xP$, $R_2' = xP_{pub}$, $\tau' = e(R_2', Q_{ID_B})$, $k' = H_2(\tau)$, $c' = E_{k'}(m')$, $h' = H_3(m', R_1', k', Q_A')$.

3) Compute $W' = \alpha P_{pub}$, set $Q_A' = (\alpha - x)P / h'$ as a public key of U_A

4) The final ciphertext is $\sigma' = (c', R_1', W')$.

5) Send the forged ciphertext and the replaced public key to the receiver.

She will meet a hard problem that if she wants to compute h', Q_A' is necessary or if she wants to computes Q_A', h' must be known. As a result, if she can succeed in forging the ciphertext, she must own the ability to solve the DL problem.

Ⅴ. THE IMPROVEMENT OF S. S. D. SELVI ET AL.' SCHEME

The scheme involves four roles: the PKG, a trust dealer, a sender group $U_A = \{M_1, M_2, \ldots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B.

Setup: Given a security parameter k, the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1, a bilinear map $e: G_1 \times G_1 \to G_2$, a secure symmetric cipher (E,D) and hash functions $H_1: \{0,1\}^* \to G_1$, $H_2: G_2 \to \{0,1\}^{n_1}$, $H_3: \{0,1\}^* \times G_1 \times \{0,1\}^* \times G_1 \to Z_q^*$. The PKG chooses a master-key $s \in_R Z_q^*$ and computes $P_{pub} = sP$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$ and keeps the master-key s secret.

Extract: Given an identity ID, the PKG computes $Q_{ID} = H_1(ID)$ and the private key $S_{ID} = sQ_{ID}$. Then PKG sends the private key to its owner in a secure way.

Keydis: Suppose that a threshold t and n satisfy $1 \le t \le n < q$. To share the private key S_{ID_A} among the group U_A, the trusted dealer performs the steps below.

1) Choose $F_1, \ldots, F_{t-1}$ uniformly at random from G_1^*, construct a polynomial $F(x) = S_{ID_A} + xF_1 + \cdots + x^{t-1}F_{t-1}$

2) Compute $S_i = F(i)$ for $i = 0, \ldots, n$. ($S_0 = S_{ID_A}$). Send S_i to member M_i for $i = 1, \ldots, n$ secretly.

3). Broadcast $y_0 = e(S_{ID_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \ldots, t-1$.

4) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{i^j}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt: Let $M_1, \ldots, M_t$ are the t members who want to cooperate to signcrypt a message m on behalf of the group U_A.

1) Each M_i chooses $x_i \in_R Z_q^*$, computes $R_{1i} = x_i P$, $R_{2i} = x_i P_{pub}$, $\tau_i = e(R_{2i}, Q_{ID_B})$ and sends (R_{1i}, τ) to the clerk C.

2) The clerk C (one among the t cooperating players) computes $R_1 = \prod_{i=1}^{t} R_{1i}$, $\tau = \prod_{i=1}^{t} \tau_i$, $k = H_2(\tau)$, $c = E_k(m)$, and $h = H_3(m, R_1, k, Q_{ID_A})$.

3) Then the clerk C sends h to M_i for $i = 0, \ldots, t$.

4) Each M_i computes the partial signature $W_i = x_i P_{pub} + h\eta_i S_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j \ne i}^{t} -j(i-j)^{-1} \bmod q$.

5) Clerk C verifies the correctness of partial signatures by checking if the following equation holds:

$$e(P, W_i) = e(R_{1i}, P_{pub})(\prod_{j=0}^{t-1} y_j^{i^j})^{h\eta_i}$$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^{t} W_i$; otherwise rejects it and requests a valid one.

6) The final threshold signcryption is $\sigma = (c, R_1, W)$.

Unsigncrypt: When receiving σ , Bob follows the steps below.

1) Compute $\tau = e(R_1, S_{ID_B})$ and $k = H_2(\tau)$.

2) Recover $m = D_k(c)$

3) Compute $h = H_3(m, R_1, k, Q_{ID_A})$ and accept σ if and only if the following equation holds:

$$e(P, W) = e(P_{pub}, R_1 + hQ_{ID_A})$$

VI. SECURITY ANALYSIS OF OUR IMPROVED SCHEME

In this section, we will give a formal proof on Unforgeability and Confidentiality of our scheme under CDH problem and DBDH problem.

Theorem 1 (Unforgeability): Our improved scheme is secure against chosen message attack under the random oracle model if CDH problem is hard.

Proof: Suppose the challenger C wants to solve the CDH problem. That is, given (aP, bP) , C should computes abP .

C chooses system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$, sets $P_{pub} = aP$, and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(ID, \alpha, Q_{ID}, S_{ID})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the j-th query, he will return $Q_{ID_j} = bP$ to Eve and sets $(ID_j, \perp, Q_{ID_j} = bP, \perp)$ on L_1 . Else C selects $\alpha_i \in Z_q^*$, computes $Q_{ID_i} = \alpha_i P$, $S_{ID_i} = \alpha_i P_{pub}$, returns Q_{ID_i} to E and sets $(ID_i, \alpha_i, Q_i, S_i)$ on L_1.

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k). If C receives a query about τ_i , selects $k_i \in Z_q^*$, returns k_i to E, and sets (τ_i, k_i) on L_2.

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h). If C receives a query about $(m_i, R_{1i}, k_i, Q_{ID_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve, and sets $(m_i, R_{1i}, k_i, Q_{ID_i}, h_i)$ on L_3.

Signcrypt query: If C receives a query about Signcrypt with message m_i , identity ID_i

1. Select $x_i \in Z_q^*, W_i \in G_1$

2. Look-up L_1, L_2, set $Q_{ID_i} = \alpha_i P$ in L_1, $k_i = k_i$ in L_2, and compute $R_i = x_i Q_{ID_i}$

3. Set $h_i = H_3(m_i, R_i, k_i, Q_{ID_i})$.

4. Return (h_i, W_i) to Eve.

Finally, Eve output a forged signcryption (m, h_i, W_i, Q_{ID_i}). If $Q_{ID_i} \neq Q_{ID_j}$, Eve fails. Else, if $Q_{ID_i} = Q_{ID_j}$, Eve succeeds in forging a signcryption.

As a result, C gains two signcryption ciphertexts which meet:

$$e(P, W_i) = e(P_{pub}, R_i + h_i Q_{ID_i})$$

$$e(P, W_j) = e(P_{pub}, R_j + h_j Q_{ID_j})$$

Thus,

$$e(P, (W_i - W_j)) = e(P_{pub}, (R_i + h_i Q_{ID_i}) - (R_j + h_j Q_{ID_j})) \quad (1)$$

Note $Q = Q_{ID_i} = Q_{ID_j}$,

(1) can be expressed as $e(P, (W_i - W_j)) = e(P_{pub}, (R_i - R_j) + (h_i - h_j)Q)$ (2)

$\because P_{pub} = aP, Q_{ID_j} = bP$

(2) can be expressed as $e(P, (W_i - W_j)) = e(aP, ((\alpha_i - \alpha_j) + (h_i - h_j))bP)$

$\therefore W_i - W_j = ((\alpha_i - \alpha_j) + (h_i - h_j))abP$

Hence, the CDH problem $abP = \dfrac{W_i - W_j}{(\alpha_i - \alpha_j) + (h_i - h_j)}$ can be computed by C with aP and bP.

Theorem 2 (Confidentiality): Our improved scheme is secure against adaptive chosen ciphertext and identity attack under the random oracle model if DBDH problem is hard.

Proof: Suppose the challenger C wants to solve the DBDH problem. That is, given (P, aP, bP, cP, τ), C should decide whether $\tau = e(P,P)^{abc}$ or not. If there exists an adaptive chosen ciphertext and identity attacker for our improved scheme, C can solve the DBDHP.

C chooses system parameters $\{G_1, G_2, n_1, e, P, P_{pub}, E, D, H_1, H_2, H_3\}$, sets $P_{pub} = aP$, and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(ID, \alpha, Q_{ID}, S_{ID})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the j-th query, he will return $Q_{ID_j} = bP$ to Eve and sets $(ID_j, \perp, Q_{ID_j} = bP, \perp)$ on L_1. Else C selects $\alpha_i \in Z_q^*$, computes $Q_{ID_i} = \alpha_i P$, $S_{ID_i} = \alpha_i P_{pub}$, returns Q_{ID_i} to E and sets $(ID_i, \alpha_i, Q_i, S_i)$ on L_1.

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k). If C receives a query about τ_i, selects $k_i \in Z_q^*$, returns k_i to E, and sets (τ_i, k_i) on L_2.

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h). If C receives a query about $(m_i, R_{1i}, k_i, Q_{ID_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve, and sets $(m_i, R_{1i}, k_i, Q_{ID_i}, h_i)$ on L_3.

Signcrypt query: If C receives a query about Signcrypt with message m_i, identity ID_i

1. Select $c_i \in Z_q^*, W_i \in G_1$

2. Look-up L_1, L_2, set $Q_{ID_i} = \alpha_i P$ in L_1, $k_i = k_i$ in L_2. Compute $R_i = c_i P$,if $ID_i \neq ID_j$. Else, if $ID_i = ID_j$, compute $R_i = cP$

3. Set $h_i = H_3(m_i, R_i, k_i, Q_{ID_i})$.

4. Return (h_i, W_i) to Eve.

After the first stage, Eve chooses a pair of identities on which he wishes to be challenged on (ID_i, ID_j). Note that Eve can not query the identity of ID_A. Then Eve outputs two plaintexts m_0 and m_1. C chooses a bit $b \in \{0,1\}$ and signcrypts m_b. To do so, he sets $R_1^* = cP$, obtains $k^* = H_2(\tau)$ from the hash function H_2, and computes $c_b = E_{k_1^*}(m_b)$. Then C chooses $W^* \in G_1$ and sends the ciphertext $\sigma^* = (c_b, R_1^*, W^*)$ to Eve. Eve can performs a second series of queries like at the first one. At the end of the simulation, she produces a bit b' for which he believes the relation σ^* =Signcrypt $(m_{b'}, \{S_i\}_{i=1,\ldots,t}, ID_j)$ holds. If $b = b'$, C outputs

$\tau = e(R_1^*, S_{ID_j}) = e(cP, abP) = e(P,P)^{abc}$. Else, C outputs $\tau \neq e(P,P)^{abc}$. So C can solve the BDDH problem.

Ⅶ. Conclusion

In this paper, we show that the threshold signcryption scheme of S. S. D. Selvi et al. is vulnerable if the attacker can replaces the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably-secure improved scheme

to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

REFERENCES

[1] Zheng Y Digital signcryption or How to achieve cost (signature & Encryption)<< cost (signature) + cost (encryption), In Proc. Advances in CRYPTO'97, LNCS 1294, pp.165-179, Springer-Verlag,1997.

[2] Bao F., Deng R H, A signcryption scheme with signature directly verifiable by public key. PKC'98 LNCS, vol.1431, pp55-59, Springer-Verlag, 1997.

[3] Chow S.S.M., Yiu S.M., Hui L.C.K., Chow K.P., Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. ICISC'03 LNCS, vol.2971, pp.352-269, Spring-Verlag, 2004.

[4] Boyen X, Multipurpose identity based signcryption: a swiss army knife for identity based cryptography. CRYPT'03 LNCS, vol.2729, pp.383-399, Springer-Varlag, 2003.

[5] Mu Y., Varadharajan V., Distributed signcryption, INDOCRYPT'00. LNCS, vol.1977, pp.155-164, Springer-Varlag, 2000

[6] Yang G., Wong D.S., Deng X., Analysis and improvement of a signcryption scheme with key privacy, ISC'05. LNCS, vol.3650, pp.218-232, Springer-Varlag, 2005

[7] SteinFeld R., Zheng Y., A signcryption scheme based on integer factorization. ISW'00. LNCS, vol 1975, pp.308-322, Springer-Varlag, 2000

[8] Libert B., Quisquater J., Efficient signcryption with key prevacy from gap Diffie-Hellman groups. PKC'04 LNCS vol.2947, pp.187-200, Springer-Varlag, 2004

[9] Baek J., Steinfeld R., Zheng Y., Formal proofs for the security of signcryption, PKC'02 LNCS vol.2274, pp.80-98, Springer-Varlag, 2002

[10] A. Shamir,"Identity-based cryptosystems and signature schemes", CRYPTO'84 LNCS 196, pp.47-53, Springer-Varlag, 1984.

[11] D. Boneh, M. Franklin, Identity-based encryption from well pairing, CRYPTO'01, LNCS 2139, pp.213-229, Springer-Varlag, 2001

[12] P.S.L.M. Barreto, B. Libert, N. Mccullagh, J.J. Quisquater, Efficient and provably-secure identity-based signatures and signcryption from bilinear maps ASIACRYPT'05, LNCS 3788, pp.515-532, Springer-Verlag, 2005

[13] X. Huang, W. Susilo, Y. Mu, E Zhang, Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in t he ubiquitous world, 19th International Conference on Advanced Information Networking and Applications, pp.649-654, Taiwan, 2005

[14] Fagen Li, Hu Xiong, Xuyun Nie, A new multi-receiver ID-based signcryption scheme for group communications, ICCCAS'2009, pp.296-300, 2009

[15] Yiliang Han, Xiaolin Gui, Multi-recipient signcryption for secure group communication, ICIEA 2009, pp.161-165.

[16] Zhengping Jin, Qiaoyan Wen, Hongzhen Du, An improved semantically-secure identity-based signcryption scheme in the standard model, Computers and Electrical Engineering 36(2010), pp.545-552,Elsevier, 2010

[17] Zhenhua Liu, Yupu Hu, Xiangsong Zhang, Hua Ma, Certificateless signcryption scheme in the standard model, Information Sciences 180(2010), pp.452-464, Elsevier, 2010.

[18] Yong Yu, Bo Yang, Ying Sun, Sheng-lin Zhu, Identity based signcryption scheme without random oracles, Computer Standards & Interfaces 31(2009), pp.56-62, Elsevier, 2009

[19] Y. Desmedt, Society and group oriented cryptography: a now concept, CRYPTO'87, LNCS 293, pp.120-127, Springer-Varlag, 1987

[20] Y. Des. Frankel, Shared generation of authenticators and signatures, CRYPTO'91, LNCS 576, pp.457-469, Springer-Varlag, 1991

[21] R. L. Rivest, A. Shamir, L. Adleman, A method for obtaining digital signatures and public-key cryptosystems, Communications of the ACM, Vol.21, No.2, pp.120-126, 1978

[22] J. Baek, Y. Zheng, Identity-based threshold signature scheme from the bilinear pairings, International Conference on Information Technology 2004, pp.124-128, Las Vegas, Nevada, USA, 2004

[23] S. Duan, Z. Cao, R. Lu, Robust ID-based threshold signcryption scheme from pairings, International Conference on Information security, pp.33-37, Shanghai, China, 2004

[24] C. Peng, X. Li, An identity-based threshold signcryption scheme with semantic security, Computational Intelligence and Security 2005, LNAI 3902, pp.173-179, Springer-Varlag, 2005

[25] B. Libert, J.J. Quisquater, Anew identity based signcryption schemes from pairings, 2003 IEEE information theory workshop, pp.155-158, Paris, France, 2003

[26] Fagen Li, Yong Yu, An efficient and Provably Secure ID-Based Threshold Signcryption Scheme, ICCCAS 2008, 488-492

[27] Selvi S.S.D., Vivek S.S, Rangan C.P, Cryptanalysis of Li et al.'s Identity-Based Threshold Signcryption Scheme, Embedded and Ubiquitous Computing 2008, pp.127-132

[28] Malone Lee J:Identity based signcryption. In: Cryptology ePrint Archive. Report 2002/098, 2002.
[29] Chow S.S.M., Yiu S.M., Hui L.C.K., Chow K.P.: Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: Lin, J.-I., Lee, D.-H. (eds.) ICISC 2003. LNCS, vol. 2971, pp.352-369. Springer-Varlag, 2004
[30] Boyen X.: Multipurpose identity based signcryption: a Swiss army knife for identity based cryptography. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp.383-399. Springer-Varlag, 2003

Wei Yuan was born in Chengde of Hebei province of China in 1984. He began the study of computer science at Jilin University in 2003 and got his bachelor degree in 2007. Then he continued his research on information security and received his master degree in 2010. Now he is a PhD candidate of the college of computer science and technology of Jilin University.
His main research interests include cryptography and information security. he have participated in several projects include two National Natural Science Foundations of China and one National Grand Fundamental Research 973 Program of China and published more than 10 research papers from 2007.

Liang Hu was born in 1968. He has his BS degree on Computer Systems Harbin Institute of Technology in 1993 and his PhD on Computer Software and Theory in 1999. Currently, he is the professor and PhD supervisor of College of Computer Science and Technology, Jilin University, China.
His main research interests include distributed systems, computer networks, communications technology and information security system, etc. As a person in charge or a principal participant, Dr Liang Hu has finished more than 20 national, provincial and ministerial level research projects of China.

Li Hongtu was born in Siping of Jilin, China on Mar. 17 1984. In 2002, Li Hongtu began the study of computer science at Jilin University in Jilin, Changchun, China. And in 2006, Li Hongtu got bachelor's degree of computer science. In the same year, Li Hongtu began the master's degree study in network security at Jilin University. After 3 years study, Li Hongtu got his master's degree in 2009. From then on, Li Hongtu began the doctor's degree in the same field of study at the same University.
From 2009, he has got a fellowship job. He worked in grid and network security laboratory as an ASSISTANT RESEACHER at Jilin University. From 2006 to now, he has published several papers.

Jianfeng Chu, **Corresponding author**, was born in 1978, Ph.D. Now he is the teacher of the College of Computer Science and Technology, Jilin University, Changchun, China. He received the Ph.D. degree in computer structure from Jilin University in 2009. His current research interests focus on information security and cryptology.
An important objective of the projects is to probe the trend of network security, which can satisfy the need of constructing high-speed, large-scale and multi-services networks. Various complex attacks can not be dealt with by simple defense. And to add mechanisms to network architecture results in decreasing performance. In a word, fundamental re-examination of how to build trustworthy distributed network should be made.

Yuyu Sun, female, born in 1977,Lecturer, Ph.D. of Jilin University. She graduated from the Department of Computer Science and Technology of Jilin University in 2005,and obtained an MA degree. From 2008,she began to start her doctorate in computer in Jilin University, now she is working in Changchun University. Her current research interests include network and information security. She mainly engaged in Teaching and research on information security and Application software development. She has participated in one National Natural Science Foundation of China, one Major Project of Chinese National Programs for Fundamental Research and Development (973 Program), five Science and technology support key project plan of Jilin Provincial Science and technology Department , three S&T plan projects of Jilin Provincial Education Department .She has Wrote 4 textbooks as yet. She has published 14 academic articles in English and Chinese, four of that has been retrieved by EI.

An Independent Set Packet Classification Algorithm Using Priority Sorting

Rong Hui-Gui, Chen Hao
School of Information Science and Engineering, Hunan University,
Changsha, 410082, China
ronghg@163.com chenhao@hnu.cn

***Abstract*—Independent set algorithms, as a kind of packet classification algorithms with space efficiency, has lower execution efficiency for the lack of priority consideration in linear matching process. In addition, new independent sets created frequently as a result of dynamic updates greatly increase its dependence on the consumption of storage space. In order to overcome these above disadvantages, an improved algorithm based on independent sets using priority sorting (ISSP) is proposed and an improvement strategy of split rule is designed for higher storage efficiency in dynamic updates. The simulation results further show that the improved algorithm, compared with IS algorithm, reduces its dependence on storage space in dynamic updates and has higher execution efficiency.**

***Index Terms*—packet classification, independent sets (IS), priority sorting, dynamic updates**

I. INTRODUCTION

With the rapid development of network technology and the emerging network applications, Internet users are demanding more for reliability, security and diversity of the network service [1]. It is necessary for routers to provide differentiated network services to meet the needs of different users, such as packet filtering firewall, traffic accounting, differentiated services, QoS and so on. Routers should have the ability of fast packet classification to support these differentiated services. Fast packet classification algorithms have become a key technology for high-speed routers and also have been the key of avoiding the router being as the bottleneck of network performance.

Packet classification algorithms in general may be divided into two categories: one group is algorithms implemented by pure hardware implementation, such as content access memory (CAM). This group algorithms have a good lookup efficiency, but it is difficult to promote since own deficiencies (bulky, high power consumption, supporting no range type of rules) [2]; the other group is through software implementation, and they are subdivided into algorithms based on Terry tree and collection location. Xuehong Sun presents a new fast packet classification algorithm based on independent sets, in the IEEE Transaction On Networking meeting [3], and this algorithm has become the most popular and efficient packet classification algorithms in recent years.

This paper deals with the traditional problems of IS packet classification algorithms and analyzes some key factors affecting the performance of IS algorithm, then an improved algorithm (ISSP) is proposed. The approved algorithm maintains the original characteristics of IS algorithm and it solves the linear matching issues by introducing a priority-sorted mechanism since the first matching rule is the final rule after sorting rather than traverse the whole rule index table. As a result, this new algorithm raise the performance efficiency; on the same time, new independent sets created frequently for dynamic updates greatly increase its dependence on the consumption of storage space, an improvement strategy by split rule is designed for higher storage efficiency in dynamic updates.

II PROBLEM DESCRIPTIONS

Independent sets originate from the concept of independent set [4] in graph theory, in which independent sets mean the subset of vertex set, and any two vertices are not connected. The idea that get independent elements together to determine the relevance have been reflected in many mathematical models, above all, it is widely applied to different scientific areas, such as fault diagnosis, computer vision, computer networks and so on. This paper focuses on the application and expansion of independent sets in packet classification fields and uses "independent sets" to distinguish from the independent set concept in graph theory.

A. Calculation of independent sets

The first step of using independent sets to packet classification is to construct several independent sets based on rule base and each independent sets should contain as many as rules for raising the storage and lookup efficiency. The rule base of packet classification may be mapped to an undirected graph according to the overlapping relationship among rules, and then the problem of constructing independent sets based on the rule library becomes that of finding independent sets in graph theory. Solving Independent sets always is a classic

Foundation item: Project (531107021115) supported by "Fundamental Research Funds for the Central Universities"; Project (61070194) supported by the National Natural Science Foundation of China.
Manuscript received: January 1, 2011; revised June 1, 2011; accepted July 1, 2011.
Corresponding author: Rong Hui-gui, Lecturer of Hunan University, Doctor of Wuhan University; Tel: +86-731- 88828148; E-mail: ronghg @163.com

doi:10.4304/jnw.6.11.1565-1571

and complex problem in graph theory [5]. This paper uses a solution for finding independent sets based greedy algorithm which generates a maximal independent set of local optimum after once iteration, and maximal independent sets of local optimum will be constructed after a number of iterations. Fig.1 illustrates packet classification rule base represented by an undirected graph, where the left part denotes rule base, the right part is its undirected graph mapped by the rule base.

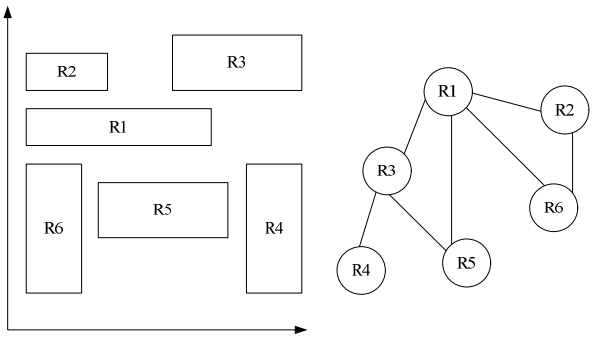


Figure 1. The rule base denoted by the undirected graph

Greedy algorithm [6] is a method with the goal to get local optimal solution instead of pursuing the overall optimal solution. Many computer algorithms have used the idea of greed, such as the knapsack problem, game theory-related issues and so on.

Given a graph G, *V* represents all of its vertices, *E* represents all the sides, *Nv (u)* represents all the vertices connected to u and S means the independent sets of G. Firstly a vertex *x* selected from G will be added to the set *S* that is initially an empty set. Then all the vertices in *Nv (x)* should be deleted from G in order to ensure that the next vertex from G is not connected with all the vertices in *S*, that is independent. The next step is to determine whether the graph G is a complete graph after *Nv (x)* being deleted. If G becomes a complete graph or only one vertex is left, then we add any vertex of G to *S* and *S* will become a maximum independent set of local optimum; if G is still not a complete graph and also have more than one vertex, then the above steps should be repeated until G becomes a complete graph or only a vertex is left. Fig.2 shows the calculation of using greedy algorithm to get the independent set.

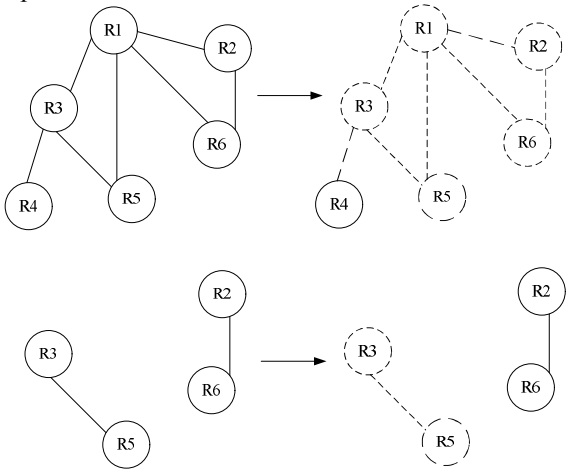


Figure 2. The process of using greedy algorithm to calculate the independent set

In Fig.2, there are six vertices in total. Firstly, add vertex R1 to the independent set S and delete R2, R3, R5 and R6 that connected with A from G, and all the deleted vertices and edges are represented with dotted lines. Only one vertex R4 is left in G after deletion, then R1 and R4 will form an independent set named I_1, thus, the first iteration is completed. At the beginning of the second iteration, vertex R1 and R4 have been removed, only the vertices R2, R3, R5 and R6 are left, then we add R3 to independent set S' and remove R6 from G, the remaining vertices R2 and R6 will form a complete graph. Then, we take any vertex together with R3 to form an independent set, if the selected vertex is R2, then R2 and R3 will form the second independent set I_2, thus, the second iteration is finished. Just like this, the third independent set composed by R3 and R6 may be found.

Although the greedy algorithm in graph theory does not guarantee that each independent set is the best overall, but will ensure the independent set of each iteration being global optimal by improving the greedy algorithm since the substantial overlapping relationship among connected vertices, in the undirected graph constructed by the rule base of packet classification.

R represents the rule base; *S* denotes independent sets, *r* means a rule and the calculation of independent sets is carried out specifically through the following two steps:

Step 1: Choose *r* with the smallest destination from *R*. If several items satisfy the condition, select one randomly to add to S and delete all the rules overlap with r from S. Then jump to step 2.

Step 2: If *R* is an empty set, then *S* is a locally optimal independent set, and the iteration is finished. If *R* is nonempty set, jump to step 1.

This method needs a pre-sort for all the rules and its time complexity is *O(nlogn)+O(n)*, where *n* is the number of rules. The method can be proved by mathematics that independent sets calculated every time is the largest independent set in current rule base.

B. Limitations of IS algorithm

IS algorithm center on using the calculations of independent sets to dispatch rules of final rule base into several independent sets. We use I_1 to denote the first independent set constructed by the rule *R*, and suppose that $R_1 = R - I_1$, which represents the remaining rules. Then, we may carry out the calculation of the independent set once again to generate I_2, and then $R_2 = R_1 - I_2$. The iteration repeats itself until R_m becomes an empty set, at last, the rules of final rule base will be stored separately in a number of independent sets $\{ I_1, I_2, \cdots\cdots, I_s \}$ and IS algorithm may be seen references[3] in detail.

For a given rule base $R = \{ r_1, r_2, \cdots\cdots, r_n \}$, the first step is to conduct a calculation of independent set based on the mentioned method in section 2.1, then use the IS algorithm to construct rule index tables for all the basic sections of B_0. Fig.3 shows the process.

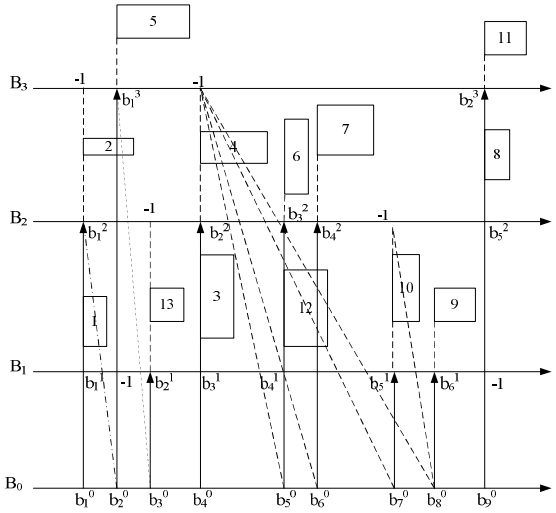


Figure 3. Constructing rule index tables of B_0.

Then, in accordance with IS algorithm, the rule index tables of all the points in B_0 is described in Fig.4:

b_9^0	-1	8	11
b_8^0	9	-1	-1
b_7^0	10	-1	-1
b_6^0	12	7	-1
b_5^0	12	6	-1
b_4^0	3	4	-1
b_3^0	13	-1	5
b_2^0	-1	2	5
b_1^0	1	2	-1

Figure 4. The rule index tables of all the points in B_0

Unlike FIS[7] algorithm to dealing with the starting point and end point of the rules, this packet classification algorithm based on independent only needs to process the starting point of the rules since no overlapped rules at the given dimension in the same set, and thus, it saved a lot of storage consumption. However, this algorithm only presents one dimensional divisions for the rule base, and as a result, there exist a larger number of rules in rule index tables of each point in B_0. After packets are located to the basic section of B_0, the algorithm should begin a linear matching for these rules, while this group algorithm needs to lookup the whole rule index table for getting the rule with highest priority as a result of ignoring the rules' priority in linear matching process, which greatly reduces lookup efficiency.

In addition, the core structure of packet classification algorithm based on independent sets consists of several independent sets, and when new rules arrive, they need to be added to the appropriate independent sets. If an independent set can meet the mutual independent condition among all rules after the new rule being added, then we add the new rule to it. If the new rule can not be added to any current independent set, IS algorithm will create a new independent set, and then add the new rule into it. Although this method is a relatively simple implementation, obviously, it greatly increases the number of independent sets and the length of the rule index tables, which greatly increased the consumptions of storage space from dynamic updates.

III PROPOSED IMPROVED ALGORITHM

From the before-mentioned discussion and analysis of IS algorithm, an improved algorithm using priority sorting is proposed in order to overcome the weakness of IS algorithm.

A. The implementation of ISSP algorithm

There usually has more than one matching rule for a packet, so we should the define the priority of the tuples to ensure the uniqueness of matching results, which means that the rule found is the highest priority [8-10]. Generally speaking, the longer the prefix is or the smaller the range is, the higher the corresponding priority is [11].

IS algorithm has lower execution efficiency due to the lack of priority consideration for rules in linear matching of data packet, in addition, new independent sets created frequently for dynamic updates greatly increase the consumption of storage space. In order to overcome these above disadvantages, an improved algorithm (ISSP) is proposed, and the detailed algorithm flow chat is described in Fig.5.

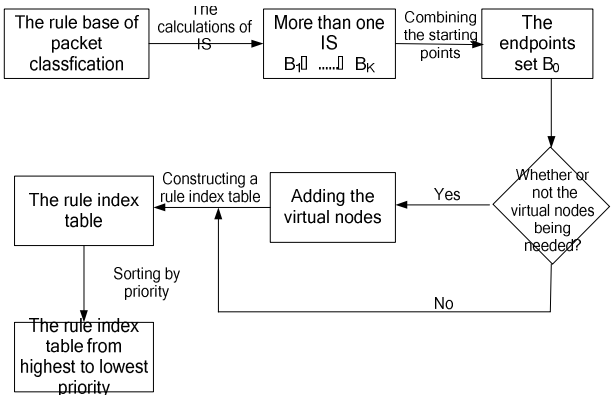


Figure 5. The detailed algorithm flow chat of ISSP

In this new algorithm, for any b_i^0 in B_0, constructing its rule index bale still needs to lookup all points of the B_k(k=1，2，……，s) to find the eligible maximum value satisfies the condition $b_p^k \le b_i^0$ for each starting point, and add the corresponding rule of b_p^k to the index table of b_i^0. When a rule being added to the index table, ISSP algorithm will conduct an insertion sort by priority for all rules exist in the index tables. Although, the rules in rule index tables are in order from highest to lowest priority, and then the linear lookup for the rule index table should only return to first rule of meeting the needs instead of traversing the entire index table.

If b_p^k is a virtual point and its corresponding rule index is given the value -1, which means no corresponding rules in B_k. Once the sorting finishes, these virtual points

will be at the end of the index table. Therefore, the lookup should stop and return the negative results when the value of -1 is met, which indicates no corresponding rules matching the data packets during searching the rule index table, since all the values of the rest indexes are -1 and have no actual rule indexes. Obviously, this improvement of filtering the index entry valued -1 is very significant when the values of -1 occupy the major positions in many rule index tables.

The concept of rule priority is basically in agreement with the longest prefix match, and it means the longer rule prefix or the smaller rule scope, the higher priority. In rare circumstances, some exceptions exist, where some rules with a larger scope have been artificially defined as a high priority, only when special traffic should be protected, such as the necessary protection of VOIP traffic to ensure the screen session smoothly in a large enterprise. Assumed that the priority assignment of the rules in Fig.3 is described as follows:

Table 1. the rules list sorted by priority

The number of rules	priority	The number of rules	priority
1	high	2	very high
3	low	4	high
5	very high	6	low
7	Very high	8	high
9	Very low	10	low
11	high	12	high
13	low		

The corresponding rule index table will become the next Fig.6:

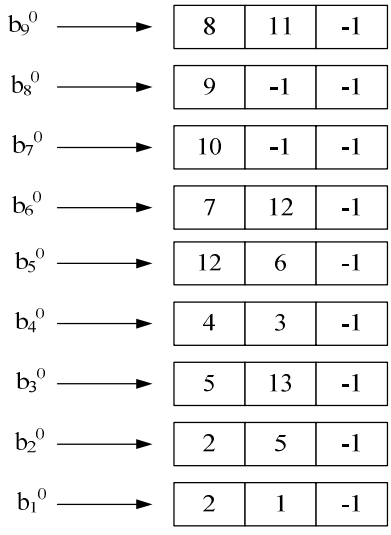


Figure 6. The rule index table after priority sorting

Obviously, ISSP algorithm may improve the lookup efficiency of rule index tables, while this improvement also undermined the overall structure of IS algorithm. In IS algorithm, the i-index in rule index table stores a rule in a independent set I_i, and after priority sorting, the rules that the *i*- index stored do not necessarily belong to I_i, only ensure that the priority of the index is larger than *i*. despite some original features of the improved structure being lost, but this improvement increases the processing speed of packet classification and only the real-time updates of rules base willget negative effect. ISSP algorithm is better than that of IS algorithm in overall if it is not required for router to support dynamic updates. Provided the need for real-time update rule base, then the original rules index table should be kept and the rule index table constructed by ISSP algorithm is also added to each point of B_0. Therefore, there are two rule index tables in each point of B_0, and one table is sorted by priority, the other maintains the original order. This strategy may increase the storage overhead, but can reduce the time-consuming, and maintaining the support for dynamic updates feature.

The core structure of packet classification algorithm based on independent sets consists of a number of independent sets. When a new rule is added to the rule library, it will be accepted finally only if the rule meets the mutual independent conditions among the rules. So, if existing independent sets are not able to accept new rules, then IS algorithm will create a new independent set for accepting new rules. Although the above method is simple to implement, but it will greatly increase the number of independent sets, at the same time, the rule index table also is enlarged. So, ISSP algorithm proposed a dynamic updates strategy by split rules. When there are no independent sets to accommodate the new rules, the new rules will be split into several sub-rules that are ensured to be added to the existing independent sets according to the actual situation of each independent set. This strategy will maximize the use of existing independent sets and the rule index table, thereby the storage space consumption caused the dynamic updates will be greatly reduced.

B. Limitations of dynamic updates in IS algorithm

b_{new} and e_{new} are used to indicate the start and end points, respectively. For determining whether new rules may be added to the existing independent sets, e_{new}, as a key value, is usually used to search the range lookup tree of B_0 and once the lookup finished, the maximum value b_x^0 less than or equal e_{new} will be returned. If the new rules are mutually independent with the rules connected to the rule *i*, in rule index table of b_x^0, then the new rules may be added to the independent set B_i, but it doesn't work conversely. If the value of the rule *i*, in rule index table is -1, then you need to compare the new rule with the former rule of B_i.

When a new rule is added to independent set, both cases may appear:

1) If the starting point of the new rule already exists in B_0, all rule index tables with $b_{new} \leq b_k^0 \leq e_{new}$ should be updated.

2) If the starting point of the new rule is not included in B_0, all rule index tables with $b_{new} \leq b_k^0 \leq e_{new}$ should be updated, then b_{new} must be added to B_0 and a rule index table should be created for it.

Based on Fig.3, Fig.7 shows the adding process of the two new rules 14 and 15. The process of the new rule 14

being added according IS algorithm is illustrated below: point b_4^0 will be returned through searching the range query tree of B_0 and the item 3 of the rule index table b_4^0 is -1, then the rule 14 will be added to B_3 since they are mutual independent by comparing the rule 14 with the previous non -1 rule namely the rule 5 in B_3. It is unnecessary for creating a new point for B_0 since the starting point of the rule 14 is b_4^0, and then the rule index table of b_4^0 should be updated from (3, 4, -1) to (3,4,14). When we add the new rule 15, the lookup tree will return the point b_7^0 through searching the range lookup tree of B_0, and then we find the item 1 of the rule index table of b_7^0 being the rule 10 which is ensured no mutual independent with the new rule 15, so we could not add rule to B_1. The item 1 of the index table is -1, and then the rule 15 should be compared with the previous non -1 rule namely the rule 7 in B_2, since they are not mutual independent, so the rule 15 could not be add to B_2 here; next, the item 3 of the index table is -1, so the rule 15 should be compared with the previous non -1 rule namely the rule 14 in B_3. The rule 15 will eventually be added to B_3 since they are mutual independent. Finally, in Rule 15, the starting point b_{new}, with no new endpoints included in B_0, should be added to B_0, and then we should create a rule index table (-1,7,15) and also update the index table b_7^0 (10, -1, 15).

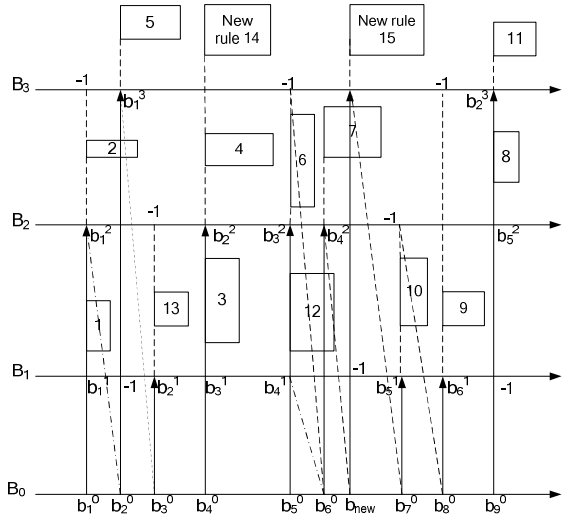


Figure 7. The structure graph of IS with new added rules

The new rules 14 and 15 may be added to existed independent sets because of finding the appropriate location, while some rules can't be added due to the overlap of range, such as the rule16 in Fig.8. When this happens, you can use a simpler approach that create a new independent set and add the new rules into it. The disadvantages of this above-mentioned strategy lie in having a relatively large memory consuming that needs space to store a new set and also increases the length of the rule index table of all points in B_0.

C. Dynamic updates strategy of ISSP

ISSP algorithm will present a new dynamic update strategy through splitting the new rules into several sub-rules will be ensured to be added to the existing independent sets.

Fig.8 is a sketch diagram of adding new rules by applying division strategy based on Fig.7. Since the range of the new rule 16 may overlap with that of rule 9 and rule 10 in B_1, overlap with rule 7 in B_2 and overlap with rule 11 in B_3, so, the new rule 16 can not be added to the existing independent set B1, B2 and B3. The dashed part of Fig. 8 means constructing a new independent set B_4 for the rule 16, and obviously the new independent set contains only one rule. After the constructing B_4, the length of rule index in all points of B_0 should increase 1 and the original 3 items are increased to 4 items.

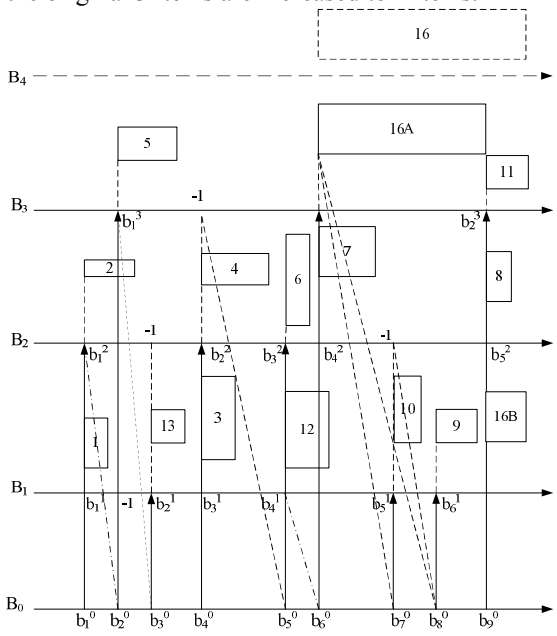


Figure 8. The add sketch based split rules

Assumed that the coverage area of rule 16 is $[begin_{16}, end_{16}]$, in Fig.8, $begin_{16} = b_6^0$. It's not difficult to find that the rule 16 may be split into two sub-rules $[b_6^0, b_9^0]$ and $[b_9^0, end_{16}]$ denoted as 16A and 16B respectively, where 16A may be added to independent set B_3 and 16B may be added to independent set B_1. Then the addition operation will be finished only by updating the rule index table of b_6^0, b_7^0, b_8^0 and b_9^0 with (12,7,16), (10, -1,16), (16,8,11). This strategy may make full use of existing space of independent sets, instead of creating new independent sets for storing one new rule, and the addition operation is a bit complicated. IS algorithm only needs to traverse the rule index table of one point in B_0, while the division rules need to traverse the rule index tables of all points from the range of $[begin_{16}, end_{16}]$ in B0

Compared with the addition operation, the deletion process only needs to traverse the endpoints of rule index table in deletion range in B_0 and the corresponding rule is assigned to -1, obviously is a bit simple. However, the independent set may no longer be the largest independent set after deletion, which means that the remaining rules of independent set with some rules removed may be added to other independent sets. In Fig.9, the dotted lines denote the rules to be deleted, and with the rules 1, 4 and 5 removed, we easily find the rule 2 and rule 7 in B_j can be

added to the independent set B_i, then they form a maximum independent set together with other rules in B_i.

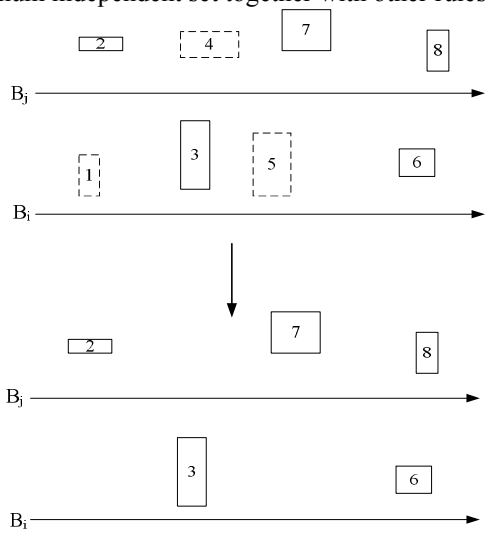


Figure 9. The sketch of the rules deletion

Notice that rebuilding an independent set cost a lot, and the rule deletion and addition are bound to exist simultaneously under the dynamic updates [12-15], so we should not cling to constructing the largest independent sets, and the vacancy with the removed rules is naturally supplemented by the new rules.

Therefore, the improved algorithm realizes the dynamic updates by splitting rules to maximize the existing independent sets and rule index tables, as a result this reduce the consumption on storage space from dynamic updates, greatly.

Ⅳ SIMULATION EXPERIMENT

Simulation platform, whose runtime environments include Pentium 4 3.06G CPU, 512MB RAM and Windows XP operating system, is programmed by C++. The rule sets adopted by simulation experiment is not randomly created but is derived from the rule tables of core routers in the real computer network (data sources：CAIDA）. Because the length of the rule index table determine the maximum number of linear matches, and the value is consistent with the number of independent sets that determined by the scale of the rule library, therefore, this article conducts some granularity analysis based on the scale of rule base (from tens of thousands to hundreds of thousands). The simulation experiments are conducted based on the rule base with a large order of magnitude since packet classification algorithms in high-speed networks should support classification rules (in millions). When the magnitude variable a of the rule base, respectively is valued 319337，96371，47096，20828, the performance of dealing with data packets in different orders of magnitude (from one million to ten million) is discussed and described between IS algorithm and ISSP algorithm. The simulation experiment is just a preliminary examination for algorithm performance with less consideration to the actual environmental factors in real networks, and the packets used in experiments are randomly generated.

Simulation program consists of several components: the rule base pretreatment of packet classification, whose main task is to extract rules' information required by algorithms and translate IP address in the form of character string into a digital representation, for instance, IP address: 202.103.96.1, its corresponding binary code is 1,100,011,001,100,111 0,110,000,000,000,001, we can convert this into a binary integer code 3395772417. The second component is the construction of independent set, whose main work is to divide the rule base into several independent sets in accordance with IS algorithm, and then generate a rule index table for each point in B_0 and sort every rule by priority. The third part is to construct the balance tree, and generate the balance lookup tree of scope for all the basic section of B_0.

The results take the running time as the contrast parameter in seconds (S), vertical axis dictates the running time, and the abscissa dictates the number of the packet (million).

When the scale of the rule base reaches a= 319337, the simulation results are showed in Fig.10:

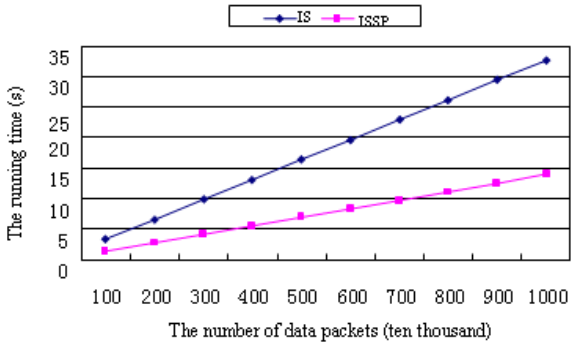


Figure 10. The simulation results of the rule base scale: a= 319337

When the scale of the rule base reaches a= 96371, the simulation results are showed in Fig.11:

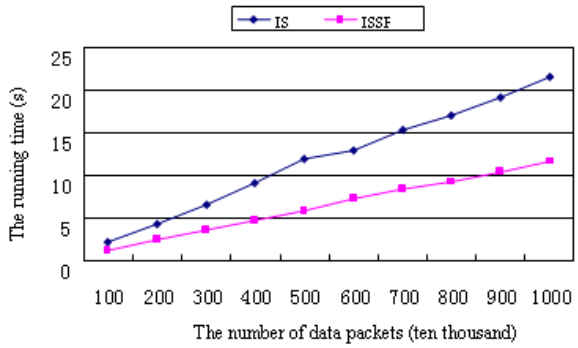


Figure 11. The simulation results of the rule base scale: a= 96371

When the scale of the rule base reaches a= 47096, the simulation results are showed in Fig.12:

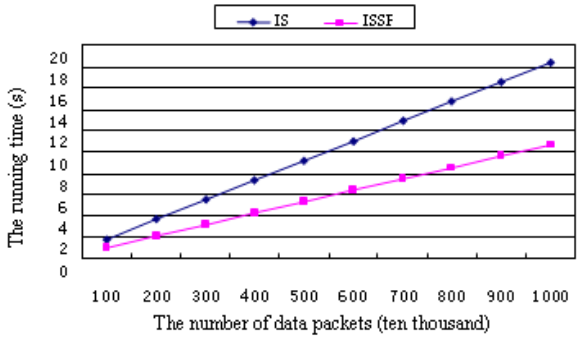


Figure 12. The simulation results of the rule base scale: *a*= 47096

When the scale of the rule base reaches *a*= 20828, the simulation results are showed in Fig.13:

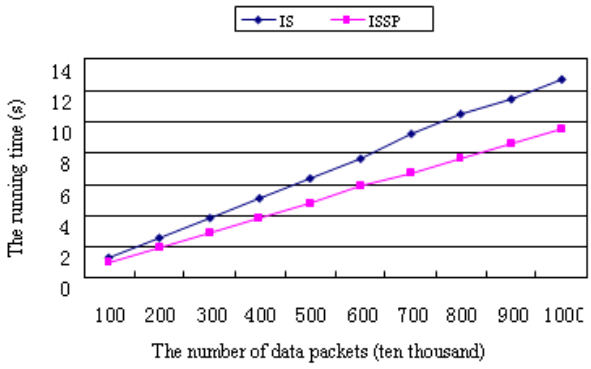


Figure 13. The simulation results of the rule base scale: *a*=20828

From the simulation results: ISSP algorithm is better than IS algorithm in running time, moreover, the larger the size of rule base, the more obvious of the advantages. As the rule base increases, the running time may be saved 20%-50% because ISSP algorithm may determine the highest priority rule matching the data packet with only a small amount of linear searches, instead of traversing the whole rule index table. While IS algorithm only determine the adopted rules after traversing the full rule index tables, and the larger of the size of rule base, the more time-consuming of the traversal. Overall, compared with IS algorithm, ISSP algorithm can run more quickly in the large magnitude rule base.

Ⅴ CONCLUSIONS

This paper analyzed the factors influencing the performance of IS algorithms and proposed an improved IS algorithm. This new algorithm maintains the original characteristics of IS algorithm instead of traversing the whole index table, as a result, the linear matching process is greatly shortened. At the same time, it analyzes the shortage that new independent sets created frequently from dynamic updates greatly increase its dependence on the consumption of storage space, and proposed an improvement strategy of split rule for higher storage efficiency in dynamic updates. The simulation results show that the improved algorithm is more efficient in running time and the split rules increase the storage efficiency in dynamic updates.

ACKNOWLEDGMENT

I would like to acknowledge the wonderful work of our team for this paper. This research is supported by the “Fundamental Research Funds for the Central Universities” (531107021115); Project (61070194) supported by the National Natural Science Foundation of China.

REFERENCES

[1] WANG Yong-gang, SHI Jiang-tao, DAI Xue-long, YAN Tian-xin. Simulated Testing and Comparison of Algorithms for Packet Classification[J]. Journal of University of Science and Technology of China, 2004，34（4）：400-409. (in Chinese)

[2] S. Zezza, E. Magli, G. Olmo, and M. Grangetto, "SEACAST: a protocol for peer-to-peer video streaming supporting multiple description coding," in Proc. of ICME 2009, New York, USA, Jun. 28--Jul. 3, 2009.

[3] Sun X H, Sartaj S. Packet Classification Consuming Small Amount of Memory. IEEE Transaction On Networking, 2005, 13(5): 1135-1145.

[4] Karp R, Wigderson A. A fast parallel algorithm for the maximal independent set problem. Journal of the ACM , 1985, 32(4): 762-773.

[5] S. Milani and G. Calvagno, "A Game Theory Based Classification for Distributed Downloading of Multiple Description Coded Videos," in Proc. of the IEEE ICIP 2009, Cairo, Egypt, Nov. 24--28, 2009.

[6] Chazelle B, Guibas L J. Fractional cascading I: A data structuring Technique. Algorithmica, 1986, 1(2): 133–162.

[7] [Geraci F, Pellegrini M, Pisati P, et al. Packet classification via improved space decomposition techniques. In Proc of INFOCOM. Miami, 2005, 13-17.

[8] SUN Yi,LIU Tong, CAI Yi-bing, HU Jin-long, SHI Jing-lin. Research on Packet Classification Algorithm [J]. Application Research of Computers，2007，24（4）：5-11. (in Chinese)

[9] Haoyu Song, Fang Hao, Murali Kodialam, T.V. Lakshman, IPv6 lookups using Distributed and Load Balanced Bloom Filter for 100Gbps Core Router Line Cards, INFOCOM, 2009.

[10] V. Pus and J. Korenek. Fast and scalable packet classification using perfect hash functions. In FPGA '09: Proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays, pages 229--236, New York, NY, USA, 2009. ACM.

[11] Gupta P, Lin S, Mckeown N. Routing Lookups in Hardware at Memory Access Speeds. In: Proc of IEEE INFOCOM. San Francisco, 1998, 1240–1247

[12] Song H Y, Jonathan T, John L. Shape Shifting Tries for Faster IP Route Lookup. In Proc of ICNP. Boston, 2005, 358-367.

[13] Kamath P, Lan K C, Heidemann J, et al. Generation of high bandwidth network traffic traces. In: Proc of International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems. FortWorth, 2002, 401-410.

[14] Song H Y, Jonathan T, John L. Shape Shifting Tries for Faster IP Route Lookup. In: Proc of ICNP. Boston, 2005, 358-367.

[15] A. G. Alagu Priya, Hyesook Lim. Hierarchical packet classification using a Bloom filter and rule-priority tries[J]. Computer Communications, 2010, 33(10): 1215-1226

Enabling Awareness Driven Differentiated Data Service in IOT

Haoming Guo
Beihang University Computer School, Beijing, China
Email: guohm@nlsde.buaa.edu.cn

Shilong Ma and Feng Liang
Beihang University Computer School, Beijing, China
Email:{slma, Fengliang} @nlsde.buaa.edu.cn

***Abstract*—IOT needs to organize sensor resources to monitor events in real world for all time. As result, huge number of data will be concentrated in the system. Due to difference of sensors' awareness on event, the concentrated data's qualities are different. The data process task's performance will be affected without discrimination of data's quality. This paper introduced an approach, called Awareness Driven Schedule (ADS) that enables involved sensor resources to provide differentiated data service by their awareness, to address the issue. In the approach, higher a sensor resource's awareness on the event is, more detailed data service it should provide. Requirement that specify relation of sensors' awareness and rules of data collect job will be submitted initially. Constant and continuous data channels are created to organize sensors. In accordance of their awareness and task initial requirement, sensor resources are scheduled to collect data and aggregate to task through the channels. Sensor resource's involvement and service depend on it's awareness and task's requirement. Upon ADS, a middleware is built for CEA's (China Earthquake Administration) SPON(Seismological Precursors Observation Net) and applied for China's earthquake research applications. In the applications, Dull data of low awareness sensors could be banned out, applications may by more efficient.**

***Index Terms:* Sensor; Resource Schedule; IOT; Web Service**

I. INTRODUCTION

In IOT (Internet of things) systems, there are thousands of sensor resources deployed all over the areas[1]. Upon web technology, the sensor resources are accessible any where and anytime. The primary work of sensor resource is to monitor environment around and collect data. Through the data, system can be aware of event's development and implement related proceedings. By the approach, IOT is constructed as new frontiers between human and real world[2][3]. Sensor can constantly and continuously provide data service for applications[4]. The mechanism, however, may lead to problems that affect system's data processing performance.

For example, in earthquake application: Earth Viberation Detect(EVD). EVD's goal is to catch exceptional vibration exactly. For the purpose, a large number of sensor resources are spread all over the area. Once exceptional vibration takes place, the sensor nearby may catch the signal. A number of sensors' data are aggregated by EVD to find out detailed information about the event.

In EVD's case, sensor resource has two distinguished features: 1.Data's quantity is huge. 2.Data quality is changing. In IOT, sensor is to watch real world constantly. It generates data continuously as it works. As result, a large number of data may be concentrated for further processing. Meanwhile data's accuracy is affected by its working conditions and relations with event. As target event changes, data accuracy may change either. Some data may be highly valuable while others may be dull to application. Because data's quality is not stable as the one in conventional web, resource's service to task should be differentiated.

The goal of sensor schedule is to enhance differentiated continuous services and dynamic resource involvement in task in accordance with resource's data quality and its awareness. In this schedule approach, task publishes data requirements. All resources are involved in accordance with whether they could provide required data. Meanwhile, layered data channels are built for tasks. Resources link to corresponding data channel to transfer data with different frequency. For example, for EVD, higher accurate the data is, more frequent it is to be transferred. If one resource's data accuracy changed, it cuts off current data channel link and recreate new link with corresponding level channel. If resource could not provide required data, it quit from the working group. Once a data's data accuracy reached task's requirement, it creates connection with related data channel and provide data to task. Through this awareness driven schedule, resources are organized in accordance with their leveled data so that unnecessary huge data transfer may be reduced while guarantee credible and continuous data service for tasks.

II. RELATED WORK

Derived from the traditional sensor network, Sensor Web is now widely used in fields such as Bio-complexity mapping of the environment, Military applications [5], flood detection[6] , traffic management[7] and etc.

doi:10.4304/jnw.6.11.1572-1577

In traditional sensor web research, coverage is one primary issue. It concerns with problem of how to optimize schedule policy to improve energy efficiency and guarantee sensor network's coverage under requirement of performance. Paper[8] introduced an approach, called Coverage Configuration Protocol(CCP), to address the issue by analyzing sensors' connectivity and coverage relations. In the approach, the policy that rules whether a sensor will be activated is set to be whether the sensor is in an area not covered by other activated sensor. Upon CCP, paper[9] introduced an algorithm to reduce unnecessary activated sensor nodes while avoiding blinding. The research introduced above address issues to create possible coverage by least sensors under requirement of energy conservation. As sensors work continuously, sensors involvement of providing data services will be dynamic in the coverage net. Paper[10] introduce an approach to address continuous time sensor scheduling problem in which part of involved sources are to be chosen to collect data at each time point. In the approach, the sensors that are chosen at a particular time are represented by controls. The control variables are constrained to take values in a discrete set, and switchings between sensors can occur in continuous time.

The researches concerns with how to organize least sensors to provide measurements in dynamic. However, the measurements of activated sensors are processed equally. All sensors are viewed as data producer with same accuracy and quality.

In 2005, the OGC (Open Geospatial Consortium) has proposed a new sensor web integrated framework: SWE (Sensor Web Enablement) [11], which has become the De facto standard in industry. SWE adopts SOAP and XML from Web Service Architecture and aims at a unified management of the heterogonous sensor resources via Internet, including discovery, access, controlling and notification with the plug-and-play feature.

Conforming to the SWE standard, NICTA Open Sensor Web Architecture (NOSA) is a software infrastructure aimed at harnessing massive computation power of grid computing in sensor web. The core middleware includes planning, notification, collection and repository services. By splitting the information sensing and processing, it harnesses the Grid Services to process the information, which not only greatly reduces the load of sensor network, but also simplifies the heterogeneous sensor network management. The architecture allows the data interoperability. However, the quality of data is not considered as a standard way for result processing.

Resource scheduling is a NP-complete question [12] in distributed systems, therefore up to now only locally optimal solution is available. According to the mentioned research above, it can be seen most of the information-driven middleware implements the scheduling by processing the data indiscriminately, this scheduling mechanism is suitable for information browsing activities, but inefficient for emergencies which involves variety of parties and monitors distributed dynamic event sources because of longer processing time and more resource. Therefore a sophisticated and effective mechanism is required for data filtering and scheduling.

III. Design of Awareness Driven Schedule

ADS's goal is to enable differentiated continuous services and dynamic resource involvement in task in accordance with resource's data quality and its awareness. In ADS, tasks register awareness requirement(AR) to Awareness Requirement Registration(ARR) and create Task Awareness Schedule(TS) in Task Awareness Scheduler Manager(TSM). In TS, data channels are defined and built for data differentiation and service forwarding. In the AR, information type and data value definition are listed. TSR searches all resources who can provide same data as defined by AR and invoke. Resources create task object handler(TOH) in local task object handler pool(THOP). Once a resource is aware of target event, it collects data and check data channels' definition from related TS. Data transfer frequencies are listed in TS's data channel definition. Resource retrieves the frequency information by which it transfers data. In TS's data channel definition, data process services are defined. All data from one channel is about to forward to the specific service. If resource's data value shift from one range to another, TS may link related data channel to resource and reassign data transfer job. If resource lost awareness of event, it cut off link to TS. If a resource finds the event, it checks TSR with the event information and retrieve related AR upon which links to TS are built. Through this approach ADS's goal is rtealized. The whole view of ADS is shown as below.

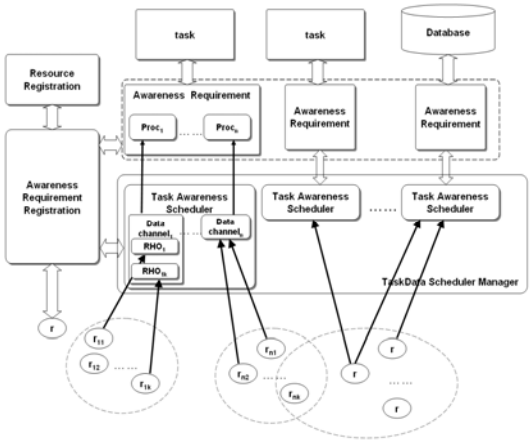


Figure 1. Whole view of ADS.

A. Definitions in ADS

Definition 1: Awareness requirement (AR). AR is task's awareness requirement definition. In AR, task specifies requirements of sensor resources to pool as working group and task's data channel definitions. It's definition is shown as below:

AR=(ID, taskID, resPTable, dataChannelList)

(1) ID is the AR's identity;
(2) taskID is current task's identity
(3) resPTable=$\{p_i|i=1,2,..n\}$, it specifies what kind of

resource task needs. It consists of a table of property. p_i =(name,type,value) name is property's name. type is property's type. While looking for resources from resource registration at the beginning, if AR's all property requirements matche one resource's properties, the resource may be included in task's initial working group.

(4) dataChannelList={dcf_i| i=1,2,……n}. dataChannelList is to specify data channels link between resource and related data process in accordance with resource's awareness or its data accuracy.

dcf=(ID, taskID, proc, maxValue, minValue, frequency, transMod, cacheSize);

proc is target data process in task to process the data with required accuracy.

maxValue and minValue are to define range of the channel. If a resource's data is within the range, the resource will be linked to the channel and the resource will send data by the frequency.

transMod = {" flow" , "periodic"}. It specifies by which way the involved sensor resources transfer data.

Definition 2: Resource Handler Object(RHO)

RHO is for task to receive data from corresponding sensor resource. Once a sensor resource is invoked, a RHO will be created and pooled in related data channel. RHO is defined as below:

RHO=(ID, taskID, resBinding, dataCache);

(1) ID is the RHO's identity. Through ID, RHOPool seek and retrieve the object.

(2) taskID is to maintain RHO's hosted task identity.

(3) resBinding is to specify the binding information. Through the information, RHO may redirect commanding messages to right resource.

(4) dataCache ={$value_i$ | i=1,2,..n}. it's used to cache data collected by resource by time order.

Defination 3: Task Handler Object(THO)

THO is created by sensor resource for data service request. In accordance with data quality, THO collects data and send data back to paired RHO. If its data range changed, RHO may change from original pool to other data channel's pool and the new channel's information is forwarded to THO to adjust its data collection job. THO's definition is shown as below:

THO ={ID, taskID, RHOID, dataCache, dcf}

(1) ID: is THO's identity.

(2) taskID is to maintain THO's hosted task identity.

(3) RHOID is paired RHO's identity.

(4) dataCache's definition is same as RHO's.

(5) dcf's definition is same as AR's

Definition 4: Task Awareness Scheduler(TS)

TS is to keep contact with resources, receive and forward data to right data process object in task and schedule resource's service. Its definition shown as below:

TS= {ARID, dcs}

(1) ARID is corresponding with AR'ID. One AR has one TS created.

(2) dcs is data channel list in TS. It's consist of a group of data channel: dcs={dc_j| j=1,2,….m };

dc= {ID, dcf ID, RHOPool}

ID is the data channel's identity, dcfID is data channel's definition identity through which data channel may retrieve information from corresponding AR. RHOPool is pool of resource handler object(RHO).

B. Registration of Task's Awareness Requirement

Awareness Requirement Registration(ARR) is to organize task's awareness requirement(AR) through which resource could involve into task's working group. Once task's AR is registered, its TS will be created in Task Awareness Scheduler Manager(TSM). TSM is consist of a group of TS as: TSM={TS_i|i=1,2,….n}.

After AR's TS is created, ARR searches for all resources corresponding to AR's "resPTable" specification. The result resources of the search are organized as initial working group of task. All resources in the working group would be invoked to create link with AR's TS.

The whole AR registration is shown as below:

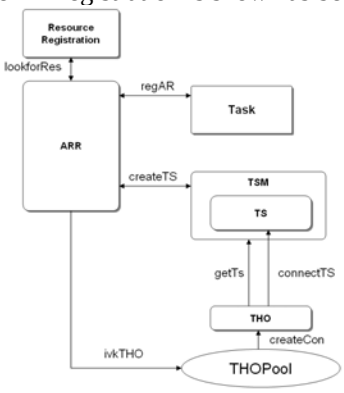


Figure 2. Communication for AR registration.

$$RegProc = Task.\overline{regAR} < AR > .ARR.\tau.$$
$$\overline{createTS} < AR > TSM.\tau.createTS(TSID)$$
$$ARR.\overline{lookFor} < resPTable@AR >$$
$$RR.lookFor(resList)ARR.invokProc.0$$
$$invokeProc = ARR.\overline{ivkTHO} < TSID > THOPool.$$
$$\tau.\overline{createCon} < TSID > THO.$$
$$\overline{getTS} < TSID > TSM\tau.getTS(TS).$$
$$createConnProc.0$$

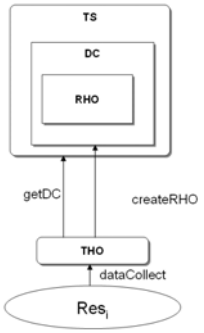


Figure 3. Communication for TS registration.

$$createConnProc = THO.\overline{getDC} < dataCache > TS.$$
$$getDC(DC).\overline{createRHO} < resBinding > DC.$$
$$createRHO(RHOID).0$$

C. Data Linkage for Resource and Task Process

After data channel connection, the resource needs to get relevant data channel's information. By the information, resource retrieves data channel's definition dcf from ARR and. In dcf, data channel's data range is defined and the resource transfer data by rules of dcf while data is within the range. Resource's THO return data to task's data channel first. Data channel forward the data to related RHO. RHO looks for data channel's definition from ARR and retrieve process object of task which is persisted in dcf. Then RHO transfer the data to the process object.

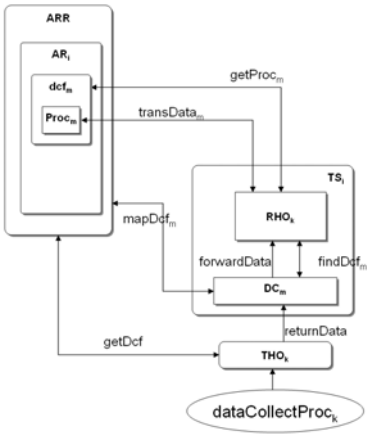


Figure 4. Communication for data collection task.

$$serProc_k = THO_k.\overline{getDcf} < dcfID@DC >$$
$$ARR.getDcf(dcf).$$
$$dataCollectProc_k.THO_K.$$
$$\overline{returnData_k} < dataCache_k > DC.$$
$$forwardData_k < dataCache_k > RHO_k.$$
$$TaskDataProc_k.0$$
$$TaskDataProc_k = RHO_k.\overline{findDcf_m} < dcfID > .DC_m.$$
$$\overline{mapDcf_m} < dcfID > ARR.$$
$$\overline{mapDcf_m}(dcf_m).\overline{findDcf_m} < dcf_m > .$$
$$RHO_k.\overline{getProc_m} < NULL > dcf_m.getProc_m(Proc_m).$$
$$RHO_k.\overline{transData_m} < dataCache_k > .0$$

In the in task's awareness requirement, task's process object is persisted. During implementation, THO transfers data back to RHO through it working data channel. RHO looks for the process object linked to hosted data channel's definition and forward data to it. In ADS, resource may provide data service constantly by this approach.

Resources are to monitor real world's event and collect data. In ADS, resource's data collection job is ruled by its linked data channel. In data channel's specification, maxValue and minValue are to define current data channel's range. If one resource's collected is within the range, it keeps data collection for current data channel. Otherwise, resource looks for new channel in current TS and collect data by the new one's rule. In data channel's definition, transMod is defined as "flow" or "periodic". If one data channel is defined as "flow", the linked resources should cache all data of the ruled intervals which will be transferred back by the frequency. If one data channel is defined as "periodic", the resources calculate average value of the ruled intervals and only the average value will be transferred back instead of whole data cache.

D. Resource Awareness Orientation

With development of monitored event, resource's awareness may change. In data channel's definition, maxValue and minValue are to define current data channel's range. Once a resource's collected data is out of current data channel's range definition. It may check ARR for new oriented data channel and shift related RHO from old hosted data channel to the new one. If resource lost awareness of the event, it will be removed from data channel. During this process, the ROH shift request message is defined as:

changeDCReq=(ID, RHOID, resBinding, taskID, oldDCID, newDCID)

In the request message, RHOID is related RHO identity. TS retrieves the object through the identity. resBinding is information about resource. taskID is current task's identity through which locates related TS. oldDCID is current linked data channel's ID and newDCID is the new data channel to link. The RHO remove request message is defined as : removeRHOReq=(ID, RHOID, resBinding, taskID, oldDCID)

The process is shown as below:

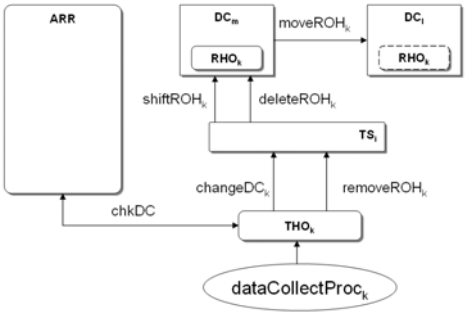


Figure 5. Communication for awareness shift.

$$AwChangeProc_k = THO_K.\overline{chkDC} < dataCache > ARR.$$
$$chkDC(chkResult).$$
$$((chkResult = NULL).\overline{removeROH_k} < removeROHReq_k > . \text{ If a}$$
$$TS_j.\overline{deleteROH_k} < RHO@removeROHReq_k > .\tau.0 +$$
$$(chkResult = dcf_l).\overline{changeDC_k} < changeDCReq_k > .TS_j$$
$$.\overline{shiftROH_k} < changeDCReq_k > DC_m.$$
$$moveRHO_k < RHO_k > .DC_l\tau.0)$$

resource begins to sense the event, it checks ARR with the data and it's own property for tasks which are require the data from ARR. ARR may return a list of available tasks' AR. The resource create connection with the tasks' TS and begin to provide data service. The check message is defined as:chkTaskReq=(ID, dataCache, pTable) ;the ARR returned message is defined as: tskResp=($TSID_1$, $TSID_2$,……$TSID_n$)
the process is shown as below:

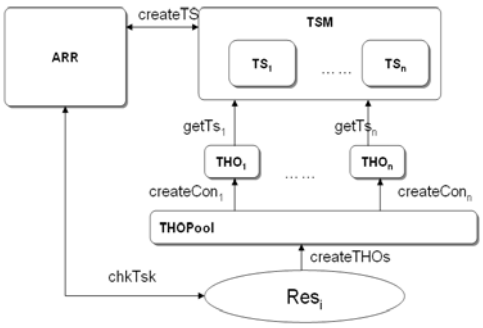


Figure 6. communication for service shift.

$$chkNewTsk = dataCollectProc.\overline{chkTsk} < chkTaskReq > ARR.$$
$$chkTsk(tskResp).\overline{createTHOs} < tskResp > .THOPool.\tau.$$
$$(createCon_1 < TSID_1 @tskResp > THO_1.\overline{getTS_1}$$
$$< TSID_1 @tskResp > TSM\tau.getTS_1(TS_1).createConnProc_1.0|$$
$$createCon_2 < TSID_2 @tskResp > THO_2.\overline{getTS_2}$$
$$< TSID_2 @tskResp > TSM\tau.getTS_2(TS_2).createConnProc_2.0|.....$$
$$createCon_n < TSID_n @tskResp > THO_n.\overline{getTS_n}$$
$$< TSID_n @tskResp > TSM\tau.getTS_n(TS_n).createConnProc_n.0)$$

IV. Application and Test

Upon ADS, a Seismological Sensor Resource Data Service System(SSRDSs) is built for CEA's Seismological General Scientific Data Platform(SGSDP) built for SPON.

In test, 12 resources are deployed to simulate application environment. The resource's data collection working frequency is about 60Hz. Two tasks were implemented for comparison. Task 1 collected all data directly from resource without discrimination. Task 2 created 4 data channel and resources provided differentiated data services. During implementation, 12 resources transferred data back to No1 task at about 360 data per second. In task 2, resources transferred at about 168 data per second. For task 2, data load was 47% of task 1. Data lost may lead to certain accuracy lost. Figure 7 shows two task's data aggregation curve. Task 2's result's accuracy is lower than task 1's. However, it was within application's accuracy requirement.

The test above shows effectiveness of ADS for data concentrated applications of IOT. Through ADS, task may organize resources to provide differentiated data service on their awareness capability that enable applications gain data within requirement of accuracy and reducing unnecessary dull data's burden.

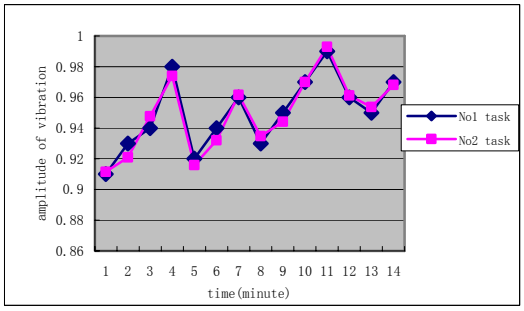


Figure 7. Comparison of ADS and conventional.

Task may gain higher accuracy by adjust data channel's setting. In another test, 500 sensor resources are organized to simulate real application of EVD. In test, by different data channel setting, application can gain different quality data. The graphic below shows details about data process node's load with different data accuracy.

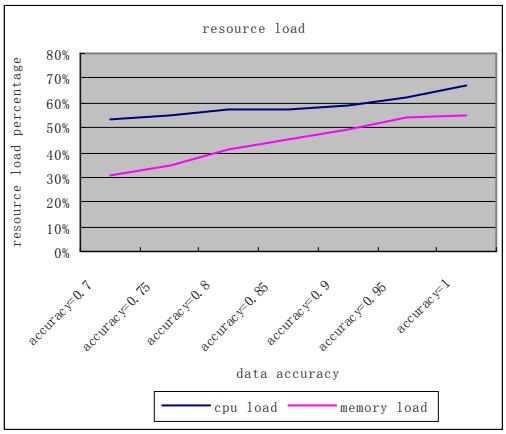


Figure 8. Resource load of ADS.

V. Conclusion and Future Work

This paper introduced an approached called (Awareness Driven Schedule)ADS. Through task's requirement, ADS organize all involved resource and enable them to provide differentiated data service to task. Higher a resource's awareness is, more detailed data it should collect and transferred. As result, low awareness of resources only need to provide limited and periodic data service. Resources' data services are differentiated with their awareness. Dull data are banned out.

In real world, event may be changing so that a task's awareness requirement over resources could be dynamic. With development of event, data channel's range and rule should change accordingly to ensure better surveillance work. As result, Awareness requirement's data channel that links process object and resources need a forecasting approach to capture change tendency of data and help task to readjust its requirement. Currently, ADS leaves the work to task. In future work, research should be done in this field to provide better data service for task.

VI. Acknowledgement

This research work was supported by China

Earthquake Administration's program for Seism-Scientific Research "Research in Online Processing Technologies for Seismological Precursory Network Dynamic Monitoring and Products" (NO. 201008002)

REFERENCES

[1] Hakima Chaouchi:The Internet of Things: Connecting Objects, Wiley-ISTE,2010
[2] Lu Yan, Yan Zhang, Laurence T. Yang, Huansheng Ning: The Internet of Things: From RFID to the Next-Generation Pervasive Networked Systems (Wireless Networks and Mobile Communications) Auerbach Publications,2008
[3] Hu W, Bulusu. N. Chou, C. T, Jha. S, "Design and evaluation of a hybrid sensor network for cane toad monitoring".ACM Trans. Sen. Netw., ACM, vol.5, pp.1-28, 2009
[4] J. Schelp and R. Winter, "Business application design and enterprise service design: a comparison", International Journal of Services Sciences, vol. 1, pp. 206--224, 2008.
[5] Akyildiz. I. F, Su.W, Sankarasubramaniam. Y and Cayirci. E, "Wireless sensor networks: a survey.Computer Networks", Computer Networks , Vol.38(4), pp.393-422, March 2002
[6] Faradjian. A,Gehrke. J and Bonnet. P, "GADT: A Probability Space ADT for Representing and Querying the Physical World". Proceedings of the 18th International Conference on Data Engineering, pp. 201-211, 2002
[7] Shih.E, Cho. S.-H, Ickes.N, Min. R , Sinha. A,Wang. A and Chandrakasan. A, "Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks" MobiCom '01: Proceedings of the 7th annual international conference on Mobile computing and networking, ACM, pp. 272-287, 2001
[8] Xiaorui. Wang, Guoliang. Xing, Yuanfang Zhang, Chenyang Lu, Robert Pless and Christopher Gill, "Integrated coverage and connectivity configuration in wireless sensor networks", SenSys '03 Proceedings of the 1st international conference on Embedded networked sensor systems , ACM, pp. 28-40, 2003
[9] Yuheng Liu, Juhua Pu, Shuo Zhang, Yunlu Liu and Zhang Xiong, "A Localized Coverage Preserving Protocol for Wireless Sensor Networks", Sensors,vol 9(1), pp. 281-302, 2009
[10] H. W. J. Lee, K. L. Teo and Andrew E. B. Lim, "Sensor scheduling in continuous time", Automatica,vol 37(12), pp. 2017-2023, 2001
[11] Botts. M, Percivall. G, Reed. C and Davidson. J, "Sensor Web Enablement: Overview And High Level Architecture".OGC White Paper, OGC, pp. 07-165, 2007
[12] D. Fernandez-Baca, "Allocating Modules to Processors in a Distributed System". IEEE Transactions on Software Engineering, vol.15, pp.1427-1436, 1989.

Haoming Guo, male, born in 1978. Post Doctor the College of Mathematics and Systems Engineering), Beihang University. His main research interests include grid, computation model in network, Data Integration, and IOT Application, etc.

Shilong Ma, male, born in 1953. Professor and PhD supervisor of the College of Computer Science and Technology, Beihang University. His main research interests include grid, computation model in network, and logic and behavior in computing, etc.

Feng Liang is currently a Ph. D student in National Laboratory for Software Development and Environment, Beihang University with the research interests in grid computing and cloud computing.

Enhancement of an Authenticated 3-round Identity-Based Group Key Agreement Protocol

Wei Yuan
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: yuanwei1@126.com

Liang Hu
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: hul@jlu.edu.cn

Hongtu Li
Department of Computer Science and Technology, Jilin University, Changchun, China
Email: li_hongtu@hotmail.com

Jianfeng Chu
Department of Computer Science and Technology, Jilin University, Changchun, China
Corresponding author, Email: chujf@jlu.edu.cn

Yuyu Sun
College of Computer Science and Technology, Jilin University, Changchun 130012, China,
Software Institue, Changchun University, Changchun 130022, China
E-mail: sunyy@ccu.edu.cn

***Abstract*—In 2008, Gang Yao et al. proposed an authenticated 3-round identity-based group key agreement protocol, which is based on Burmester and Desmedt's protocol proposed at Eurocrypt 94. However, their protocol can only prevent passive attack. If the active attack is allowed, the protocol is vulnerable and an internal attacker can forge her neighbor's keying material. It is obvious that the protocol do not achieve the aim of authentication. In this paper, we discuss the flaws to attack this protocol and propose an enhanced provably-secure protocol based on their protocol. At last, we make a detailed security analysis of our enhanced authenticated identity-based group key agreement protocol.**

***Index Terms*—authentication, identity-based, key agreement, bilinear pairing, cryptanalysis, attack**

I. INTRODUCTION

Secure and reliable communications [1] have become critical in modern society. The centralized services such as file sharing, can be changed into distributed or collaborated system based on multiple systems and networks. Basic cryptographic functions such as data confidentiality, data integrity, and identity authentication are required to construct these secure systems.

Key agreement protocol [2] [3] [4] allows two or more participants, each of whom has a long-term key respectively, to exchange information over a public communication channel with each other. However, the participants can not ensure others' identity. Though Alice wants to consult a session key with Bob, Alice can not distinguish it if Eve pretends that she is Bob. The authenticated key agreement protocol overcomes this flaw and makes unfamiliar participants to ensure others' identities and consult a common session key in the public channel.

A.Shamir [5] introduced an identity-based public key cryptosystem in 1984, in which a user's public key can be calculated from his identity and defined hash function, while the user's private key can be calculated by a trusted party called Private Key Generator (PKG). The identity-based public key cryptosystem simplifies the program of key management and increases the efficiency. In 2001, Boneh and Franklin [6] found bilinear pairings positive applications in cryptography and proposed the first practical identity-based encryption protocol with bilinear pairings. Soon, the bilinear pairings become important tools in constructing identity-based protocols and a number of identity-based encryption or signature schemes [7], [8], [9], [10], [11] [12] and authenticated key agreement protocols [13], [14], [15], [16] [17] were proposed.

In 2008, Gang Yao, Hongji Wang, and Qingshan Jiang [18] proposed an authenticated 3-round identity-based group key agreement protocol. The first round is for identity authentication, the second round is for key agreement, and the third round is for key confirmation. Their protocol is based on the protocol of Burmester and Desmedt [19] which was proposed at Eurocrypt 94. They declared the proposed protocol provably-secure in the random oracle model.

doi:10.4304/jnw.6.11.1578-1585

In this paper, we show that an authenticated 3-round identity-based group key agreement protocol proposed by Gang Yao et al. is vulnerable: an internal attacker can forge her neighbors' keying material. Then we propose an improved provably-secure protocol based on Burmester and Desmedt's as well. At last, we summarize several security attributes of our improved authenticated identity-based group key agreement protocol.

II. PRELIMINARIES

A. Security attributes

To get a rational key agreement protocol, Marko Hölbl, Tatjana Welzer and Boštjan Brumen defined some security attributes which have to be fulfilled by their secure authenticated key agreement protocol. Assume A, B and C are three honest entities. It is desired for authenticated key agreement protocol to possess the following security attributes [15]:

1. Known-Key Security. A unique secret session key should be generated in each round of a key agreement protocol. Each session key generated in one protocol round is independent and should not be exposed if other secret session keys are compromised, i.e. the compromise of one session key should not compromise other session keys.

2. Forward Secrecy. If long-term private keys of one or more of the entities are compromised, the secrecy of previously established session keys should not be affected. We say that a protocol has forward secrecy if some but not all of the entities' long-term keys can be corrupted without compromising previously established session keys, and we say that a protocol has perfect forward secrecy if the long-term keys of all the participating entities may be corrupted without compromising any previously established session key.

3. Key-Compromise Impersonation Resilience. Suppose that the long-term secret key of one participating entity is disclosing (e.g. A). Obviously, an adversary who knows this secret key can impersonate this entity to other participating entities (e.g. A to B and C). However, it is desired that this disclosure does not allow the adversary to impersonate other entities (e.g. B and C) to the entity whose long-term secret key was disclosed (e.g. A).

4. Unknown Key-Share Resilience. After the protocol ran, one entity (e.g. A) believes she shares a key with the other participating entities (e.g. B and C), while those entities (e.g. B and C) mistakenly believe that the key is instead shared with an adversary. Therefore, a rational authenticated key agreement protocol should prevent the unknown key-share situation.

5. Key Control. The key should be determined jointly by all participating entities (e.g. A, B and C). None of the participating entities can control the key alone.

The inclusion of identities of the participating entities and their roles in the key derivation function provide the resilience against unknown key share attacks and reflection attacks. The inclusion of transcripts in the key derivation function provides freshness and data origin authentication.

B. Bilinear pairing

Let P denote a generator of G_1, where G_1 is an additive group of large order q and let G_2 be a multiplicative group with $|G_1| = |G_2|$. A bilinear pairing is a map $e: G_1 \times G_1 \to G_2$ which has the following properties:

1.Bilinearity:

Given $Q,W,Z \in G_1$, $e(Q,W+Z) = e(Q,W) \cdot e(Q,Z)$ and $e(Q+W,Z) = e(Q,Z) \cdot e(W,Z)$. There for any $q,b \in Z_q$: $e(aQ,bW) = e(Q,W)^{ab} = e(abQ,W) = e(Q,abW) = e(bQ,W)^a$.

2. Non-degenerative:

$e(P,P) \neq 1$, where 1 is the identity element of G_2.

3. Computable:

If $Q,W \in G_1$, one can compute $e(Q,W) \in G_2$ in polynomial time efficiently.

C. Computational problems

Let G_1 and G_2 be two groups of prime order q, let $e: G_1 \times G_1 \to G_2$ be a bilinear pairing and let P be a generator of G_1.

- Discrete Logarithm Problem (DLP)
  Given $P,Q \in G_1$, find $n \in Z_q$ such that $P = nQ$ whenever such n exists.
- Computational Diffie-Hellman Problem (CDHP)
  Given $(P, aP, bP) \in G_1$ for $a,b \in Z_q^*$, find the element abP.
- Bilinear Diffie-Hellman Problem (BDHP)
  Given $(P, xP, yP, zP) \in G_1$ for $x,y,z \in Z_q^*$, compute $e(P,P)^{xyz} \in G_2$

D. Introduction of BR security model

To describe the security model for entity authentication and key agreement aims, M. Bellare and P. Rogaway proposed the BR93 model [13] for two-party authenticated key agreement protocol in 1993 and the BR95 model [14] for three-party authenticated key agreement protocol in 1995. In BR model, the adversary can control the communication channel and interact with a set of $\Pi^i_{U_x,U_y}$ oracles, which specify the behavior between the honest players U_x and U_y in their i^{th} instantiation. The predefined oracle queries are described informally as follows:

• Send (U_x, U_y, i, m): The adversary sends message m to the oracle $\Pi^i_{U_x,U_y}$. The oracle $\Pi^i_{U_x,U_y}$ will return the session key if the conversation has been accepted by U_x and U_y or terminate and tell the adversary.

• Reveal (U_x, U_y, i): It allows the adversary to expose an old session key that has been previously accepted. After receiving this query, $\Pi^i_{U_x,U_y}$ will send this session key to the adversary, if it has accepted and holds some session key.

• Corrupt (U_x, K): The adversary corrupts U_x and learns all the internal state of U_x. The corrupt query also allows the adversary to overwrite the long-term key of corrupted principal with any other value K.

• Test (U_x, U_y, i): It is the only oracle query that does not correspond to any of the adversary's abilities. If $\Pi^i_{U_x,U_y}$ has accepted with some session key and is being asked a Test(U_x, U_y, i) query, then depending on a randomly chosen bit b, the adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Freshness. The notion is used to identify the session keys about which adversary should not know anything because she has not revealed any oracles that have accepted the key and has not corrupted any principals knowing the key. Oracle $\Pi^i_{A,B}$ is fresh at the end of execution, if, and only if, oracle $\Pi^i_{A,B}$ has accepted with or without a partner oracle $\Pi^i_{B,A}$, both oracle $\Pi^i_{A,B}$ and its partner oracle $\Pi^i_{B,A}$ have not been sent a Reveal query, and the principals A and B of oracles $\Pi^i_{A,B}$ and $\Pi^i_{B,A}$ (if such a partner exists) have not been sent a Corrupt query.

Security is defined using the game G, played between a malicious adversary and a collection lection of $\Pi^i_{U_x,U_y}$ oracles and instances. The adversary runs the game simulation G, whose setting is as follows.

Phase 1: Adversary is able to send any Send, Reveal, and Corrupt oracle queries at will in the game simulation G.

Phase 2: At some point during G, adversary will choose a fresh session on which to be tested and send a Test query to the fresh oracle associated with the test session. Note that the test session chosen must be fresh. Depending on a randomly chosen bit b, adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Phase 3: Adversary continues making any Send, Reveal, and Corrupt oracle queries of its choice.

Finally, adversary terminates the game simulation and outputs a bit b', which is its guess of the value of b. Success of adversary in G is measured in terms of adversary's advantage in distinguishing whether adversary receives the real key or a random value. A wins if, after asking a Test (U_x, U_y, i) query, where $\Pi^i_{U_x,U_y}$ is fresh and has accepted, adversary's guess bit b' equals the bit b selected during the Test (U_x, U_y, i) query.

A protocol is secure in the BR model if both the validity and indistinguishability requirements are satisfied:

• Validity. When the protocol is run between two oracles in the absence of a malicious adversary, the two oracles accept the same key.

• Indistinguishability. For all probabilistic, polynomial-time (PPT) adversaries A, AdvA (k) is negligible.

III. Review of Gang Yao et al.'s Protocol

Let $U_1,\ldots,U_n$ be n participants, and PKG be the private key generator. Let ID_i be the identity of U_i. Suppose that G_1 and G_2 are two cyclic groups of order q for some large prime q. G_1 is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1, and $e: G_1 \times G_1 \to G_2$ be a bilinear pairing.

In Gang Yao et al.'s protocol, the following two steps prepare the system parameters:

Setup:

The PKG chooses a random number $s \in Z_q^*$ and set $R = sP$. The PKG also chooses $H_0 : \{0,1\}^* \to G_1^*$ to be a Map-to-Point hash function, and H is a cryptographic hash function. Then the PKG publishes system parameters $\{q, G_1, G_2, e, P, R, H_0, H\}$, and keeps s as its master key.

Extract:

Given a public identity $ID \in \{0,1\}^*$, the PKG computes the public key $Q = H_0(ID) \in G_1$ and generates the associated private key $S = sQ$. The PKG passed S as the private key to the user via some secure channel.

Let n users $U_1,\ldots,U_n$ with respective public keys $Q_i = H_0(ID_i)(1 \le i \le n)$ decide to agree upon a common secret key. $S_i = sQ_i$ is the long term secret key of U_i sent by the PKG on submitting U_i 's public identity $(1 \le i \le n)$. Let U denote $U_1 \| \cdots \| U_n$.

We assume that U_1 is the protocol initiator. The protocol may be performed in three rounds as follows:

Round 1: Identity Authentication

- Every participant U_i generates a random number $r_i \in Z_q^*$, computes

$E_i = r_i P$, $F_i = H(U, e(E_i, R))S_i + r_i R$

And broadcasts E_i and F_i.

- After receiving every E_j and F_j $(1 \le j \le n, j \ne i)$, U_i verifies that none of them equals 1. If the check succeeds, U_i verifies whether

$$e(\sum_{j \ne i} F_j, P) = e(\sum_{j \ne i} H(U, e(E_j, R))Q_j + E_j, R)$$

holds or not. If the verification succeeds, U_i continues with the next round. Otherwise, the protocol execution is terminated and a notification of failure will be broadcasted.

Round 2: Key Agreement

- U_i computes $T = H_0(ID_1 \| E_1 \| \cdots \| ID_n \| E_n)$, then he computes Y_i, X_i as follows.

$Y_i = r_i T$, $X_i = r_i(E_{i+1} - E_{i-1} + T)$,

And broadcasts X_i and Y_i.

- After receiving every X_i and Y_i $(1 \le j \le n, j \ne i)$, U_i verifies whether

$$e(\sum_{j \ne i} Y_j, P) = e(\sum_{j \ne i} E_j, T)$$

holds or not. IF the verification succeeds, U_i continues with the next round. Otherwise, the protocol execution is terminated and a notification of failure will be broadcasted.

Round 3: Key Confirmation

- U_i computes the keying material Z_i as

$$Z_i = e(nr_i E_{i-1} + \sum_{j=1}^{n-1}(n-1-j)(X_{i+j} - Y_{i+j}), R),$$

then he computes

$$C_i = H(i \| U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i)$$

and broadcasts C_i.

- After receiving every C_j $(1 \le j \le n, j \ne i)$, U_i computes the session key as

$K_i = H(U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i \| C_1 \| \cdots \| C_n)$.

Otherwise, U_i terminates the protocol execution and a notification of failure will be broadcasted.

IV. CRYPTANALYSIS OF GANG YAO ET AL.'S PROTOCOL

In our view, two important principles should be attention: Before we want to protect a message, we should know whether it really needs to be protected. After finishing our protocol, we should ensure all the valuable messages have been protected. In Gang Yao et al.'s protocol, to derive the session key, $K_i = H(U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_i \| C_1 \| \cdots \| C_n)$, all the parameters except Z_i can be gained from the broadcast messages. It is important to ensure the transmitted messages are not modified, forged, or deleted by attackers. In the round 1, E_i $(1 \le i \le n)$ should be protected and in the round 2, X_i and Y_i should be protected. However, only E_i and Y_i here are protected in Gang Yao et al.'s

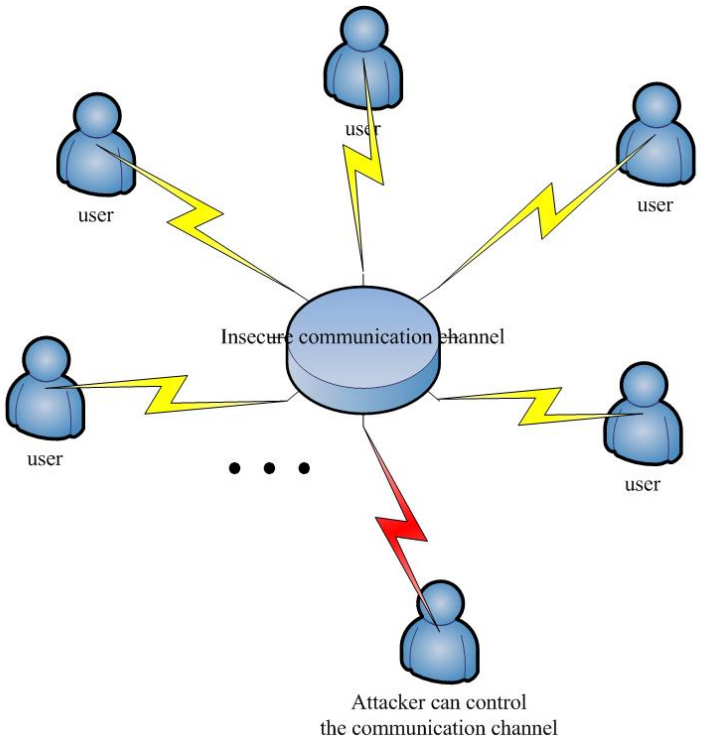


Figure 1. Attacking model.

protocol. Because the equation $e(\sum_{j\neq i} Y_j, P) = e(\sum_{j\neq i} E_j, T)$, where $T = H_0(ID_1 \| E_1 \| \cdots \| ID_n \| E_n)$, user U_i can ensure that Y_j is not modified or forged but X_i is transmitted without any verification. Actually, both X_{i+j} and Y_{i+j} are needed in the equation $Z_i = e(nr_i E_{i-1} + \sum_{j=1}^{n-1}(n-1-j)(X_{i+j} - Y_{i+j}), R)$ in the round 3. Thus, X_i can be replaced by $X_i = Y_i$. Then Z_i can be expressed as $Z_i = e(nr_i E_{i-1}, R)$. Due to the characteristic of bilinear pairing: $Z_i = e(nr_i E_{i-1}, R) = e(nr_i r_{i-1} P, sP)$ $= e(r_{i-1} sP, r_i P)^n = e(r_{i-1} R, E_i)^n$ That is, with the random number r_i, any user U_i can generate U_{i+1}'s keying material Z_{i+1}.The attacking model is described as the figure 1.

As a result, an attacker who can control the communication channel has the ability to intercept and forge all the X_i s. If a malicious user U_e wants to forge U_{e+1}'s keying material C_{e+1}, she can compute $Z_{e+1} = e(r_e R, E_{e+1})^n$ and $C'_{e+1} = H(e+1 \| U \| E_1 \| \cdots \| E_n \| X_1 \| \cdots \| X_n \| Y_1 \| \cdots \| Y_n \| Z_{e+1})$. Finally, she can broadcast C'_{e+1} to replace C_{e+1}.

V. Improvement of Gang Yao et al.'s protocol

In this section, we first review Bermester and Desmedt's group key exchange protocol. Then we propose a non-authentication protocol based on their protocol with bilinear pairing. Finally, we improve the non-authentication group key agreement protocol to an authentication group key agreement protocol

A. Bermester and Desmedt's group key exchange protocol

Let n be the size of the group, the Bermester and Desmedt's group key exchange protocol works as follows:

- Each participant U_i chooses a random number x_i and broadcasts $z_i = g^{x_i}$;
- Each participant computes $Z_i = z_{i-1}^{x_i}$ and $Z_{i+1} = z_i^{x_{i+1}} = z_{i+1}^{x_i}$, and broadcasts $X_i = Z_{i+1} / Z_i$;
- Each participant computes his session key as $K_i = Z_i^n X_i^{n-1} X_{i+1}^{n-2} \cdots X_{i+n-2}$.

It is easy to see that each U_i can compute the same session key $K_i = \sum_{j=1}^{n} Z_j = g^{x_1x_2 + x_2x_3 + \cdots + x_nx_1}$

B. Non-authentication protocol transformed from Bermester and Desmedt's protocol

G_1 and G_2 are two cyclic groups of order q for some large prime q. G_1 is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1, $e: G_1 \times G_1 \to G_2$ be a bilinear pairing and n be the size of the group, the non-authentication protocol works as follows:

- Each user U_i chooses a random number $r_i \in Z_q^*$ and broadcasts $z_i = r_i P$
- Each user U_i computes $Z_i = r_i z_{i-1}$, $Z_{i+1} = r_i z_{i+1}$, and broadcasts $X_i = Z_{i+1} - Z_i$
- Each player U_i can computes his session key as: $K_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$

It is easy to see that for each U_i, $K_i = \sum_{j=1}^{n} Z_j = (r_1r_2 + r_2r_3 + \cdots + r_nr_1)P$

C. Our authenticated identity-based group key agreement protocol

Let $U_1, \ldots, U_n$ be n participants, and PKG be the private key generator. Let ID_i be the identity of U_i. Suppose that G_1 and G_2 are two cyclic groups of order q for some large prime q. G_1 is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1, and $e: G_1 \times G_1 \to G_2$ be a bilinear pairing.

Our protocol is described as follows:

Setup:

The PKG chooses a random number $s \in Z_q^*$, sets $R = sP$, chooses two hash functions, H_0 and H, where $H_0: \{0,1\}^* \to G_1^*$. Then the PKG publishes system parameters $\{q, G_1, G_2, e, P, R, H_0, H\}$, and keeps the master key s as a secret.

Extract:

Given a public identity $ID \in \{0,1\}^*$, the PKG computes the public key $Q = H_0(ID) \in G_1$ and generates the associated private key $S = sQ$. The PKG outputs S as the private key to the user via some secure channel.

Let n users $U_1, \ldots, U_n$ with respective public key $Q_i = H_0(ID_i)(1 \le i \le n)$ decide to agree upon a common secret key. $S_i = sQ_i$ is the long term secret key of U_i sent by the PKG on submitting U_i's public identity $(1 \le i \le n)$. Let U denote $U_1 \| \cdots \| U_n$.

We assume that U_1 is the protocol initiator. The protocol may be performed in three rounds as follows:

- Round 1:

Each participant U_i chooses a random number $r_i \in Z_q^*$, computes $z_i = r_i P$, $B_i = H(ID_i, z_i)$, $v_i = B_i S_i$ and broadcasts (ID_i, z_i, v_i).

- Round 2:

After receiving each $(ID_i, z_i, v_i)\ (1 \le i \le n)$, each user can compute

$B_i = H(ID_i, z_i)$, $Q_i = H_0(ID_i)(1 \le i \le n)$

and verify whether the equation

$e(v_i, P) \stackrel{?}{=} e(B_i Q_i, R)$

sets or not. If the equation sets, U_i can ensure that (ID_i, z_i, v_i) is not modified or forged by attackers. Then he computes $Z_i = r_i z_{i-1}$, $Z_{i+1} = r_i z_{i+1}$, $X_i = Z_{i+1} - Z_i$, $C_i = H(ID_i, X_i)$, $w_i = C_i S_i$ and broadcasts (ID_i, X_i, w_i).

- Round 3:

After receiving each $(ID_i, X_i, w_i)\ (1 \le i \le n)$, each user can compute

$C_i = H(ID_i, X_i)$, $Q_i = H_0(ID_i)(1 \le i \le n)$

and verify whether the equation

$e(w_i, P) \stackrel{?}{=} e(C_i Q_i, R)$

sets or not. If the equation sets, U_i can ensure that (ID_i, X_i, w_i) is not modified or forged by attackers. Then he computes the keying material

$D_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$

Actually,

$$D_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$$
$$= \sum_{j=1}^{n} Z_j = (r_1 r_2 + r_2 r_3 + \cdots + r_n r_1) P$$

Then each user computes the session key as

$K_i = H(U \| z_1 \| \ldots \| z_n \| X_1 \| \ldots \| X_n \| D_i)$

VI. Security analysis and security attributes

Theorem 6.1 Any modification can be found by the short signature if the hash function H is collision resistance.

Proof: In the function of $e(v_i, P) \stackrel{?}{=} e(B_i Q_i, R)$, the parameters P and R are public, which can not be forged or modified, and $B_i = H(ID_i, z_i)$, $Q_i = H_0(ID_i)$ are computed by the receiver. Though v_i and z_i may be modified by the attacker, the collision resistance hash function H will make it impossible to gain suitable pairs of v_i and z_i to pass the verification function. So if attacker modifies any elements of (ID_i, z_i, v_i), other users can find it. The function $e(w_i, P) \stackrel{?}{=} e(C_i Q_i, R)$ has a similar situation with $e(v_i, P) \stackrel{?}{=} e(B_i Q_i, R)$. That is why any modification can be found by the short signature. □

Theorem 6.2 The attacker can't obtain the session key from the intermediate messages if CDH problem is hard.

Proof: Suppose the challenger C wants to solve the CDH problem. That is, given (aP, bP) , C should compute abP . In our protocol, the intermediate messages transmitted in the public channel are (ID_i, z_i, v_i) in the first round and (ID_i, X_i, w_i) in the second round. The efficient elements are z_i and X_i , and other elements are used to protect them. Supposed that attacker can obtain the session key $K_i = H(U \| z_1 \| \ldots \| z_n \| X_1 \| \ldots \| X_n \| D_i)$. That is, she can obtain the keying material D_i .For $D_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$, she can obtain Z_i according to the equation $Z_i = \left. D_i - \left[(n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}\right] \right/ n$, where X_i and n had been obtained by the attacker. As it is known to us, $z_i = r_i P$ and $Z_i = r_i z_{i-1} = r_i r_{i-1} P$. Define $z_i = aP$ and $z_{i-1} = bP$, which is given to the attacker. If she can obtain the session key, she can compute $Z_i = abP$ and she solves the CDH problem. That is to say, if CDH problem is hard, the attacker can't obtain the session key. □

VII. Conclusions

In this paper, we show that an authenticated 3-round identity-based group key agreement protocol proposed by Gang Yao et al. is vulnerable: an internal attacker can forge her neighbors' keying material. Then we propose an improved provably-secure protocol based on Burmester and Desmedt's protocol as well. At last, we summarize some security attributes of our improved authenticated identity-based group key agreement protocol.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

REFERENCES

[1] Sandeep S. Kulkarni, Bezawada Bruhadeshwar, Key-update distribution in secure group communication, Computer Communications 33 (2010) 689–705.

[2] New Directions in Cryptography W. Diffie and M. E. Hellman, IEEE Transactions on Information Theory, vol. IT-22, (1976) 644–654.

[3] The First Ten Years of Public-Key Cryptography Whitfield Diffie, Proceedings of the IEEE, 76 (5) (1988) 560–577.

[4] Jianjie Zhao, Dawu Gu, Yali Li, An efficient fault-tolerant group key agreement protocol, Computer Communications 33 (2010) 890–895.

[5] A.Shamir, Identity-based Cryptosystems and Signature Schemes. Advances in Cryptology, CRYPTO'84, LNCS 196, Springer-Verlag, Berlin, 1984, pp.47-53

[6] D.Boneh, M.Franklin, Identity-based Encryption from the Weil pairing, Advances in Cryptology, CRYPTO'2001, LNCS 2139, Springer-Verlag, Berlin, , 2001, pp.213-229.

[7] Ji-Jian Chin, Swee-Huay Heng, and Bok-Min Goi,An efficient and provable secure identity-based identification scheme in the standard model, LNCS 5057, Springer-Verlag, Berlin, 2008, pp.60-73.

[8] Zhenhua Liu, Yupu Hu, Xiangsong Zhang, Hua Ma, Certificateless signcryption scheme in the standard model, Information Sciences 180 (2010) 452-464.

[9] Jianhong Zhang, Yixian Yang, Xinxin Niu, Shengnan Gao, Hua Chen, Qin Geng, An Improved Secure Identity-Based On-Line/Off-Line Signature Scheme, ISA 2009, LNCS 5576, Springer-Verlag, Berlin,2009, pp. 588–597 .

[10] Ting-Yi Chang, An ID-based group-oriented decryption scheme secure against adaptive chosen-ciphertext attacks, Computer Communications 32 (2009) 1829-1836.

[11] Aggelos Kiayias, Hong-Sheng Zhou, Hidden Identity-Based Signatures, LNCS 4886, Springer-Verlag, Berlin, 2007, pp.134–147.

[12] Chun-Ta Li, On the Security Enhancement of an Efficient and Secure Event Signature Protocol for P2P MMOGs, ICCSA 2010, LNCS 6016, pp.599–609, 2010.

[13] R. Lu, Z. Cao, A new deniable authentication protocol from bilinear pairings, Applied Mathematics and Computation 168 (2005) 954-961.

[14] R. Lu, Z. Cao, S. Wang and H. Bao, A New ID-based Deniable Authentication Protocol,. Informatics 18 (1) (2007) 67-78.

[15] T. Cao, D. Lin and R. Xue, An Efficient ID-based Deniable Authentication Protocol from pairings, AINA'05, 2005, 388-391.

[16] J.S. Chou, Y.L. Chen, J.C. Huang, An ID-Based Deniable Authentication Protocol on Pairings, Cryptology ePrint Archive: Report (335), 2006.

[17] Jung Yeon Hwang, Kyu Young Choi, Dong Hoon Lee, Security weakness in an authenticated group key agreement protocol in two rounds, Computer Communications 31 (2008) 3719–3724.

[18] G. Yao, H. Wang, and Q. Jiang, An Authenticated 3-Round Identity-Based Group Key Agreement Protocol, the third International Conference on Availability, Reliability, and Security, pp.538-543, ACM, 2008.

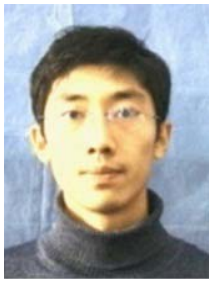
[19] M. Burmester and Y. Desmedt, A Secure and Efficient Conference Key Distribution System, EUROCRYPT'94, LNCS 950, Springer-Verlag, Berlin, 1994, pp.275-286.

Wei Yuan was born in Chengde of Hebei province of China in 1984. He began the study of computer science at Jilin University in 2003 and got his bachelor degree in 2007. Then he continued his research on information security and received his master degree in 2010. Now he is a PhD candidate of the college of computer science and technology of Jilin University.

His main research interests include cryptography and information security. he have participated in several projects include two National Natural Science Foundations of China and one National Grand Fundamental Research 973 Program of China and published more than 10 research papers from 2007.

Liang Hu was born in 1968. He has his BS degree on Computer Systems Harbin Institute of Technology in 1993 and his PhD on Computer Software and Theory in 1999. Currently, he is the professor and PhD supervisor of College of Computer Science and Technology, Jilin University, China.

His main research interests include distributed systems, computer networks, communications technology and information security system, etc. As a person in charge or a principal participant, Dr Liang Hu has finished more than 20 national, provincial and ministerial level research projects of China.



Hongtu Li was born in Siping of Jilin, China on Mar. 17 1984. In 2002, Li Hongtu began the study of computer science at Jilin University in Jilin, Changchun, China. And in 2006, Li Hongtu got bachelor's degree of computer science. In the same year, Li Hongtu began the master's degree study in network security at Jilin University. After 3 years study, Li Hongtu got his master's degree in 2009. From then on, Li Hongtu began the doctor's degree in the same field of study at the same University.

From 2009, he has got a fellowship job. He worked in grid and network security laboratory as an ASSISTANT RESEACHER at Jilin University. From 2006 to now, he has published several papers. The list of published articles or books is as follows:

"Identity -Based Short Signature Without Random Oracles Model", International Conference of ICT Innovation and Application-ICIIA2008, Guangzhou, China, 2008.

"Registration and private key distribution protocol based on IBE", the 5th International Conference on Frontier of Computer Science and Technology-FCST2010, Changchun, China, 2010.

"Certificateless authenticated key agreement protocol against KCI and KRA", The 2011 International Conference on Network Computing and Information Security-NCIS'11 and the 2011 International Conference on Multimedia and Signal Processing-CMSP'11, Guilin, China, 2011.

Expect network security, he also interested in grid computing, wireless networks, intrusion detection and so on. From 2006 to now, he have participated in or led several projects include two National Natural Science Foundations of

China and one National Grand Fundamental Research 973 Program of China.

Jianfeng Chu, corresponding author, was born in 1978, Ph.D. , Now he is the teacher of the College of Computer Science and Technology, Jilin University, Changchun, China. He received the Ph.D. degree in computer structure from Jilin University in 2009. His current research interests focus on information security and cryptology.

An important objective of the projects is to probe the trend of network security, which can satisfy the need of constructing high-speed, large-scale and multi-services networks. Various complex attacks can not be dealt with by simple defense. And to add mechanisms to network architecture results in decreasing performance. In a word, fundamental re-examination of how to build trustworthy distributed network should be made.

Yuyu Sun, female, born in 1977,Lecturer, Ph.D. of Jilin University. She graduated from the Department of Computer Science and Technology of Jilin University in 2005,and obtained an MA degree. From 2008,she began to start her doctorate in computer in Jilin University, now she is working in Changchun University. Her current research interests include network and information security. She mainly engaged in Teaching and research on information security and Application software development. She has participated in one National Natural Science Foundation of China, one Major Project of Chinese National Programs for Fundamental Research and Development (973 Program), five Science and technology support key project plan of Jilin Provincial Science and technology Department , three S&T plan projects of Jilin Provincial Education Department .She has Wrote 4 textbooks as yet. She has published 14 academic articles in English and Chinese, four of that has been retrieved by EI.

A Task Scheduling Strategy in Heterogeneous Multi-sinks Wireless Sensor Networks

Liang Dai
School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China
Email: ldai1981@gmail.com

Hongke Xu
School of Electronic and Control Engineering, Chang'an University, Xi'an 710064, China
Email: xuhongke@chd.edu.cn

Ting Chen
School of Information Engineering, Chang'an University, Xi'an 710064, China
Email: tchenpcn@126.com

***Abstract*—Using multiple sinks in a wireless sensor network can significantly decrease the amount of energy spent on communication, so it has been paid much attention in recent years. In this paper, we introduce a new divisible load scheduling strategy to solve the problem how to complete the tasks within the possibly shortest time in multi-sinks wireless sensor network. In this strategy, the tasks are distributed to wireless sensor network based on the processing and communication capacity of each sensors by multiple sinks. After received the sub-tasks, the intra-cluster sensors perform its tasks simultaneously, and send its results to cluster head sequentially. By removing communications interference between each sensor, reduced makespan and improved network resource utilization achieved. Cluster heads send fused data to sinks sequentially after fused the data got from intra-cluster sensors, which could overlap the task-performing and communication phase much better. A unique scheduling strategy that allows one to obtain closed form solutions for the optimal finish time and load allocation for each node in heterogeneous clustered networks is presented. And solutions for an optimal allocation of fractions of task to sensors in the network are also obtained via bi-level programming. Finally, simulation results indicate this strategy reasonably distributes tasks to each node in multi-sinks wireless sensor networks, and effectively reduces the time-consuming of task completion. Compared to the traditional single-sink structure, makespan is reduced by 20%, and the energy-consuming of sensors is more balanced.**

***Index Terms*—wireless sensor networks, heterogeneous, divisible load theory, task scheduling, multiple sinks**

I. Introduction

In recent years, it's discovered that the stability and effectiveness of wireless sensor networks faced a huge threat if there is only one sink as data management center. In view of this situation, multi-sinks wireless sensor networks become a new hotspot [1-2]. As shown in Figure 1, Data acquisition in single-sink sensor networks might have issues in scalability. As the size of sensor networks grows, the distances between the sink and the responding sensors become larger. This leads to a greater energy consumption for query-flooding and data-collection between sensors and the sink, leading to a possible reduction in the lifetime of the sensors. Hence, we need to design energy-efficient data acquisition mechanisms that scale with the size of the network. One solution is to simultaneously deploy multiple sinks in the sensor network.

Owing to the wireless sensor network node with limited energy, the task should be completed within the shortest possible amount of time. Divisible load theory [3] provides an effective solution to wireless sensor networks for task scheduling [4-7]. Different from other heuristic solutions of task scheduling problem in wireless sensor networks [8-9], this scheme can get not only the optimal solution, but also the analytic solution, thus ensuring the consistency of the results of scheduling.

Divisible load scheduling algorithms were applied to wireless sensor networks in [4-7]. Although the authors derived closed-form solutions to obtain the optimal finish time, the network topology discussed in those papers is single-level tree structure. While in wireless sensor networks, as compared with the single-level tree structure, clustered structure (multi-level tree structure) has a great of advantages [10].

Therefore, we present a task scheduling algorithm(DMTA) based on divisible load theory in multi-sinks wireless sensor networks The goal of this algorithm is to minimize the overall execution time (hereafter called makespan) and fully utilize network resources, by finding an optimal strategy of splitting the original tasks received by sink into a number of sub-tasks as well as distributing these sub-tasks to the clusters in the right order, and through the proposed mechanism to encourage collaboration.

doi:10.4304/jnw.6.11.1586-1593

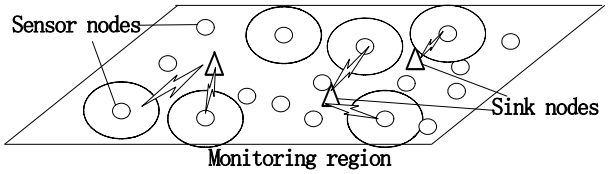


Figure1. Multi-sinks wireless sensor networks

II. RELATED WORKS AND OUR CONTRIBUTIONS

Due to the nature of distributed sensing in wireless sensor networks, it is expected that divisible load theory will play an important role in providing an optimal solution for load distribution under WSN environments, especially those having various energy and computational constraints. Divisible load theory was firstly applied to wireless sensor networks for the task scheduling analysis in [4]. The type of application this work involves measurements from a sample space where each sensor is assigned a certain nonoverlapping part of the sample space to measure. For instance, the sample space may consist of a very large frequency range and performing measurements by a single sensor may be both time and energy consuming. Although the authors derived closed-form solutions to obtain the optimal finish time for three particular sensor networks, their single level (single-hop) model is not scalable to a large sensor network. While in wireless sensor networks, as compared with the single-level tree structure, clustered structure (multi-level tree structure) has a great of advantages [10]. In [5], based on [4], Liu analyses the delay of measurement and communication in single-level wireless sensor networks. LIU applied divisible load theory to wireless sensor mesh networks to analysis task scheduling, the task issuing and the results reporting are also based on single-level tree [6]. Kijeung considers a single channel cluster composed of single cluster head and n sensors in the wireless network (star topology). This is a simple network scenario where the intra-cluster sensors can report their own sensory data directly to the clusterhead via a single channel. But in a hierarchical wireless sensor network, the network sensor nodes are usually partitioned into multiple clusters. Single cluster means very little in wireless sensor networks [7]. Existing research efforts on multi-cluster three tier hierarchical wireless sensor networks model based on four scenarios (single/multi channel, with/without front-end) [11-13]. However, the network models in those papers are assumed to be homogeneous.

To the best of our knowledge, no paper has given a satisfactory solution to the case where both the sensor network is heterogeneous and with multiple sinks, and measurement results transfer to the source is explicitly considered.

In this paper, the task scheduling problem of heterogeneous clustered wireless sensor networks with multiple sinks is formulated and analyzed in detail. The major contributions of this paper are: we present a task scheduling strategy in heterogeneous clustered wireless sensor networks with multiple sinks. The goal of this strategy is to minimize the overall execution time (hereafter called makespan) and fully utilize network resources, by finding an optimal strategy of splitting the original tasks received by sinks into a number of sub-tasks as well as distributing these sub-tasks to the clusters in the right order. The strategy consists of two phases: intra-cluster task scheduling and inter-cluster task scheduling. Intra-cluster task scheduling deals with allocating different fractions of sensing tasks among sensor nodes in each cluster; inter-cluster task scheduling involves the assignment of sensing tasks among all clusters. This strategy builds from eliminating transmission collisions and idle gaps between two successive data transmissions. By removing performance degradation caused by communication interference and idles, the reduced finish time and improved network resource utilization can be achieved. With the proposed strategy, the optimal finish time and the most reasonable load allocation ratio on each node could be derived.

In wireless sensor networks, cluster head is responsible for data exchange for SINK and in-cluster nodes. In order to reduce energy consumption caused by transmitting redundant data, lower latency and prolong the survival period, cluster head needs fuse the data [14]. A new estimation method for data fusion — information utilization constant is introduced[7] in this paper. Information utilization constant is based on a technique of information accuracy estimation. Through estimating accuracy of information, cluster head can know the approximate percentage of data fusion.

III. PROBLEM DESCRIPTION

Wireless sensor networks construct clusters several times in its life cycle. Each cluster will have a set-up phase and a steady-state phase. We discuss our task scheduling strategy in a steady-phase phase.

The original tasks received by sinks are divided into two stages: inter-cluster task scheduling and intra-cluster task scheduling. First, inter-cluster task scheduling partitions the entire tasks into each cluster, and then the sub-tasks in a cluster is assigned to each intra-cluster sensor nodes by intra-cluster task scheduling.

According to divisible load theory, to remove performance degradation caused by communications interference, sinks sends each round's tasks to cluster head sequentially. After each cluster finishing its tasks and fusing the data, the cluster heads also send this round's results to sinks sequentially. That in every moment only allows sinks node sends sub-tasks to a cluster head, or a cluster head return fusion data to the sinks.

Divisible load theory is characterized by the fine granularity of loads. There is also no precedence relation among the data elements. Such a load may be arbitrarily partitioned and distributed among sensors and links in a system. So without loss of generality, we use double–sinks wireless sensor networks as model to analyze the task scheduling problem in multi-sinks wireless sensor networks.

The network topology discussed in this paper is shown in Fig. 2.

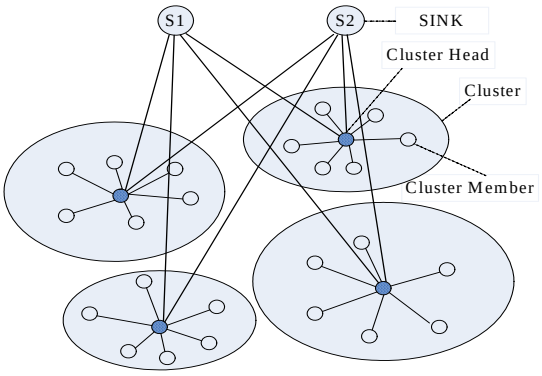


Figure2. Network topology

There are two sinks (S_1 and S_2) and k clusters ($Cluster_i$, $i=1,....,k$) in the network. Each cluster head were expressed as Ch_i , $i=1,....,k$. Within $Cluster_i$, there are n_i , $i=1,....,k$ nodes expressed as n_{ij} $(i=1,....k; j=1,...n_i)$ respectively. Communication links between Cluster head and the two sinks are expressed as l_{1i} and l_{2i} , $i=1,....,k$ respectively. Communication links between intra-cluster nodes and cluster heads are expressed as l_{ij} $(i=1,....k; j=1,...n_i)$ respectively.

The following notations will be used throughout this paper:

L_s : Total load originated from sink s , $(s=1,2)$

α_i : The total fraction of load that is assigned by the sinks to cluster head i , $(i=1,....,k)$

α_{1i} : The fraction of load that is assigned to cluster head i by the first sink;

α_{2i} : The fraction of load that is assigned to cluster head i by the second sink;

$$\alpha_i = \alpha_{1i} + \alpha_{2i},\ (i=1,....,k) \tag{1}$$

$\alpha_{1i,j}$: The fraction of load that is assigned to intra-cluster node n_{ij} in cluster i by the first sink;

$\alpha_{2i,j}$: The fraction of load that is assigned to intra-cluster node n_{ij} in cluster i by the second sink;

By definition we can see:

$$\sum_{i=1}^{k} \alpha_i = 1 \tag{2}$$

$$\sum_{j=1}^{n_i} \alpha_{1i,j} = \alpha_{1i} \tag{3}$$

$$\sum_{j=1}^{n_i} \alpha_{2i,j} = \alpha_{2i} \tag{4}$$

ω_i : A constant that is inversely proportional to the processing (data fusion) speed of cluster head Ch_i .

$y_{i,j}$: A constant that is inversely proportional to the measuring speed of intra-cluster node n_{ij} in the network.

z_{1i} : A constant that is inversely proportional to the speed of link between the first sink and the i th cluster head in the network

z_{2i} : A constant that is inversely proportional to the speed of link between the second sink and the i th cluster head in the network

$z_{i,j}$: A constant that is inversely proportional to the speed of link between the cluster head Ch_i in the network.

T_{ms} : Measurement intensity constant. This is the time it takes the intra-cluster node n_{ij} to measure the entire load when $y_{i,j}=1$. The entire assigned measurement load can be measured on the intra-cluster node n_{ij} in time $y_{i,j}T_{ms}$

T_{cm} : Communication intensity constant. This is the time it takes to transmit the entire processing load over a link when $z_i=1$. The entire load can be transmitted over the i th link in time z_iT_{cm}

T_{cp} : Data fusion intensity constant. This is the time it takes to fuse the entire load on a cluster head when $\omega_i=1$. The entire load can be fused on cluster head Ch_i in time ω_iT_{cp} .

φ_i : The information utility constant of cluster head Ch_i .

The operation process of the entire application is as follows:

1. Sink firstly divided the general task and assigned the sub-tasks to each cluster head.
2. Each cluster head partitioned the tasks it received then distributed to the nodes within its cluster.
3. Intra-cluster nodes performed measurement while reported the results to the cluster head.
4. Cluster head fused the data it received from intra-cluster nodes while sent the fused data to the sink node.

IV. Optimal Scheduling Algorithm

Wireless sensor networks construct clusters several times in its life cycle. Each cluster will have a set-up phase and a steady-state phase. We discuss our multi-rounds task scheduling algorithm in a steady-phase phase.

The original tasks received by sink are divided into two stages: inter-cluster task scheduling and intra-cluster task scheduling. First, inter-cluster task scheduling partitions the entire tasks into each cluster, and then the sub-tasks in a cluster is assigned to each intra-cluster sensor node by intra-cluster task scheduling. To improve

overlap of communication with computation, inter-cluster task scheduling assigned sensing tasks among all clusters in multiple rounds.

According to divisible load theory, to remove performance degradation caused by communications interference, sinks sends tasks to cluster head sequentially. After each cluster finishing its tasks and fusing the data, the cluster heads also send this round's results to SINK sequentially. That in every moment only allows SINK node sends sub-tasks to a cluster head, or a cluster head return fusion data to the sinks.

Two generic techniques for solving linear divisible load schedule problems are linear equation solution and linear programming. Analytical closed form solutions have the advantage of giving insight into system dependencies and tradeoffs. Furthermore, analytical solutions, when they can be realized, usually require only a trivial amount of calculation. Linear programming has the advantage of being able to handle a wide variety of constraints and producing numerical solutions for all types of linear models. Alternately one can often, though not always, set up a set of linear equations that can be solved either numerically or, in special cases, analytically.

In this subsection A, a typical closed form solution for task scheduling of heterogeneous wireless sensor networks is achieved. In subsection B, a representative task scheduling problem with bi-level programming solution is discussed.

A. A closed form solution

A.1 **Intra-cluster task scheduling**

Fig.3 illustrates the timing diagram for a set of sensor nodes, indexed from n_1 to n_k , in one cluster. From Fig.3, it can be observed that there is no time gap between every two successive nodes because the divisible workload can be transferred in the cluster. All sensor nodes start to measure data at the same time. Once the previous node finishes transmitting data, the other one completes its measuring task and starts to report its data. As a result, the proposed timing diagram minimizes the finish time by scheduling the measuring time and reporting time of each senor node. Moreover, since the intra-cluster scheduling tries to avoid the transmission conflicts at the cluster head, energy spent on retransmission are conserved.

The working time of a sensor node can be divided into two parts: measuring time and reporting time.

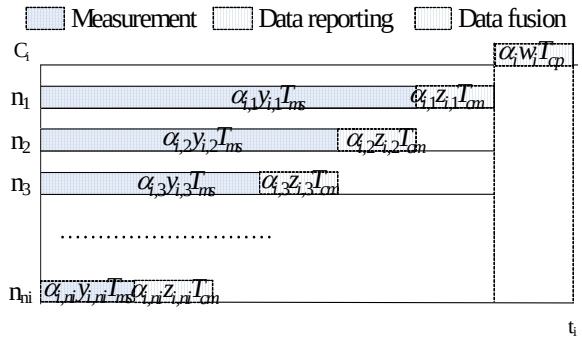


Figure 3. Timing diagram for intra-cluster task-processing

The task scheduling model considered in this paper is shown in Figure 3. The intra-cluster nodes began collecting data at the same time, and report the results collected to cluster head.

In order to fully utilize the link bandwidth, and avoid the waiting between different nodes, intra-cluster nodes completed reporting data collected to cluster head synchronously. Cluster head received the reported data from intra-cluster nodes, then fused those data, and sent the fused results to sink.

Similarly, in order to remove the performance degradation caused by idle, and to improve efficiency, cluster head completed reporting the fused data to the sink node.

For Cluster head C_i , based on the timing diagram shown in Fig. 3, one can write the following set of equations:

$$\alpha_{1i,j-1} y_{i,j-1} T_{ms} = \alpha_{1i,j} y_{i,j} T_{ms} + \alpha_{1i,j} z_{i,j} T_{cm}, \quad i = 2,3,...k \tag{5}$$

A general expression for the above set of recursive equations can be written as

$$\alpha_{1i,j} = s_{i,j} \alpha_{1i,j-1} \tag{6}$$

where $s_{i,j} = y_{i,j-1} T_{ms} / (y_{i,j} T_{ms} + z_{i,j} T_{cm})$ and $i = 2,3,...k$

The above recursive equation for $\alpha_{1i,1}$ can be rewritten in terms of α_{1i} only as

$$\alpha_{1i,1} = \alpha_{1i} / (1 + \sum_{j=2}^{n_i} \prod_{k=2}^{j} s_{i,k}) \tag{7}$$

The cluster head will use the above value of $\alpha_{1i,1}$ to obtain the amount of data that has to be measured by the rest of the $n_i - 1$ sensors by using

$$\alpha_{1i,j} = \alpha_{1i} \prod_{k=2}^{j} (s_{i,k}) / (1 + \sum_{j=2}^{n_i} \prod_{k=2}^{j} s_{i,k}) \tag{8}$$

The minimum measuring and reporting time of the first sink's sub-task α_{1i} will then be given as

$$t_{1i} = \alpha_{1i} (y_{i,1} T_{ms} + z_{i,1} T_{cm}) / (1 + \sum_{j=2}^{n_i} \prod_{k=2}^{j} s_{i,k}) + \alpha_{1i} w_i T_{cp} \tag{9}$$

Similarly we can get the minimum measuring and reporting time of the second sink's sub-task α_{2i} is :

$$t_{2i} = \alpha_{2i} (y_{i,1} T_{ms} + z_{i,1} T_{cm}) / (1 + \sum_{j=2}^{n_i} \prod_{k=2}^{j} s_{i,k}) + \alpha_{2i} w_i T_{cp} \tag{10}$$

A.2 **Inter-cluster task scheduling**

After cluster heads fused the cluster's measured data, cluster heads can sent the fused data to sinks concurrently because each cluster head has a separate channel to the sinks.

In order to remove the performance degradation caused by idle, and to improve efficiency, as shown in Fig. 4, we can get

$$\varphi_i\alpha_{2i}z_{2i}T_{cm} = t_{2i} + \varphi_i\alpha_{1i}z_{1i}T_{cm} \quad (11)$$

In eq. (11) and (12), we make

$$(y_{i,1}T_{ms} + z_{i,1}T_{cm}) / (1 + \sum_{j=2}^{n_i}\prod_{k=2}^{j} s_{i,k}) + w_iT_{cp} = s_i$$

, then take s_i to eq. (13), we can get

$$\alpha_{1i} = r_i\alpha_{2i} \quad (12)$$

where $r_i = \varphi_i z_{2i}T_{cm} / (s_i + \varphi_i z_{1i}T_{cm})$

The total tasks cluster head Ch_i get is :

$$\alpha_i = \alpha_{1i} + \alpha_{2i} \quad (13)$$

From Fig. 4 one can see that:

$$\alpha_i s_i + \varphi_i\alpha_{1i}z_{1i}T_{cm} = \alpha_{i+1}s_{i+1} + \varphi_{i+1}\alpha_{1,i+1}z_{1,i+1}{}^1T_{cm} \quad (14)$$

From eq. (14) to eq. (16), we can get

$$\alpha_i l_i = \alpha_{i+1}l_{i+1} \quad (15)$$

where $l_i = s_i + r_i\varphi_i z_i{}^1T_{cm} / (1 + r_i)$

Now using the eq. (1), one can solve for α_i as

$$\alpha_i = (1/l_i) / \sum_{i=1}^{k}(1/l_i) \quad (16)$$

Hereto, we can get that the tasks cluster head Ch_i and the intra-cluster nodes within it received from the first sink α_i^1 and $\alpha_{i,j}{}^1$. Similarly, the tasks from the second sink α_i^2 and $\alpha_{i,j}{}^2$. And the total task execution time

$$T_f = t_{2i} + t_{1i} + \varphi_i\alpha_{1i}z_{1i}T_{cm} \quad (17)$$

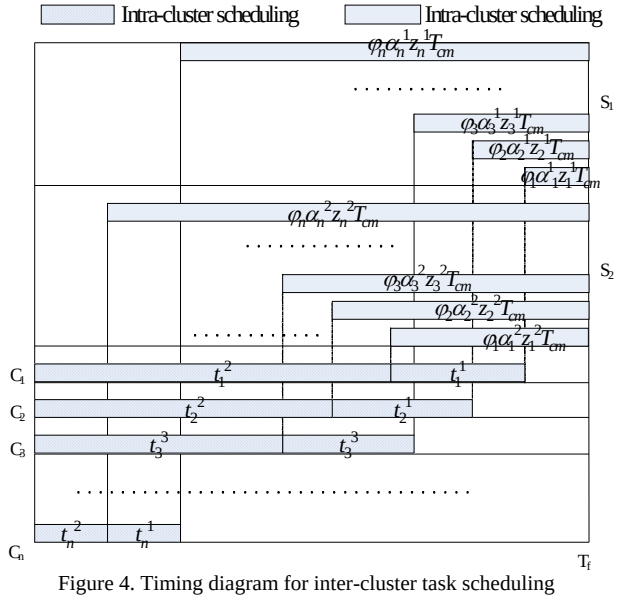


Figure 4. Timing diagram for inter-cluster task scheduling

B. A bi-level programming method

In this subsection, a bi-level programming model is constructed in the task scheduling problem of wireless sensor networks.

We can regard the task scheduling problem as a Leader-Follower problem.

The upper-level can be described as the load allocation ratio of sinks allocated to each cluster head satisfying the divisible load theory, which make the makespan minimum. The lower-level can be described as the load allocation ratio of cluster head allocated to each intra-cluster sensor divisible load theory, which make the intra-cluster task completion times minimum.

The problem of minimizing the total task finish time in scheduling algorithm is described below:

$$t_{2i} + t_{1i} + \varphi_i\alpha_{1i}z_{1i}T_{cm} \le T_f \quad (18)$$

So, for the upper-level programming, the mathematical model is as follow:

$$\begin{aligned} &Min\ T_f \\ &subject\ to\ (Cluster\ head\ C_i) \\ &t_{2i} + t_{1i} + \varphi_i\alpha_{1i}z_{1i}T_{cm} \le T_f,\ i = 1,...,k \\ &\sum_{i=1}^{k}\alpha_i = 1,\ \alpha_i \ge 0,\ i = 1,...,k \end{aligned} \quad (19)$$

When the upper-level programming achieves optimal, the most reasonable load allocation ratio α_i on each cluster head could be derived. According to the load allocation ratio α_i on each cluster head, the optimal load allocation ratio $\alpha_{i,j}$ on each intra-cluster sensor can be determined by the lower-level programming.

For the lower-level programming, the mathematical model is as follow:

$$\begin{aligned} &Min\ \ t_i\ \ i = 1,...,k \\ &subject\ to\ (Sensor\ S_{i,j}) \\ &\sum_{k=1}^{j}\alpha_{i,k}z_{i,k}T_{cm} + \alpha_{i,j}y_{i,j}T_{ms} \le t_i,\ j = 1,...,n_i \\ &\sum_{j=1}^{n_i}\alpha_{i,j} = \alpha_i,\ \alpha_{i,j} \ge 0,\ j = 1,...,n_i \end{aligned} \quad (20)$$

From the above, a bi-level programming model is constructed in the synthetic Problems of task scheduling for wireless sensor networks. The most reasonable load allocation ratio α_i on each cluster head could be fixed by the upper-level programming, and the lower-level programming established the most suitable load allocation ratio $\alpha_{i,j}$ on each intra-cluster sensor. In the above programming, α_i and T_f are the target function and the decision variable of upper level respectively, and

$\alpha_{i,j}$ and t_i are the target function and the decision variable of lower level respectively.

On minimum makespan T_f as the target function, the most reasonable load allocation ratio α_i on each cluster head as the decision variable, the programs to realize the optimization hauling project of minimum expenses and to output various forms are compiled according to the demand.

V. Wireless Energy Use

In this section, the energy model of the OTSA-WSN algorithm is presented in detail and the equations of energy consumption of individual sensor nodes are derived. The model is based on first-order radio model [10].

There are three kinds of energy consumption in the wireless sensor network: measurement, data fusion, and communication. Because nodes in the sensor networks cooperate with each other via data transmission, energy consumption of communications exists in sensor nodes, cluster heads and sink. It is not necessary for cluster heads and sinks to perform any sensing task. Thus, there is no energy cost for cluster heads due to the measurement of these nodes, while the additional energy cost of cluster heads attributes to data fusion. The energy to sense, fuses, and transmits a unit sensory data are denoted by e_s, e_p and e_{tx}, respectively. Sensor nodes also consume the energy of e_{rx} to receive one unit of data. The distance between the sender and the receiver is d.

The energy use for each kind of nodes is outlined as follows:

Energy use for individual sensor nodes j in cluster i:

$$E_{i,j} = \alpha_{i,j}(e_s + e_{tx}d^2), i = 1,\cdots,k,\ j = 1,\cdots,n_i \tag{21}$$

Energy use for individual cluster head:

$$E_i = \alpha_i(e_{rx} + e_p + \varphi_i e_{tx}d^2),\ i = 1,\cdots,k \tag{22}$$

Energy use for sink:

$$E_{SINK} = \sum_{i=1}^{k} \alpha_i \varphi_i e_{tx} \tag{23}$$

VI. Performance Evaluation

In this section, we investigate the effects of different measurement/communication speed under homogeneous network environment on the total task finish time (makespan) and energy consumption of every intra-cluster nodes, and compare the 2-sinks model to the traditional single sink structure.

In the simulation, the following energy parameters are adopted: transmitting a unit of sensor reading over a unit distance takes e_{tx}=200nJ, receiving one unit of sensor reading consumes e_{rx}=150nJ, measuring one unit of sensor reading needs e_s=100nJ, fusing one unit of observation consumes e_p=20nJ and the distance between the sender and the receiver is d=100m. There are 30 sensor nodes in each cluster.

The simulation results are shown in Figure 5 to Figure 7.

Firstly, the makespann against the number of clusters are plotted in Fig. 4. In Fig. 4(a), the value of measurement speed is chosen from 0.8 to 1.6, while communication speed is fixed to 1.0. This figure shows that measurement speed almost does not affect the makespan because sensing takes a small fraction of the entire execution time. Fig. 4(b) shows that when the communication speed of nodes increases, the makespan of a given task is reduced. It can be found that the five lines in Fig. 4(b) converge when the number of clusters becomes large.

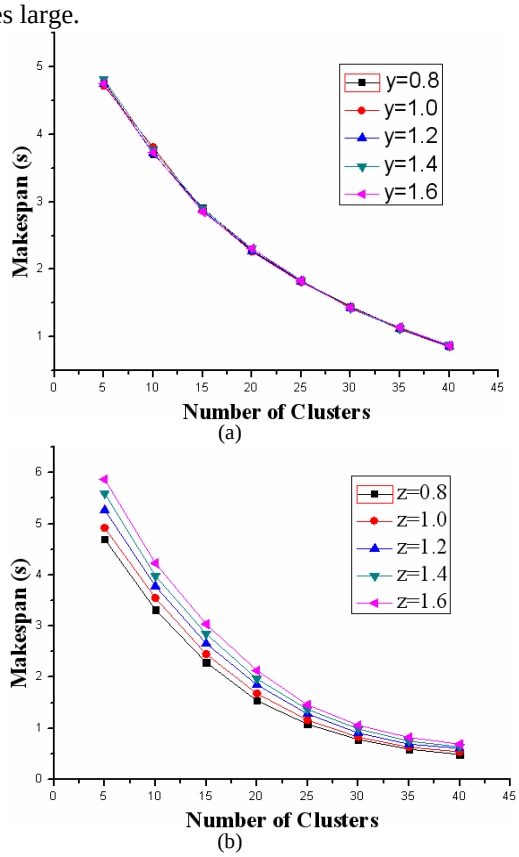
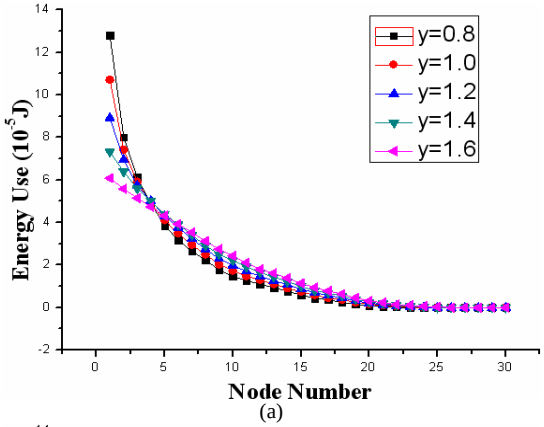


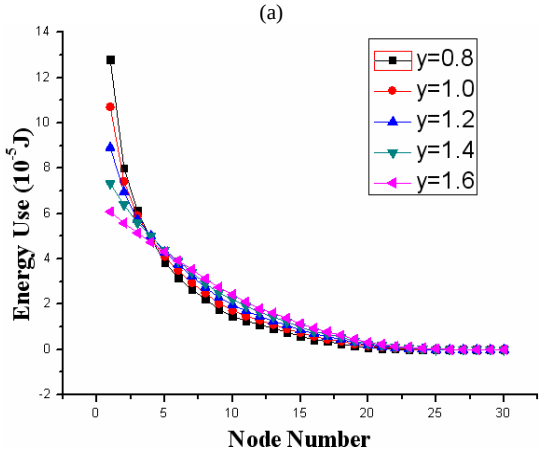
Figure 5. Impact of measuring speed and bandwidth on the makespan

Next, the second simulation is about the energy consumption of intra-cluster nodes. Sinks and cluster heads are not taken into account because generally, sinks has no energy constraint and the chosen cluster heads have the possibly enough energy. The network is configured with 20 clusters. Without loss of generality, the intra-cluster nodes in the first cluster are chosen to study the energy consumption, as shown in Fig.5. Fig. 5(a) shows the higher the intra-cluster node's measuring speed, the more evenly the tasks allocated to each nodes, hence the smaller the energy consumption of the nodes. Fig. 5(b)

presents the larger communication speed between senders and receivers, the smaller the energy consumption of the intra-cluster nodes.



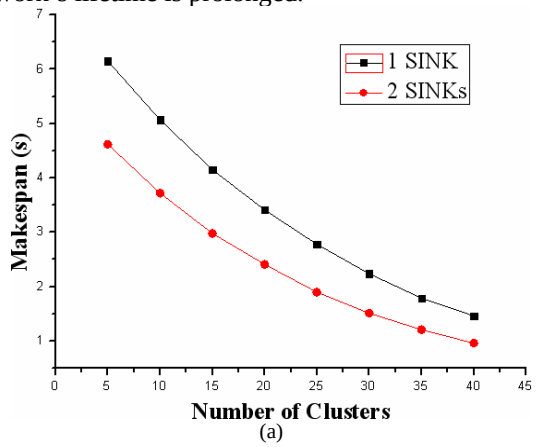
(a)



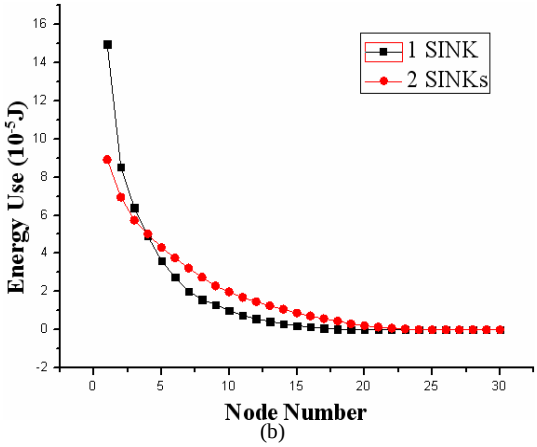
(b)

Figure 6. The impact of measuring speed and bandwidth on the energy consumption in intra-cluster nodes

Then, Fig.6 reflects the comparison of time-consuming and energy-consuming of two network architecture in dealing with the same task. In the simulation, we supposed that: $y=z=w=1.0$. As can be seen from Fig. 6(a), the task completion time is reduced by 20% in network with 2 sinks due to better computation and communication overlap. Fig. 6(b) shows that the energy-consuming of sensors is more balanced, so the network's lifetime is prolonged.



(a)



(b)

Figure 7. Comparison of time-consuming and energy-consuming of the two network architecture

VII. CONCLUSIONS

As the nodes in wireless sensor network have limited energy, so the tasks should be completed as soon as possible. In this paper, we present a new task scheduling strategy in heterogeneous clustered wireless sensor networks with multiple sinks based divisible load theory, to solve the problem how to complete the tasks within the possibly shortest time. In this strategy, the tasks are distributed to wireless sensor network based on the processing and communication capacity of each sensors by multiple sinks. After received the sub-tasks, the intra-cluster sensors perform its tasks simultaneously, and send its results to cluster head sequentially. By removing communications interference between each sensor, reduced makespan and improved network resource utilization achieved. Cluster heads send fused data to sinks sequentially after fused the data got from intra-cluster sensors, which could overlap the task-performing and communication phase much better. The strategy consists of two phases: intra-cluster task scheduling and inter-cluster task scheduling. Intra-cluster task scheduling deals with allocating different fractions of sensing tasks among sensor nodes in each cluster; inter-cluster task scheduling involves the assignment of sensing tasks among all clusters. Solutions for an optimal allocation of fraction of task to sensors in heterogeneous wireless sensor networks are obtained via closed-form solution and bi-level programming solution, respectively.

ACKNOWLEDGMENT

The authors thank the editors and the anonymous reviewers for their valuable comments that helped to improve the paper. The work was supported by the National Natural Science Foundation of China (No.60972047), and the 111 project (No.B08038).

REFERENCES

[1] V. Shah-mansouri, A. Mohsenian-rad, "Lexicographically Optimal Routing for Wireless Sensor Networks With Multiple Sinks," *IEEE Transactions on Vehicular Technology*, 2009, 58(3): 1490 – 1500.
[2] K. Yuen, L. Ben, B. C. Li, "A Distributed Framework for Correlated Data Gathering in Sensor Networks," *IEEE Transactions on Vehicular Technology*, 2008, 57(1) :578 – 593.
[3] V. Bharadwaj, D. Ghose, T. G. Robertazzi, "Divisible load theory: A new paradigm for load scheduling in distributed systems," *Cluster Computing*, 2003, 6(1), pp.7-18.
[4] M. Moges, T. G. Robertazzi, "Wireless sensor networks: scheduling for measurement and data reporting," *IEEE Transactions on Aerospace and Electronic Systems*, 2006, 42(1), 327-340.
[5] H. Liu, X. Yuan, M. Moges, "An Efficient Task Scheduling Method for Improved Network Delay in Distributed Sensor Networks," *In Proceedings of TridentCom*, 2007, (pp.1-8). Orlando, FL, US: IEEE.
[6] H. Liu, J. Shen, X. Yuan, M. Moges, "Performance Analysis of Data Aggregation in Wireless Sensor Mesh Networks," *In Proceedings of Earth & Space 2008*, (pp.1-8), Akron, OH, USA: IEEE.
[7] C. Kijeung, T. G. Robertazzi, "Divisible Load Scheduling in Wireless Sensor Networks with Information Utility Performance," *In Proceedings of IPCCC*, 2008, (pp.9-17), Austin, Texas, USA: IEEE.
[8] Z. Zeng, A. Liu, D. Li, "A Highly Efficient DAG Task Scheduling Algorithm for Wireless Sensor Networks," *In Proceedings of ICYCS*, 2008, (pp.570–575). Zhang Jia Jie , Hunan , China: IEEE.
[9] J. Lin, W. Xiao, F. L. Lewis, et al, "Energy-Efficient Distributed Adaptive Multisensor Scheduling for Target Tracking in Wireless Sensor Networks," *IEEE Transactions on Instrumentation and Measurement*, 2009, 58(6), pp.1886 – 1896.
[10] W. Heinzelman, A. Chandrakasan, "An application-specifid protocol architecture for wireless microsensor networks," *IEEE Transaction on Wireless Communications*, 2002, 1(4), pp. 660-670.
[11] C. Kijeung, T. G. Robertazzi, "Divisible Load Scheduling in Clustered Wireless Sensor Networks," *Stony brook university*, 2009.
[12] X. Li, H. Kang and J. Cao, "Coordinated Workload Scheduling in Hierarchical Sensor Networks for Data Fusion Applications," Journal of Computer Science and Technology, vol. 23, 2008, pp. 355-364.
[13] X. Li, X. Liu and H. Kang, "Sensing Workload Scheduling in Sensor Networks Using Divisible Load Theory," *The 50th Annual IEEE Global Telecommunications Conference*, Washington DC, 2007, pp. 785-789.
[14] X. Tang, J. Xu, "Optimizing Lifetime for Continuous Data Aggregation With Precision Guarantees in Wireless Sensor Networks," *IEEE/ACM Transactions on Networking*, 2008, 16(4), pp. 904 – 917.

Liang Dai was born in 1981. He was graduated from Xidian University with Ph. D in Communication and Information System in 2011. He is with Chang'an University from 2011.

His research interests include wireless sensor networks and digital signal processing in mobile communication.

Hongke Xu received the B.Sc. degree in traffic control and management from Chang'an University, Xi'an, China, in 1985, the M.Sc. degrees in computer engineering from Xidian University, Xi'an, in 1993, and the Ph.D. in traffic control and management from Chang'an University.

He is currently a Professor with the School of Electronic and Control Engineering, Chang'an University. His major research interests are in the fields of traffic control and management, and ITS.

Ting Chen was born in 1982. She was graduated from Xidian University with Ph. D in Communication and Information System in 2011. She is with Chang'an University from 2011. Her research interests include wireless communication networks, corss-layer design, QoS guarantee mechanism, and etc..

Visual Important-Driven Interactive Rendering of 3D Geometry Model over Lossy WLAN

Bailin Yang
Department of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
Email: {ybl}@mail.zjgsu.edu.cn

Zhiyong Zhang and Xun Wang
Department of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China
Email: { zzy, wx}@mail.zjgsu.edu.cn

***Abstract*—In this paper, we propose a visual important-driven interactive rendering method for 3D model over 802.11 WLAN for overcoming the shortcomings of wireless network's narrow bandwidth, high transmission error rates and mobile devices' low power supply. This paper first proposes an efficient simplification method based on an improved visual important region detection technique. Then, we develop an efficient hybrid FEC and MAC-Lite model transmission protocol which will transmit the model data by their importance respectively. Finally, we propose a real-time interactive rendering method by an efficient model coding. Experimental results demonstrate that we can obtain better rendering result among lossy environment and gain real-time interactive rendering result.**

***Index Terms*—Visual detection, Model Simplification, FEC, MAC_Lite, interactive rendering**

I. INTRODUCTION

With both the mature of the mobile network infrastructure and wide use of mobile handheld devices, 3D applications based on mobile devices among wireless network have got rapid development. However, the shortcomings of narrow bandwidth, high transmission error rates among wireless network and the limitations of limited power supply and low computing performance for mobile devices cannot meet the requirements for the real-time interactive rendering of the 3D model on mobile devices.

Recently, the technologies for model transmission and interactive rendering have been received more attentions from researchers. The typical technique is the progressive coding and transmission method [1] for 3D model that can transmit the model data on demand of the user's quality requirement for 3D model. Usually, the model should be simplified into progressive model. Thus, how to simplify this model efficient is more important. Now, researchers are paid more attention to simplify the model from the aspect of visual optimization.

As we knew, lossy wireless network is considerably different from wired networks. The transmission over lossy wireless links stays challenging due to narrow bandwidth, fading and obstacles which result in high transmission error rates.

To address these problems, retransmission is scheduled. Evidently, this retransmission mechanism affects the network's throughput and end-to-end delay badly. Factually, most dropped packets are caused by bit-errors in the frame during transmission in the wireless network and a packet with this kind of errors can still be utilized for some error-tolerant transmission such as audio, video and graphics. In this regard, a new transport layer protocol called UDP Lite[2] which is tailored for real-time applications over error-prone networks has been proposed. Unlike UDP, UDP Lite allows for partial checksums that only covers part of a datagram, and will therefore deliver packets that have been partially corrupted. However, in WLAN, lots of corrupted packets are discarded in the MAC layer without reaching UDP layer. The CRC in the MAC layer also should be altered to allow corrupted frames being passed to higher layers, which is known as Mac-Lite [3]. Similar to UDP-Lite, the coverage of Mac-Lite's checksum can be set freely.

What's more, the key problem of interactive rendering for 3D model on mobile device is how to decrease the transmitted data during the transmission and lower the computing and rendering task in mobile device.

In this paper, we propose a simplification method based on an improved salient detection method. In the implementation, we present some optimization techniques to accelerate the progressive mode reconstruction. Then, we propose an efficient hybrid FEC and MAC-Lite model transmission protocol which will transmit the important graphics data and less important data by modified protocol based on FEC and sensitive data dependent Mac-Lite respectively. Finally, a real time transmission and interactive rendering method by an effective model coding method is proposed. In this method, the mobile client only executes the low-lever rendering operation such that shortens the waiting time before the rendering.

The rest of the paper is organized as follows. Section II is the related work. Section III describes perceptually-based progressive model construction method. Section IV

Corresponding author: Wang Xun

doi:10.4304/jnw.6.11.1594-1601

designs a hybrid transmission protocol. Section V explains the progressive transmission and interactive real time rendering method. The experimental result is shown in section VI. Finally, we summarize our work in section VII.

II. RELATED WORK

A. Simplification and Visual important Computation

Researchers have presented many simplified methods which aim to decrease the visual quality differences of the simplified model and source model. The typical method are QEM[4] and improved QEM[5]. However, those methods focus only on the geometry features and the visually important area cannot be preserved longer while simplifying.

In order to solve this problem, the simplified methods from the aspect of visual optimization are proposed. Lindstrom [6] proposed the simplified method based on the CSF model. Luebke and Hallen[7] proposed a method that employed the visual psychology model to control the 3D model simplification procedure. Qu [8] etc. al proposed a visual mask computing method, which would direct the simplifying for the textured 3D model. Unfortunately, above methods did not consider the topology information of the mode itself.

As we knew, the idea of salient region has been developed to help identify distinct spatial regions from their surrounding neighborhoods. Also, the saliency technique has been applied to 3D models. In general, the detection of 3D salient regions can be treated as an extension of identifying salient regions on a 2D image. Based on the model developed for 2D image [9], Lee et al. [10] proposed the idea of mesh saliency as a measure of regional importance for 3D models based on the center-surround mechanism and feature integration theory [11]. This method focuses only on the curvature of each vertex. Gal et al. [12], however, computed a salient region based not only on the curvature, but also the variance of curvature, the number of curvature changes and the size relative to the whole object. By these saliency methods, they achieved better simplification result.

B. UDP_Lite and MAC_Lite Transmission Protocol

Recently, there are lots of literatures about UDP Lite or Mac-Lite applications. In Ref.[13-17], UDP Lite is deployed to transmit multimedia data. Errors in the sensitive part of a multimedia packet should result in dropped packets, while errors in the insensitive part are forwarded to application layer. To allow packets containing errors to be forwarded to the UDP layer, the 802.11 MAC level errors checking feature is completely disabled. Regarding to WLAN, however, the MAC level checksum cannot be completely disabled due to the high bit error rates during transmission. Moreover, the MAC layer plays much important role than UDP layer because the data can be forwarded to the destination by the MAC protocol even without UDP protocol in WLAN.

Mac-Lite is used to transmit voice in WLAN [18]. The checksum only covers headers data such as MAC header, IP header and UDP header, but for voice data, no checksum is applied on it. The experiments results show that compared with the original CRC checking scheme, better performance of networks is achieved. In [19], the authors use different coverage of MAC layer's checksum to transmit speech and compare their experimental results. For video transmission, video coding technology is adopted to divide the video data into different parts according to their different importance and then use Mac-Lite to transmit it [20]. In order to transmit data correctly and fast by Mac-Lite, the forward error-correcting (FEC) technology is used [21]. If the partial checksum detects errors in important data such MAC header data, no retransmission but correcting it instead.

However, there is no discussion in literatures about adopting Mac-Lite or its modified version to transmit mesh of 3D model in WLAN.

C. Streaming and Interactive Rendering

Different from the desktop pc device over wired network, the main shortcoming for the mobile device is the limited power supply and computing ability. Therefore, the key problem of 3D model rendering on mobile device is how to decrease the transmitted data from the server to client and lower the computing and rendering task in mobile device.

Luo[22] proposed the progressive transmission and model simplification methods for mobile device. However, these methods needs local reconstruction operations which will take up lots of computing costs and cause the rendering delay at the client. Actually, this method is not good for mobile device. Thus, we can translate the 3D model into image or video and adopt the successful image or video coding technique to transmit the model data. For example, reference [23] and [24] respectively proposed the MPEG-4 coding and JPEG 2000 coding methods to transmit the 3D model. Unfortunately, these methods are not suitable for the application of 3D model representation in mobile e-commerce because these method can not obtain the whole 3D model data but the static images.

III. PERCETUALLY-BASED PROGRESSIVE MODEL CONSTRUCTION

Loosely speaking, a salient region of a model is the area that is distinct from its surroundings. In this paper, we propose a saliency computation method to effectively obtain salient regions of a model. Similar to [10], the saliency map is created by center-surround mechanism. Usually, center-surround differences are calculated as an across-scale difference between coarse and fine scales. For each scale, a filter window to include neighbouring vertices samples should be designed.

The implementation of our saliency computation method is depicted as follows:

Step 1: Compute the mean curvature MC at each vertex v_i ($i = 1...n$, n is the number of vertices of the mesh).

Step 2: Define the local filter window for vertex v_i and choose its neighboring vertex set $NS(v_i)$.

Step 3: According to $NS(v_i)$, calculate the Gaussian weighted average $GW(v_i)$ at different scales.

Step 4: Get the difference $DGW_{m,n}(v_i)$ between the two scales m and n for v_i and then compute the geometry feature map $G_l(m,n)$.

Step 5: Make use of the non-linear suppression operator to combine the feature maps $G_l(m,n)$ into the final geometry saliency map $\overline{GF}$.

In step 1, we use the method proposed in [15] to get the MC. Then, we utilize the local filter design method to acquire $NS(v_i)$.

Given the $GW(V_i)$ of each vertex and radius r, its Gaussian-weighted average is

$$GW(V_i,r)=\frac{\sum_{x\in W_r} MC(x)\exp\left[-\|x-V_i\|^2/(2r^2)\right]}{\sum_{x\in W_r}\exp\left[-\|x-V_i\|^2/(2r^2)\right]} \tag{1}$$

Then, each feature map $G_l(m,n)$ is calculated as:

$$G_l(m,n)=\left|GW(V_i,r_m)-GW(V_i,r_n)\right| \tag{2}$$

Finally, those feature maps will be combined into one geometry map by the nonlinear suppression operator. We improve the method proposed in [10] by not only acquiring the block salient region but also the details, such as the exact boundary of the salient region. In our case, we take the mean curvature of each vertex into consideration while combining the above four scales into the final salient region. We also adjust the weight α and β_l to get the final geometry feature map using the following formula.

$$\overline{GF}(V_i)=\alpha \mathrm{N}(MC)+\sum_{l=1}^{4}\beta_l \mathrm{N}(G_l) \tag{3}$$

To preserve the visually important vertices longer, we will adopt above salient detection method. By this method, we can get the salient importance values $S(V_i)$ of each vertex. We have modified the QSlim algorithm [2] by weighting the quadrics with mesh saliency.

After the creation of simplification metric, the collapsed queue(CQ) is initialized and the collapsed operations are executed for importing the new vertices and edge pairs(EP). Thus, we can build the full collapsed queue namely the vertex split list. Meanwhile，two data structures of the collapsed record stack and split record stack will be introduced to meet the needs of interactive rendering. In the following, we explain the optimization tactics in the implementation.

1) Initialization of the CQ. According to $w(v_i)$, the suitable EPs are chosen and the CQ is built. Usually, the CQ adopts the heap data structure. This structure is simple from logical. However, experimental result demonstrates this method is slowly while lots of EPs are appeared. In this paper, we will adopt the dynamical array structure. Different from the heap data structure, the sorting operation is executed after all the insertion operations. By this method, the whole sorting time is saved.

2) Executing the collapsed operation and building the final CQ. From the initial CQ, we can find the collapsed edge(CE) with the smallest value of $w(v_i)$ and generate the new vertices and EP. Clearly, the new vertices and EP will effect the sorting operation of this dynamical array. In our implementation, the CQ will not be sorted immediately. Factually, these new vertices and EP will effects the sorting of dynamic array. Thus, we do not carry out the sorting for the CQ immediately but pushes these collapsed vertices into CQ and sorts the dynamic array again. Experimental result shows that this method will improve the collapse speed 30%-40% and does not affect the model's simplification result.

3) Introducing of collapsed record stack (CRS) and split record stack (SRS). In order to achieve the interactive rendering, the server should provide the function of switch between different resolutions of model quickly. However, the existed simplification method [1] will consume a great deal of collapse and split operations while switching from one resolution to another resolution. Therefore, our method presents the CRS and SRS data structure, which will record each collapsed and split operation and push them into the CRS and SRS while executing the simplification operation. While the model needs the switch between different resolutions, we just fetch these records from the CRS or SRS and execute the corresponding rendering operation.

IV. TRANSMISSION PROTOCOL DESIGN

The basic idea of our modified protocol is to formulate the transmission protocol according to the different importance of the 3D model.

Progressive Mesh, as a good solution to the transmission of 3D models over network, is represented by a base mesh M0 followed by an ordered list of vertex split (VSplit), which is in the form of {M0, {VSplit 1,VSplit 2, . . . ,VSplitn}}. There exists dependency relationship among these VSplit operations. In practice, these VSplits will be packed into packets for transmission over networks. Hence, these packets also have dependency relationship. Consequently, VSplits could not be rendered unless their dependent VSplits arrived at the client. If some of the received packets are dependent on the lost packet, the client will endure a rendering delay since the lost packet retransmission will be invoked. On the contrary, if no or just a small number of the received

packets are dependent on the lost packets, the client could render more vertices at a given period of time and the delay will be reduced.

Thus, in our past work [29], we have presented a novel packetization scheme that is to decrease the dependencies among packets. In this packetization method, two steps will be performed. First, a Non-Redundant Directed Acyclic Graph (NR-DAG) will be constructed to encode all the necessary dependencies among the VSplit operations. Second, a Global Graph Equipartition Packing Algorithm (GGEPA) is applied to minimizing the dependencies among different partitions while separating the whole dependency DAG into k equal size partitions.

Though this method can decrease the dependencies among these packets, the dependencies are still existed. If the dependencies between one packet with other packets are higher, more VSplits, which are included in the dependent packets, should wait this packet be arrived at the client side. Thus, we here regard this packet are rendering-importance packet. As we knew, if the VSplits belongs to the base mesh or upper levels, the packets that contain these VSplits are also rendering important packets. Unfortunately, these packets maybe not have many packets that dependent on them. To assign this kind of packets those have many dependent packets and the packets those have in the upper level of our model, we will deal with them in a unified way.

In our GGEPA, we will record each packet's dependencies noted as PD. As we knew, the NR-DAG we built is a graph. We will translate them into a tree structure thus all nodes will have been arranged as level by level. Manifestly, the nodes in the upper level are the parents of the lower level's nodes. To assign each packet with a rendering important (RI) value, we browse this tree level by level with depth-first visiting method and calculate each packet's RI value. While finishing this depth-first visiting, we can obtain each packet's RI. However, this method only records the RI between neighboring levels. To obtain the RI among all the levels, we should add all the children's RI into their parents. Now, we can give each node with an accurate RI value.

While the packet is packed into frames, we will assign them with the perceptually importance value RI. It means if the frame's RI is high, it is important data. According to the different importance of the frame, the MAC layer uses two different ways, MAC-FEC and MAC-Lite protocol, to transmit them respectively. The details of both methods are as follows.

a) **MAC-FEC**. For visual important data, to ensure the data transmitted correctly, the forward error-correcting (FEC) technology is employed in the MAC layer, as shown in Fig.1. When a frame has arrived, checksum mechanism is used to check it. If the checksum failure, retransmission is not used but FEC for error correction. While using FEC, the actual data transmitted is larger than the original base mesh data because the additional redundancy data is added. However, the ratio of amount of base mesh data in the entire model is much low, so using this method to transmit the base model does not affect the speed of the entire model transmission obviously. By this method, it can guarantee that a base model is transmitted correctly.

b) **Mac-Lite**. For less important model data, we adopt the Mac-Lite rather than traditional MAC.

However, the key of the Mac-Lite is to set the coverage of checksum for a frame. Usually, all headers information should be covered because of the following reasons:

(1) If there are bit-errors in MAC header, the frame may be sent to other destinations because of source and destination address information in it.

(2) If there are bit-errors in IP header, the packet will be discarded when it is forwarded in the IP layer, because the IP layer also has checksum mechanism which covers the IP header.

(3) If there are bit-errors in the UDP header, the packet may be transmitted to other applications because the UDP header contains the source and destination port information.

Nevertheless, the checksum just covering the headers data is not enough. Factually, the data can be divided into topology data and geometry data. While the topology information is lost during transmission, visual errors to the rendered model, such as the surface self-intersection, will be incurred. Thus the topology data should be transmitted as safely as the frame headers information. Therefore, the coverage of Mac-Lite checksum is the summary of MAC header (28bytes), IP header (20bytes) and UDP header (8bytes) and topology data as shown in Fig. 2.

Fig. 1. MAC frame with FEC

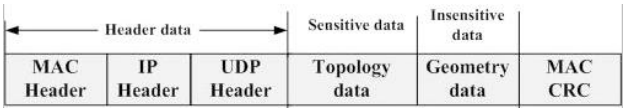


Fig. 2. MAC frame with VSplits

V. Progressive Transmission and Interactive Rendering

In this section, we presented a progressive transmission and interactive rendering method on mobile devices based on above progressive model. Different from the typical interactive rendering method, we will design a tactics for computing task allocated on the server side and client side respectively.

A. Computing Task Assign and Rendering

To reduce the mobile client's computing and storage burden, we will save and run the multi-resolution model in the server side as client does. For the client, it would only execute the rendering operation. First of all, we construct the multi-resolution model with the method in section 2, and the server will run and save this model. At the same time, using the CRS and SRS appeared in the

multi-resolution modeling to achieve the rapid switch of different resolution for the model. Secondly, when the client needs certain resolution model, the server makes use of the CRS and SRS to obtain the collapsed edge and split vertex, which will be formed into vertex index array and sent to the client. Finally, the client will execute the rendering procedure after obtaining this vertex index array and the data stored at the client already.

In the following, we will describe the real-time interactive rendering procedure from the aspect of the server and mobile client.

1) Server store the multi-resolution model data and response the client's request.

a) Responses the client's request and sends the model's geometry information to mobile client.

b) Server run and store the multi-resolution model the client needed.

c) Response the request for the certain resolution model from the client. Making use of the CRS and SRS to compute the multi-resolution model and get the vertex index array at this certain resolution.

d) Sends the vertex index array to mobile client.

e) Return to Step *c* and waits the request for another resolution model of the client.

2) Mobile client make requests for the server to obtain the certain resolution model and rendering locally according to the returned model.

a) Client makes a rendering request to server.

b) Receives the model geometry information and stores them locally.

c) According to the user's request, sends the rendering request for certain resolution.

d) Receiving the vertex index array from the server side and rendering them without any reconstruction operation locally.

e) Return to Step *c*;

During the interactive rendering procedure, it can be seen that our method just transmits a few vertex index array such that the total data over the network is decreased manifestly. More importantly, our method need not perform the local reconstruction operation thus that the waiting time for rendering of the model is decreased aggressively (see the experimental result as shown in Table 2). By this method, we can achieve interactive rendering for 3D model at the mobile client side.

VI. Experimental Result and Discussion

A. TestBed Design

We adopt the C/S model to validate our method. The PC server will transmit the data to mobile client through the D-Link wireless router among the Wireless 802.11.b network. The average network bandwidth is 0.5MB/s. The network layout is shown in Figure 2. At the server side, the multi-resolution model is created by our simplified method. In the mobile, we will adopt the rendering library M3D [25]we developed before which conforms the OpenGL ES specification. As we knew, the wireless network is varied so that the experimental result for the transmission time are measured as the average of 10 times.

To verify our transmission performance, what's more, we use the ns-2 to build the simulation testbed. In our simulation model, the nodes have no mobility. This is primarily because our interest in his paper is to focus on the effectiveness of the modification of the MAC layer. Three nodes A, B and C are used as an ad hoc network and the topology is shown in Fig 2. In order to set the different loss rate in the PHY layer, the Gilbert Error model [27,28]should be added. The bit errors generated by this model are introduced to MAC frame.

The test models in our experiment that are stored by PLY format is Lauran, Bunny, and Horse as shown in Fig 4. The number of total vertices and corresponding storage space of each model, the ratio for the geometry information and the ratio for the topology information is shown in Table I.

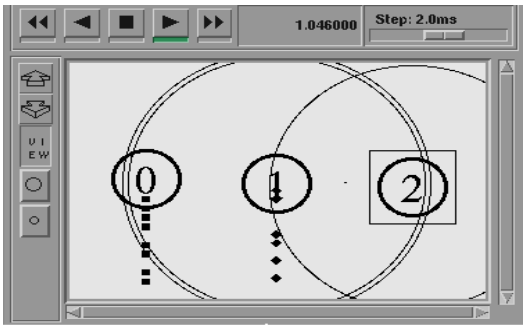


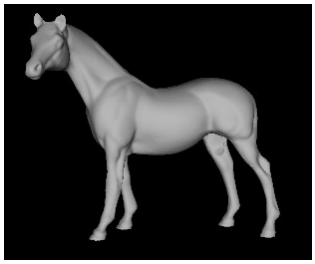
Fig. 3. Topology in NS2.33

B. Transmission Protocol Performance Analysis

Before the transmission of model, we will adopt the presented method in this paper to encode the 3D model with base mesh and a sequence of *VSplit* . Each *VSplit* including only the basic topology and geometry information is 30-byte quantity. Thus, one packet whose max size is 512-byte defined in this paper will contain roughly 17 *VSplit* .

(a) Laurana

(b) Bunny



(c) Horse

Figure 4. Test models

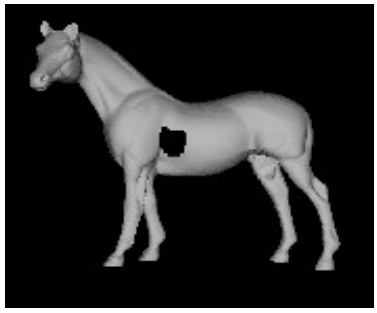
We will report the experimental result for Horse model, which will be divided into 15233 *VSplit* . What's more, the size of Horse model's base mesh and details *VSplit* is about 0.31M Bytes and 2.76M Bytes respectively.

In this paper, we only compare model quality received at the client side while the packet loss rate is 5%. Fig. 5 demonstrates the close-up wireframe view of horse model. In order to show comparative results obviously, we have to control the location of the packet loss rate. Then packets losses happen from the same area on the model. In these simulations, when part of a packet is lost and unrecovered, the packet is discarded. However, in order to show how many packets are lost at different packet-loss rates, we used a visualization trick by discarding part of the received packets in case the packets were lost. It can be seen that our transmission method can achieve the rendering result than the traditional MAC with retransmission mechanism.

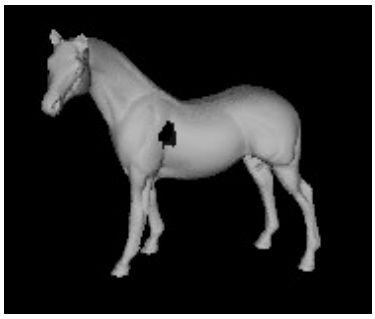
C. Comparsions for Interactive Rendering

This subsection compares the proposed interactive rendering method with the typical method in reference [1]. We will compare the sum of transmission time from server to client and the rendering time on mobile device while transmitting the base model, which occupied 20% of the total model data, 60% and 100% of the model data. In the typical method, the whole time includes the transmission time, local reconstruction time and rendering time. Fortunately, our method's whole time just includes the transmission time and rendering time. For the certain resolution model, the rendering time is assured so that the rendering time is not listed in Table II.

Since our method adopts the smart coding technique and computing task assignment, the transmitted data over network can be decreased and the local reconstruction time would be cut off. Thus, our method can achieve interactive rendering result. For example, while the full model are transmitted from the server to client and display, the time our method consumed are 41%，21% and 37% in contrast to the typical method.



(a) MAC with retransmission



（b）Ours

Fig. 5. Close-up wireframe view of horse model when loss rate is 5%.

VII. CONCLUSION AND FUTURE WORK

This paper proposes a visual important interactive rendering method for 3D model over 802.11 WLAN. This paper first proposes an efficient simplification method based on an improved saliency detection technique. By the introducing of the data structure including collapsed record stack and split record stack, we can finish the construction of multi-resolution model. Then, we develop an efficient hybrid FEC and MAC-Lite model transmission protocol which will transmit the model data by their importance respectively. Finally, we propose a real-time interactive rendering method by an efficient model coding. For decreasing the transmitted model data over the wireless network, we proposed an efficient model coding method and computing task assign method. By this method, we can transmit the model from the server to client quickly. What's more, the mobile client can save the local reconstruction operation which would consume lots of CPU resource.

In the future work, we will adopt the geometry compression technique which will decrease the model data aggressively. Also, the dynamical transmission mechanism that just transmits the part of model user can see will also reduce the model data transmitted over the wireless network.

TABLE I. The Ratio for Geometry and Topology Information of Each Model

3D Models	Vertices number/total data（KB）	Ratio for Geometry Inforation	Ration for topology information
Laurana	14499/1334	32.5%	67.5%
Bunny	20376/2963	32%	68%
Horse	16029/1382	33%	67%

TABLE II. The Comparsions for Our Method and Typical Method of Tranmission Time and Rendering Time

	Typical Method （Full model transmission time and model reconstrucion time）	Our method Full model transmision time （the ratio comparing for the typical method）
Laurana	13889	5768 （41%）
Bunny	18892	6891 （37%）

ACKNOWLEDGMENT

This work is supported by in part Zhejiang natural science key foundation (Grant No. Z1080232, Z1101340), national natural science foundation of China (Grant No. 60873218), and the Scientific Research Fund of Zhejiang Provincial Education Department (Z201018041).

REFERENCES

[1] H. Hoppe, “Progressive meshes,” *In: SIGGRAPH 96 Proceedings*, pp.99–108, 1996.

[2] http://www.ietf.org/rfc/rfc3828.txt.

[3] S. A. Khayam, S. Karande, M. Krappel, and H. Radha. Cross-layer protocol design for real-time multimedia applications over 802.11b networks. Proc. IEEE International Conference on Multimedia and Expo, July 2003, vol.2, pp. II- 425-8.

[4] M. Garland and P. Heckbert, “Surface simplification using quadric error metric,” *In: Proceedings of ACM SIGGRAPH’97*, pp.209-215, 1997.

[5] P. Lindstrom, G. Turk, “Fast and memory efficient polygonal simplification,” *In: Proceedings of the IEEEVisualization’98*, pp.279-284,1998.

[6] P. Lindstrom, “Model simplification using image and geometry-based metrics”, *PhD Thesis*, Georgia Inst. of Technology, 2000.

[7] D.P. Luebke and B. Hallen, “Perceptually-driven simplification for interactive rendering, ”*In: Proceedings of 12th Eurographics Workshop Rendering Techniques (EGRW ’01)*, pp.223-234, 2001.

[8] L.J. Qu, Gary W. Meyer, “Perceptually guided polygon reduction,” *IEEE Transactions on Visualization and Computer Graphics*, vol.14(5), pp. 015-1029, 2008.

[9] L. Itti, C. Koch, E. Niebur “A model of saliency-based visual attention for rapid scene analysis, ”*In: IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20(11), pp.1254-1259, 1998.

[10] C.H. Lee, A. Varshney, and D.W. Jacobs, “Mesh saliency,” *In: Proceedings of ACM SIGGRAPH ’05*, pp. 659-666, 2005.

[11] A. M. Treisman, G. Gelade “A feature-integration theory of attention”, *Cognitive Psychology*, vol. 12(1), pp. 97-136, 1980.

[12] R. Gal, D. Cohen-OR, “Salient geometric features for partial shape matching and similarity,” *ACM Transaction on Graphics*, vol 25(1), pp.130-150, 2006.

[13] L. Larzon, M. Degermark, and S. Pink. “UDP lite for real time multimedia applications”. *in Proc. IEEE International Conference on Communications*, June 1999.

[14] A. Singh, A. Konrad, and A. D. Joseph. “Performance evaluation of UDP lite for cellular video”. *Proc. ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video*, June 2001.

[15] H. Zheng and J. Boyce. “An improved UDP protocol for video transmission over Internet-to-Wireless networks.” *IEEE Transactions on Multimedia*, Sep. 2001, vol. 3(3), pp. 356-365.

[16] H. Zheng. “Optimizing wireless multimedia transmissions through cross layer design”. *Proc. IEEE International Conference on Multimedia and Expo*, July 2003.

[17] S. A. Khayam, S. Karande, H. Radha, and D. Loguinov. ”Performance analysis and modeling of errors and losses over 802.11b LANs for high bit rate real-time multimedia”. *Signal Processing: Image Communication*, August 2003, vol 18(7), pp. 575-595.

[18] A. Servetti and J. D. Martin. “802.11 MAC protocol with selective error detection for speech transmission”. *Proc. 3rd International Workshop on QoS in Multiservice IP Networks*, Catania, Italy, February 2005, pp. 509-519.

[19] I.Chakeres, H.Dong, E.M.Belding-Royer, A.Gersho, and J. D.Gibson. “Allowing errors in speech over wireless lans”. *In Proceedings of the 4th Workshop on Applications and Services in Wireless Networks* (ASWN), Boston, MA, August 2004, 1643-1657.

[20] E.Masala, M.Bottero, and J.De Martin. “MAC-level partial checksum for H.264 video transmission over 802.11 ad hoc wireless networks”. *Proc. IEEE 61st Vehicular Technology Conference*, May 2005, pp. 2864 - 2868.

[21] S. A. Khayam, S. S. Karande, M. U. Ilyas and H. Radha. “Header detection to improve multimedia quality over wireless networks”. *IEEE Transactions on Multimedia*, 2007, vol. 9 (2), pp. 377-385.

[22] X. N. Luo, G. F. Zheng. “Progressive Meshes Transmission over a Wired-to-Wireless Network. ” *Wireless Networks*, vol 14(1), pp. 47-53, 2008.

[23] L. Cheng, A. Bhushan, R. Pajarola, and M. E. Zarki, “Real-time 3d graphics streaming using mpeg-4,” *In: Proceedings of IEEE/ACM Workshop On Broadband Wireless Services and Application*, 2004.

[24] N.-S. Lin, T.-H. Huang, and B.-Y. Chen, “3D model streaming based on jpeg 2000”, *IEEE TCE*, , vol. 53(1), pp.182-190, 2007.

[25] Bailin Yang, Lu Ye, Zhigeng Pan, Guilin Xu. “An optimized soft 3D mobile graphics library based on JIT backend compiler,” *In: Proceedings of the 4th International Conference on Entertainment Computing（ICEC）*, Sanda, Japan, pp.67-75, 2005.

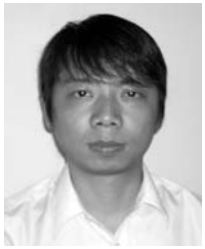
[26] G. Taubin “Estimating the tensor of curvature of a surface from a polyhedral approximation,” *In: Proceedings of IEEE International Conference on Computer Vision*, pp.902-907, 1995.

[27] E. N. Gilbert. “Capacity of a burst-noise channel”. *Bell Systems Technical Journal, 1960*, vol. 39, pp. 1253-1265.

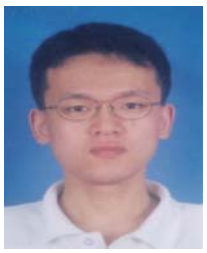
[28] E. O. Elliot. “Estimates of errors rates for codes on burst-noise channels”. *Bell Systems Technical Journal*, 1963, vol. 42, pp. 1977-1997.

[29] Bailin Yang, Frederick W.B. Li, Zhigen Pan, Xun Wang.

"An Effective Error Resilient Transmission Scheme for Progressive Mesh Transmission over Unreliable Networks". *Journal of Computer Science and Technology*. 2008, 23(6): 1015-1025.



Bailin Yang received received the Doctor's degree in department of computer science from Zhejiang University in 2007. He is a associate professor in the department of computer and electronic engineering of Zhejiang Gongshang University. His researchinterests are in moblie graphics, realtime rendering and mbile game.



Zhiyong Zhang received the master's degree in department of Mechanical and Energy Engineering from Zhejiang University, Hangzhou, P.R. China in 2001. In 2005, He received the Doctor's degree in department of computer science from Zhejiang University. Now, He is a associate professor in the department of computer and electronic engineering of Zhejiang Gongshang University. His research interests are in information retrieval, pattern recognition, and statistical machine learning.

Xun Wang received the Doctor's degree in department of computer science from Zhejiang University. He is a professor in the department of computer and elect-ronic engineering of Zhejiang Gongshang University. His research interests are in multimedia information retrieval, pattern recognition, and mobile networks. and statistical machine learning.

Secure Identity-based Threshold Broadcast Encryption in the Standard Model

Leyou Zhang
Department of Mathematical Science, Xidian University, Xi'an, 710071, China
Email: leyouzhang77@yahoo.com.cn

Qing Wu
School of Automation, Xi'an Institute of Posts and Telecommunications, Xi'an, China
Email: xidianwq@yahoo.com.cn

Yupu Hu
Key Laboratory of Computer Networks and Information
Security, Ministry of Education, Xidian University, Xi'an, 710071, China
Email: yphu@mail.xidian.edu.cn

***Abstract*—The general threshold broadcast encryption is not suitable for the networks with the constraints of computation and energy. In this paper, two constructions of the proper threshold broadcast encryption to these networks are proposed. In the proposed schemes, any user can dynamically join the system as a possible recipient, and the sender can dynamically choose the set of recipients *S* and the threshold value *t*. The new schemes achieve constant size private keys and *O*(*n*-*t*)-size ciphertexts. In addition, these schemes achieve full security in the standard model. Finally, we also show that they are provable security under *n*+1-Weak Decision Bilinear Diffie-Hellman Exponent (*n*+1-wDBDHE) assumption and the static assumptions respectively.**

***Index Terms*—Broadcast Encryption, Identity-based Threshold broadcast encryption, Dual encryption technique, Provable security, Standard model**

I. INTRODUCTION

Broadcast Encryption (BE) was introduced by Fiat and Naor in [1]. In a broadcast encryption scheme a broadcaster encrypts a message for some subset *S* of users who are listening on a broadcast channel. Any user in *S* can use his private keys to decrypt the broadcasts. Any user outside the privileged set *S* should not be able to recover the message. The threshold broadcast encryption(TBE) problem is generalization of the concept of broadcast encryption. It was first introduced by Ghodosi et al. [2]. TBE has some advantages over traditional threshold encryptions. It is specified as follows:

(1) The trusted party is eliminated and the system can be set up by individual users independently;

(2) The broadcaster can choose the privileged set and the threshold value at the time of encryption which allows a certain dynamism in the system.

Manuscript received January 1, 2011; revised June 1, 2011; accepted July 1, 2011.

Corresponding author: Leyou Zhang, Email: leyouzhang77@yahoo.com.cn

Identity-Based encryption is originally proposed by Shamir[3], which a major advantage is that it allows one to encrypt a message by using recipient's identifiers such as an email address. Now it has been an active area. The first practical identity-based encryption (IBE) scheme was proposed in 2001 by Boneh and Franklin [4], which was provably secure against adaptive chosen ciphertext attack in random oracle model. Then, many other kinds of identity-based encryption were proposed [5-9]. Identity-based cryptography significantly reduces the system complexity and the cost for establishing and managing the public key authentication framework known as PKI (Public Key Infrastructure). As a result, we focus on the construction of identity-based threshold broadcast encryption (IBTHBE) in this paper. To the best of our knowledge, very few works have dealt with this problem. In [10], Chai and Cao *et al* propose a scheme based on identity. But the length of the ciphertexts is $n+1$ and the security relies on the random oracles. Vanesa Daza *et al* propose another scheme [11]. However, its security is still relying on the random oracles. The recent work [12] has short ciphertexts, but the security of their scheme based on the identity (IBTBE) is also relying on the random oracles. In [13], authors also proposed an efficient scheme in the standard model. But this scheme only achieves a weak security -selective-identity security.

As a natural extension of the efforts to improve schemes in the standard model, we propose two new efficient identity-based threshold broadcast encryption schemes in this paper. The proposed schemes are constructed in the standard model. In our schemes, the broadcaster can choose the privileged set and the threshold value at the time of encryption. In addition, under the full security model, the security of the first scheme is reduced to the n+1-Weak Decision Bilinear Diffie-Hellman Exponent (n+1-wDBDHE) assumption and the security of the second scheme is reduced to the static assumptions.

II. PRELIMINARIES

doi:10.4304/jnw.6.11.1602-1609

In this section, some definitions are given as follows:

A. Bilinear Groups

We briefly review bilinear maps and use the following notations: Let G and G_1 be two (multiplicative) cyclic groups of prime order *p*. A bilinear map is a map $e: G \times G \to G_1$ with the properties:

1. Bilinearity: for all $u, v \in G$, $a, b \in Z_p$, we have $e(u^a, v^b) = e(u, v)^{ab}$.

2. Non-degeneracy: $e(g, g) \neq 1$.

3. Computability**:** There is an efficient algorithm to compute $e(u, v)$ for all $u, v \in G$.

B. Decisional bilinear Diffie-Hellman Exponent assumption (BDHE)

The decisional bilinear Diffie-Hellman Exponent (BDHE) problem is defined as follows. Algorithm *B* is given as input a random tuple

$$(g, h_0, y_1, \cdots, y_n, y_{n+2}, \cdots, y_{2n+2}, T),$$

where $y_i = g^{\alpha^i}$. Algorithm *B*'s goal is to output 1 when $T = e(g, h_0)^{\alpha^{n+1}}$ and 0 otherwise. Let $TU = (g, h_0, y_1, \cdots, y_n, y_{n+2}, \cdots, y_{2n+2})$. Algorithm *B* that outputs $b \in \{0,1\}$ has advantage ε in solving decision BDHE in *G* if

$$|Pr[B(TU, e(g,h_0)^{\alpha^{n+1}}) = 0] - Pr[B(TU, T) = 0]| \leq \varepsilon.$$

Definition 1 The (t, ε) decisional BDHE assumption holds if no *t*-time algorithm has a non-negligible advantage ε in solving the above game.

C Identity-based Threshold Broadcast Encryption (IBTBE)

More formally, a threshold broadcast encryption scheme consists of five algorithms.

Setup The randomized *Setup* algorithm takes as input a security parameter *k* and outputs some public parameters *params*, which will be common to all the users of the system.

Extract The key generation algorithm is run by each user ID_i. It takes as input some public parameters params and returns a correspondence private key d_{ID_i}.

Threshold Encryption The encryption algorithm takes as input a set of public keys corresponding to a set *P* of *n* receivers, a threshold *t* satisfying $1 \leq t \leq n$, and a message *M*. The output is a ciphertext *C*, which contains the description of *P* and *t*.

Partial Decryption Partial Decryption algorithm takes as input a ciphertext *C* for the pair (*P*, *t*) and a secret key d_{ID_i} of a receiver. The output is a partial decryption value k_i or a special symbol $\perp$.

Decryption The deterministic final decryption algorithm takes as input a ciphertext *C* for the pair (*P*, *t*) and *t* partial decryptions corresponding k_i to receivers in some subset $S \subset P$. The output is a message *m* or a special symbol $\perp$.

D Security Model

Concerning the security of the identity-based cryptography, there are mainly two definitions:

- Full security, which means that the attacker can choose adaptively the identity he wants to attack (after having seen the parameters);
- Selective-ID security, which means that the attacker must choose the identity he wants to attack at the beginning, before seeing the parameters. The Selective-ID security is thus weaker than full security.

To define the notion of chosen ciphertext secure identity-based broadcast threshold decryption scheme (IND-fullID-CCA) in the full security model, let us consider the following game between an adversary *A* and a challenger:

Setup The challenger runs Setup. Then challenger gives the resulting common parameter to *A*, and keeps master key secret. *A* issues the threshold parameters (*n*, *t*).

Phase 1 *A* issues private key extraction and decryption queries adaptively. The adversary *A* adaptively issues queries $q_1, \cdots, q_{s0}$, where q_i is one of the following:

- On a private key extraction query upon ID_i, the challenger runs *Extract* to generate the private key associated to ID_i, then sends it to *A*.
- On a decryption queries, the challenger runs *Decryption* to generate decryption shares and gives them to *A*.

Challenge When *A* decides that phase 1 is over, it submits a set of identities S^*, a threshold value *t* and two messages (M_0, M_1)on which it wants to be challenged. The adversary's choice of S^* is restricted to the identities that he did not request a private key for in Phase 1. The challenger runs *Encrypt* algorithm to obtain (Hdr^*, *K*) =*Encrypt*(S^*, *PK*, *t*) and returns them to *A*. Note, *A* may already learned about the private keys of at most $t-1$. There is the natural constraint that *S** contains at most $t-1$ corrupted identities.

Phase 2 The adversary continues to issue queries $q_{s0+1}, \cdots, q$, where q_i is one of the following:

- *Extraction query* (ID_i), as in phase 1;
- *Decryption query*, as in phase 1, but with the constraint that $Hdr \neq Hdr^*$. The challenger responds as in phase 1.

Guess Finally, the adversary *A* outputs a guess $b' \in \{0,1\}$ and wins the game if $b = b'$.

We say that if the above indistinguishability game allow no decryption oracle query, then the IBTBE scheme is only chosen plaintext(IND-fullID-CPA) secure. There have been many methods to convert an IND-fullID-CPA scheme to an IND-fullID-CCA scheme. Therefore, we only focus on constructing the IND-fullID-CPA scheme in this paper.

III NEW CONSTRUCTIONS (I)

A. Our Construction

Let $S = \{ID_1, \cdots, ID_n\}$ be n users, where $ID_i = \{v_{i1}, \cdots, v_{il}\}$ is an n-bit string and v_{ij} is an $\frac{n}{l}$-bit string. Our construction works as follows:

Setup To generate the system parameters, the PKG picks randomly generators g, g_2 in G and an element α from Z_p. Note that any user ID_i will be associated to a different element t_i. This can be done by defining $t_i = f(ID_i)$ for some n-1 degree polynomial function $f(x)$, where $f(0) = \alpha$. PKG sets $T_i = g^{t_i}$ for $1 \le i \le n$ and $g_1 = g^{\alpha}$. Then it chooses randomly h-length vector $\mathbf{g}_3 = (g_{31}, \cdots, g_{3n})$ and vectors $(\mathbf{U}_1, \cdots, \mathbf{U}_n)$ in G, where $\mathbf{U}_i = (U_{i1}, \cdots, U_{il})$. The public parameters PK are

$$PK = (g, g_1, g_2, T_1, \cdots, T_n, \mathbf{g}_3, \mathbf{U}_1, \cdots, \mathbf{U}_n)$$

and α is master key.

Extract(ID_i) To generate a private key for a user $ID_i \in Z_p$, the PKG first defines $F_j(x) = g_{3j} \prod_{i=1}^{l} u_{ji}^{x_i}$ for $1 \le j \le n$ and $x = (x_1, \cdots, x_l)$. Then it picks random $r_i \in Z_p$, and outputs the private key:

$$d_{ID_i} = (d_{i1}, d_{i2}) = (g_2^{t_i} (\prod_{i=1}^{n} F_j)^{r_i}, g^{r_i}),$$

where $F_i = F(ID_i) = g_{3i} \prod_{j=1}^{l} u_{ij}^{v_{ij}}$.

Threshold Encryption To encrypt a message M for a set $S = \{ID_1, \cdots, ID_n\}$ of n players, with threshold $t \le n$ for the decryption, the idea is to set up an (n, N)-threshold secret sharing scheme, where $N = 2n - t$. The n public keys $(T_1, \cdots, T_n)$ of users implicitly define a $n-1$ degree polynomial. The idea is to compute the values of this polynomial in the points $x = 0$ (This will lead to obtain the value of g_1).Then a sender acts as follows:

- Select a random element $s \in Z_p^*$ and compute

$$C_1 = g^s, C_2 = e(g_1, g_2)^s M, C_3 = (\prod_{i=1}^{n} F_j)^s.$$

- Choose a set $\overline{S}$ of $n-t$ dummy players, such that $\overline{S} \cap S = \phi$. For each user $ID_i' \in \overline{S}$, compute $T_i' = \prod_{ID_i \in S} T_i^{\lambda_{ij}}$ and $K_i = \frac{1}{e(T_i', g_2^s)}$, where λ_{ij} denotes the Lagrange coefficients.
- The ciphertexts are $(C_1, C_2, C_3, \{K_i\}_{ID_i' \in \overline{S}})$.

Note: $K_i = \frac{1}{e(T_i', g_2^s)} = \frac{1}{e(g^{t_i'}, g_2^s)}$ by using Lagrange interpolation where $t_i' = f(ID_i')$.

Partial Decryption Given the ciphertexts $(C_1, C_2, C_3, \{K_i\}_{ID_i' \in \overline{S}})$, the receiver $ID_i \in S$ with his corresponding private d_{ID_i} computes as follows:

$$K_i = \frac{e(C_3, d_{i1})}{e(d_{i0}, C_1)} = \frac{1}{e(g^{t_i}, g_2)^s}.$$

Decryption Given the valid ciphertexts $(C_1, C_2, C_3, \{K_i\}_{ID_i' \in \overline{S}})$, a subset $S_1 \subset S$ with $|S_1| = t$ and corresponding t partial decryption K_j, the algorithm computes with the whole set $S' = S_1 \bigcup \overline{S}$ as follows: $K = \prod_{ID_i \in S'} K_i^{\lambda_{i0}} = \frac{1}{e(g_1, g_2)^s}$, $M = K \cdot C_2$.

Efficiency In our scheme, the size of ciphertexts is $O(n-t)$ and the size of private key is constant as it consists of two group elements. This is the first efficient construction which has full security in the standard model for the identity-based threshold broadcast encryption. In addition, if the values $e(g_1, g_2)$ and $e(T_i, g_2)$ can be precomputed and cached, so no pairing computations are needed at the phase of *Threshold Encryption*. Table 1 gives the efficiency comparison between ours and the others IBTBEs.

Note: R.O. denotes the random oracles. C-Size is the size of ciphertext and pk is the private keys. SM denotes the security model. Full and s-ID are full security and selective-identity model.

B Security Analysis

Theorem 1 Suppose the n+1-wDBDHE assumption holds. Then the proposed scheme above is semantically secure against selective identity, chosen plaintext attacks(IND-fullID-CPA).

Proof Suppose an adversary A has advantage ε in attacking our scheme. Using A, we build an algorithm B that solves the decision n+1-wDBDHE problem in G with the advantage ε. For a generator $g \in G$ and $\alpha \in Z_p$, set $Y_i = g^{\alpha^i}$. Algorithm B is given as input a random tuple $(g, g_0, T_1, \cdots, T_n, T)$ where $g_0 \in G$. Algorithm B's goal is to output 1 when $e(g, g_0)^{\alpha^{n+1}}$ and 0 otherwise. Algorithm B works by interacting with A in a threshold full security game as follows:

Initial A outputs a set $\tilde{S}$ of identities that he wants to corrupt, where $|\tilde{S}| \le n-1$.

Setup B sets $m = 2\max\{2q, 2^{\frac{n}{T}}\} = 4q$ where q is the maximum query time for private query.

- First, B selects $n-1$ random integers $\alpha_1, \alpha_2, \cdots, \alpha_{n-1} \in Z_p$. Let $f(x)$ be the degree $n-1$ polynomial implicitly defined to satisfy $f(0) = \alpha$ and $f(ID_i) = \alpha_i$ for $ID_i \in \tilde{S}$, note that B does not know f since it does not know α. For $ID_i \in \tilde{S}$, B computes

TABLE I. THE COMPARISON OF THE EFFICIENCY WITH THE OTHERS IDTHBE

Schemes	Assumption	C-Size	pk Size	Parings	WithoutR.O.	S.M
[10]	DBDH	n +1	1	1 +2 t	NO	Full
[11]	DBDH	$n-t$	1	3 t +2 t	NO	Full
[12]	DMBDH	n	2	(0+ t)+1	NO	Full
[13]	n+1-DBDHE	$O(n-t)$	2	2	YES	s-ID
Ours	n+1-wDBDHE	$O(n-t)$	2	0+ t	YES	Full

$T_i = g^{\alpha_i}$. Otherwise, B computes $\alpha_i = f(ID_i) = \lambda_0\alpha + \sum_{j=1}^{n-1}\lambda_j\alpha_j$ with the Lagrange coefficients λ_j . Note that these Lagrange coefficients are easily calculated since they do not depend on f . Then B sets $T_i = g_1^{\lambda_0}\prod_{ID_j\in\tilde{S}} T_j^{\lambda_j}$.

• Next, B chooses random l-length vectors $\mathbf{X_1},\cdots,\mathbf{X_n}\in Z_m$, where $\mathbf{X}_i = \{x_{i1},\cdots,x_{il}\}$ for $1\le i\le n$. Furthermore, B chooses random $u_1,\cdots,u_n$ from Z_m and $z_1,\cdots,z_n$ from Z_p , l -bit vectors $\mathbf{Y_1},\cdots,\mathbf{Y_n}$ from Z_p where $\mathbf{Y}_i = \{y_{i1},\cdots,y_{il}\}$ for $1\le i\le n$. Then it selects k_j for $1\le j\le n$ randomly from $\{0,1,\cdots,\mu_l\}$, where $\mu_l = l(2^{\frac{n}{T}}+1)$. Some functions are defined as follows:

$$F_j(ID_j) = p + mk_j - u_j - \sum_{i=1}^{l} v_{ji}x_{ji},\ J_j(ID_j) = z_j + \sum_{i=1}^{l} v_{ji}y_{ji},$$

$$K_j(ID_j) = \begin{cases} 0 & if\ u_j + \sum_{i=1}^{l} v_{ji}x_{ji} = 0 \bmod m \\ 1 & otherwise \end{cases},\ 1\le j\le l .$$

• Then B constructs a set of public parameters for the scheme by making the following assignments. B takes as input a tuple $TU = (g, g_0, T_1,\cdots,T_n, T)$ where g, g_0 are random generators of G and $Y_i = g^{\alpha^i}$ for some random $\alpha\in Z_p$. Then B chooses a random $b\in Z_p$ and assigns: $g_1 = Y_1 = g^{\alpha}$, $g_2 = g^{\alpha^n}g^b$, $g_{3j} = Y_{n-j+1}^{p+mk_j-u_j}g^{z_j}$, $u_{ji} = Y_{n-j+1}^{-x_{ji}}g^{y_{ji}}$, $1\le j\le n, 1\le i\le l$. It provides A the

$PK = (g, g_1, g_2, T_1,\cdots,T_n, \mathbf{g}_3, \mathbf{U}_1,\cdots,\mathbf{U}_n)$.

Furthermore, this assignment means that the master secret will be $g_2^{\alpha} = (g^{\alpha^n}g^b)^{\alpha}$ which is unknown to B. Using the definition of the public parameters, it shows that $F_j = g_{3j}\prod_{i=1}^{l} u_{ji}^{v_{ji}} = Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)}$.

Phase 1 A issues private key extraction and decryption queries adaptively. The adversary A adaptively issues queries $q_1,\cdots,q_{s0}$, where q_j is one of the followings: Suppose the adversary A issues a query for an identity $ID_j = \{v_{j1},\cdots v_{jl}\}$. B checks whether $K_j(ID_j)\neq 0$. It aborts if there is no such j.

Otherwise, it answers the query as follows:(by using a similar procedure as in case of the scheme [6]) B first computes the Lagrange coefficients $\lambda_0,\lambda_1,\cdots,\lambda_{n-1}$ such that $t_j = f(ID_j) = \lambda_0\alpha + \sum_{j=1}^{n-1}\lambda_j\alpha_j$. Let $S = \{ID_j\}\bigcup\tilde{S} = \{ID_1',\cdots ID_n'\}$ where $ID_j' = ID_j$. B selects a random $r'\in Z_p$ and generates the private key:

$$d_{ID_j} = (d_{j0}, d_{j1}) = (g_2^{t_j}(\prod_{i=1}^{n} F_i)^r, g^r),$$

where $r = r' - \frac{\lambda_0\alpha^j}{F_j(ID_j)}$. One can obtain that d_{ID_j} is a properly simulated private key for the identity ID_j . In fact,

$$d_{j0} = g_2^{t_j}(\prod_{i=1}^{n} F_i)^r = (g^{\alpha^n}g^b)^{\sum_{j=0}^{n}\lambda_j\alpha_j}(\prod_{i=1}^{n} F_i)^r$$
$$= Y_{n+1}^{\lambda_0}Y_n^{\sum_{j=1}^{n-1}\lambda_j\alpha_j}g^{b\sum_{j=1}^{n-1}\lambda_j\alpha_j}Y_1^{b\lambda_0}F_j^r(\prod_{i=1,i\neq j}^{n} F_i)^r, \quad (1)$$

Where

$$F_j^r = (Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^r = (Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^{r'-\frac{\lambda_0\alpha^j}{F_j(ID_j)}}$$
$$= (Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^{r'}(Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^{-\frac{\lambda_0\alpha^j}{F_j(ID_j)}}$$
$$= Y_{n+1}^{-\lambda_0}Y_j^{-\frac{\lambda_0 J_j(ID_j)}{F_j(ID_j)}}(Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^{r'},$$

$$F_i^r = (Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^r = (Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^{r'-\frac{\lambda_0\alpha^j}{F_j(ID_j)}}$$
$$= (Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^{r'}(Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^{-\frac{\lambda_0\alpha^j}{F_j(ID_j)}}$$
$$= (Y_{n+j-i+1}^{-\frac{\lambda_0 F_i(ID_i')}{F_j(ID_j)}}Y_j^{-\frac{\lambda_0 J_i(ID_i')}{F_j(ID_j)}})(Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^{r'}$$
$$= ((Y_nY_{j-i+1})^{-\frac{\lambda_0 F_i(ID_i')}{F_j(ID_j)}}Y_j^{-\frac{\lambda_0 J_i(ID_i')}{F_j(ID_j)}})(Y_{n-i+1}^{F_i(ID_i')}g^{J_i(ID_i')})^{r'}.$$

By using (1), one can obtain the followings.

$$d_{j0} = Y_{n+1}^{\lambda_0}Y_n^{\sum_{j=1}^{n-1}\lambda_j\alpha_j}g^{b\sum_{j=1}^{n-1}\lambda_j\alpha_j}Y_1^{b\lambda_0}F_j^r(\prod_{i=1,i\neq j}^{n} F_i)^r$$
$$= Y_{n+1}^{\lambda_0}Y_n^{\sum_{j=1}^{n-1}\lambda_j\alpha_j}g^{b\sum_{j=1}^{n-1}\lambda_j\alpha_j}Y_1^{b\lambda_0}Y_{n+1}^{-\lambda_0}Y_j^{-\frac{\lambda_0 J_j(ID_j)}{F_j(ID_j)}}\bullet$$
$$(Y_{n-j+1}^{F_j(ID_j)}g^{J_j(ID_j)})^{r'}(\prod_{i=1,i\neq j}^{n} F_i)^r$$

$$= Y_{n+1}^{\lambda_0} Y_n^{\sum_{j=1}^{n-1}\lambda_j\alpha_j} g^{b\sum_{j=1}^{n-1}\lambda_j\alpha_j} Y_1^{b\lambda_0} Y_{n+1}^{-\lambda_0} Y_j^{\frac{\lambda_0 J_j(ID_j)}{F_j(ID_j)}} (Y_{n-j+1}^{F_j(ID_j)} g^{J_j(ID_j)})^{r'} \bullet$$

$$\prod_{i=1,i\neq j}^{n} ((Y_n Y_{j-i+1})^{-\frac{\lambda_0 F_i(ID_i')}{F_j(ID_j)}} Y_j^{-\frac{\lambda_0 J_i(ID_i')}{F_j(ID_j)}} (Y_{n-i+1}^{F_i(ID_i')} g^{J_i(ID_i')})^{r'}$$

$$= Y_n^{\sum_{j=1}^{n-1}\lambda_j\alpha_j} g^{b\sum_{j=1}^{n-1}\lambda_j\alpha_j} Y_1^{b\lambda_0} Y_j^{-\frac{\lambda_0 J_j(ID_j)}{F_j(ID_j)}} (Y_{n-j+1}^{F_j(ID_j)} g^{J_j(ID_j)})^{r'} \bullet$$

$$\prod_{i=1,i\neq j}^{n} ((Y_n Y_{j-i+1})^{-\frac{\lambda_0 F_i(ID_i')}{F_j(ID_j)}} Y_j^{-\frac{\lambda_0 J_i(ID_i')}{F_j(ID_j)}} (Y_{n-i+1}^{F_i(ID_i')} g^{J_i(ID_i')})^{r'}.$$

Since Y_{n+1} cancels out, all the terms in this expression are known to *B*. Thus, *B* can compute the first private key component d_{j0}. *B* computes $Y_j^{-\frac{\lambda_0}{F_j(ID_j)}} g^{r'} = g^r$. Then the second private key component is obtained. Finally, $d_{ID_j} = (d_{j0}, d_{j1})$ is given to *A*.

Challenge *A* outputs two same-length messages M_0, M_1, a threshold value t and a set of identities $S^* = (ID_1^*, ID_2^*, \cdots, ID_n^*)$ on which it wishes to be challenged. Note that $|S^* \bigcap \tilde{S}| \leq t-1$. *B* first checks whether there exists a $j \in \{1,\cdots,n\}$ such that $F_j^*(ID_j^*) \neq 0$, then *B* will abort. Otherwise, *B* picks a random $\gamma \in \{0,1\}$ and constructs the challenge ciphertexts as follows: $C^* = (C_1^*, C_2^*, C_3^*, \{K_i\})$

$$= (g_0, M_\gamma e(g_1, g_0^b) T, g_0^{\sum_{i=1}^{n} J_i(ID_i^*)}, \{K_i\}_{ID_i \in S_0}).$$

where S_0 is a set of *n* - *t* dummy users. In addition, K_i is computed in the following manner: *B* first chooses a set S_0 of $n-t$ dummy users such that $S_0 \bigcap S^* = \phi$. For each dummy user $ID_i \in S_0$, *B* computes the Lagrange coefficients λ_{ji} with $1 \leq j \leq n$ such that $t_j' = f(ID_j) = \sum_{ID_i^* \in S^*} \lambda_{ji} \alpha_i'$, where α_i' is known to *B* since *B* can compute it by using $(\alpha_1, \alpha_2, \cdots, \alpha_{n-1})$ and satisfies $g^{\alpha_i'} = T_i$. Then *B* computes $T_j' = \prod_{ID_i^* \in S^*} T_i^{\lambda_{ji}}$. Finally, *B* computes $K_i = 1 / e(g_0^{\sum_{i=1}^{n}\lambda_{ji}\alpha_i'}, g_2)$.

Let $g_0 = g^\mu$ for some unknown $\mu \in Z_p$. If $T = e(g, g_0)^{\alpha^{n+1}}$, one can obtain that C^* is a valid encryption for M_γ. In fact, $C_1^* = g^\mu$, $C_2^* = M_\gamma e(g_1, g_0^b) T$

$$= M_\gamma e(g_1, g_0^b) e(g, g_0)^{\alpha^{n+1}} = M_\gamma (e(g_1, g^b) e(g^{\alpha^{n+1}}, g))^\mu$$

$$= M_\gamma (e(g_1, g^b Y_n))^\mu = M_\gamma e(g_1, g_2)^\mu,$$

$$C_3^* = g_0^{\sum_{i=1}^{n} J_i(ID_i^*)} = \prod_{i=1}^{n} g_0^{J_i(ID_i^*)} = (\prod_{i=1}^{n} g^{J_i(ID_i^*)})^\mu$$

$$= (\prod_{i=1}^{n} Y_{n-i+1}^{F_i(ID_i^*)} g^{J_i(ID_i^*)})^\mu = (\prod_{i=1}^{n} F_i)^\mu.$$

and $$K_i = \frac{1}{e(g_0^{\sum_{i=1}^{n}\lambda_{ji}\alpha_i'}, g_2)} = \frac{1}{e(g^{\sum_{i=1}^{n}\lambda_{ji}\alpha_i'}, g_2)^\mu}$$

$$= \frac{1}{e(\prod_{i=1}^{n} T_i^{\lambda_{ji}}, g_2)^\mu} = \frac{1}{e(T_i', g_2)^\mu} = \frac{1}{e(T_i', g_2^\mu)}.$$

If T is a random element of G_1, C^* gives no information about *B*'s choice of γ.

Phase 2 The adversary continues to issue queries and *B* responds as in phase 1.

Guess *A* outputs a guess $\gamma' \in \{0,1\}$ and wins the game if $\gamma' = \gamma$. If $\gamma' = \gamma$, *B* will output 1 to indicate that *B* solves the *n*+1-wDBDHE problem, otherwise it outputs 0 to mean that it learns nothing from C^*. When *A* outputs 1, it means $|\Pr(\gamma = \gamma') - \frac{1}{2}| \geq \varepsilon$. Otherwise $\Pr(\gamma = \gamma') = \frac{1}{2}$. Therefore, we have

$$|\Pr(B(TU, e(g, g_0)^{\alpha^{n+1}}) = 0) - \Pr(B(TU, T) = 0)|$$

$$\geq |\frac{1}{2} \pm \varepsilon - \frac{1}{2}| = \varepsilon.$$

IV NEW CONSTRUCTIONS (II)

The first construction achieves full security in the standard model. But the size of public keys is too large and the computation cost of the private key is expensive. In addition, the hardness assumption in our scheme is strong.

As a natural extension to improve the first scheme, we propose anther scheme in this section. It is based on the dual encryption technique[14-17]. In addition, the security of the proposed scheme is reduced to three static (i.e. non *q*-based) assumptions.

A. Dual encryption technique

Recently, a new technique is applied to IBE. It is called Dual Encryption Technique. In a dual system[14,15], ciphertexts and keys can take on two forms: normal or semi-functional. Semi-functional ciphertexts and keys are not used in the real system, they are only used in the security proof. A normal key can decrypt normal or semi-functional ciphertexts, and a normal ciphertext can be decrypted by normal or semi-functional keys. However, when a semi-functional key is used to decrypt a semi-functional ciphertext, decryption will fail. More specifically, the semi-functional components of the key and ciphertext will interact to mask the blinding factor by an additional random term. Security for dual systems is proved using a sequence of games which are shown to be indistinguishable. The first game is the real security game (with normal ciphertext and keys). In the next game, the ciphertext is semi-functional, while all the keys are normal. For an attacker that makes *q* key requests, games 1 through *q* follow. In game *k*, the first *k* keys are semi-functional while the remaining keys are normal. In game *q*, all the keys and the challenge ciphertext given to the attacker are semi-functional. Hence none of the given keys are useful for

decrypting the challenge ciphertext. At this point, proving security becomes relatively easy. Waters[14] first proposed a broadcast encryption scheme based on this new technique. However, the proposed scheme is not based on identity and also inefficient since its cost of decryption is dependent on depth of users set.

B. Composite Order Bilinear Groups

Composite order bilinear groups were used in [14-16]. In this paper, the outputs is ($N= p_1p_2p_3$, G, G_1, e), where p_1, p_2, p_3 are distinct primes, G and G_1 are cyclic groups of order N. A bilinear map e is a map $e: G \times G \to G_1$ with the following properties:

(i) Bilinearity: for all $u, v \in G$, $a, b \in Z_N$, we have $e(u^a, v^b) = e(u, v)^{ab}$;

(ii) Non-degeneracy: $\exists\ g \in G$ such that $e(g, g)$ has order N in G_1.

(iii) Computability: there is an efficient algorithm to compute $e(u, v)$ for all $u, v \in G$.

C. Static Hardness Assumption

In this section, we give our complex assumption. These assumptions have been used in [14,15].

Assumption 1(Subgroup decision problem for 3 primes) Given ($N= p_1p_2p_3$, G, G_1, e), select randomly $g \in G_{p_1}$, $X_3 \in G_{p_3}$, $T_1 \in G_{p_1p_2}$, $T_2 \in G_{p_1}$ and set D=($N= p_1p_2p_3$, G, G_1, e, g, X_3). It is hard to distinguish T_1 from T_2. The advantage of an algorithm is defined as

$$Adv_1 = |\Pr[A(D, T_1) = 1] - \Pr[A(D, T_2) = 1]|.$$

Definition 2 Assumption 1 holds if Adv_1 is negligible.

Assumption 2 Given ($N=p_1p_2p_3$, G, G_1, e), pick randomly $g, X_1 \in G_{p_1}$, $X_2, Y_2 \in G_{p_2}$, $X_3, Y_3 \in G_{p_3}$, set D=($N= p_1p_2p_3$, G, G_1, e, g, X_1X_2, X_3, Y_2Y_3). Then select $T_1 \in G$, $T_2 \in G_{p_1p_3}$ at random. It is hard to distinguish T_1 from T_2. The advantage of an algorithm is defined as $Adv_2 = |\Pr[A(D, T_1) = 1] - \Pr[A(D, T_2) = 1]|$.

Definition 3 Assumption 2 holds if Adv_2 is negligible.

Assumption 3 Given ($N=p_1p_2p_3$, G, G_1, e), pick randomly $g \in G_{p_1}$, $X_2, Y_2, Z_2 \in G_{p_2}$, $X_3 \in G_{p_3}$, $\alpha, s \in Z_N$, set D=($N=p_1p_2p_3$, G, G_1, e, g, $g^\alpha X_2, X_3, g^s Y_2, Z_2$). Then compute $T_1 = e(g, g)^{\alpha s}$ and pick randomly $T_2 \in G_1$. It is hard to distinguish T_1 from T_2. The advantage of an algorithm is defined as

$$Adv_3 = |\Pr[A(D, T_1) = 1] - \Pr[A(D, T_2) = 1]|.$$

Definition 4 Assumption 3 holds if Adv_3 is negligible.

C. Construction

We give an initial construction at first. It works as follows:

Let S = { $ID_1, \cdots, ID_n$ } be n players where $ID_i \in Z_p$. These users want to form an ad hoc network. Our construction works as follows:

Setup: To generate the system parameters, the PKG picks randomly generators { $g, g_2, h, h_1, \cdots, h_n$ } in G and an element α from Z_p. Note that any user ID_i will be associated to a different element t_i. This can be done by defining $t_i = f(ID_i)$ for some n-1 degree polynomial function $f(x)$, where $f(0) = \alpha$. PKG sets $T_i = g^{t_i}$ for $1 \le i \le n$ and $g_1 = g^\alpha$. The public parameters PK are

$$PK = (g, g_1, g_2, T_1, \cdots, T_n, h, h_1, \cdots, h_n)$$

and α is master key.

Extract(ID_i) : To generate a private key for a user $ID_i \in Z_p$, the PKG picks randomly $r_i \in Z_p$ and also chooses random elements $R_{i0}, R'_{i0}, R_{i1}, \cdots, R_{i(i-1)}, R_{i(i+1)}, \cdots, R_{in} \in G_{p_3}$. Then it computes private keys as follows: $d_{ID_i} = (d_{i0}, d', d_{i1}, \cdots, d_{i,i-1}, d_{i,i+1}, \cdots, d_{in})$

$$= (g_2^\alpha (hu_i^{ID_i})^{r_i} R_{i0}, g^{r_i} R'_{i0}, u_1^{r_i} R_{i1}, \cdots, u_{i-1}^{r_i} R_{i(i-1)}, u_{i+1}^{r_i} R_{i(i+1)}, \cdots, u_n^{r_i} R_{in}).$$

Threshold Encryption: To encrypt a message M for a set S = {$ID_1, \cdots, ID_n$} of n players, with threshold $t \le n$ for the decryption, the idea is to set up an (n, N)-threshold secret sharing scheme, where $N = 2n - t$. The n public keys ($T_1, \cdots, T_n$) of users implicitly define a $n-1$ degree polynomial. The idea is to compute the values of this polynomial in the points $x = 0$ (This will lead to obtain the value of g_1).Then a sender acts as follows:

- Select a random element $s \in Z_p^*$ and compute $C_1 = g^s$, $C_2 = e(g_1, g_2)^s M$ and $C_3 = (\prod_{i=1}^{n} h_i^{ID_i} h)^s$.
- Choose a set $\overline{S}$ of $n - t$ dummy players, such that $\overline{S} \cap S = \phi$. For each user $ID'_i \in \overline{S}$, compute $T'_i = \prod_{ID_i \in S} T_i^{\lambda_{ij}}$ and $K_i = \frac{1}{e(T'_i, g_2^s)}$, where λ_{ij} denotes the Lagrange coefficients.
- The ciphertexts are $(C_1, C_2, C_3, \{K_i\}_{ID'_i \in \overline{S}})$.

Note: $K_i = \frac{1}{e(T'_i, g_2^s)} = \frac{1}{e(g^{t'_i}, g_2^s)}$ by using Lagrange interpolation where $t'_i = f(ID'_i)$.

Partial Decryption: Given the ciphertexts $(C_1, C_2, C_3, \{K_i\}_{ID'_i \in \overline{S}})$, the receiver $ID_i \in S$ with his corresponding private d_{ID_i} computes as follows:

$$K_i = \frac{e(C_3, d'_{i0})}{e(d_{i0}\prod_{j=1,j\neq i}^{n} d_{ij}^{ID_j}, C_1)} = \frac{1}{e(g^{t_i}, g_2)^s}.$$

Decryption: Given the valid ciphertexts $(C_1, C_2, C_3, \{K_i\}_{ID'_i \in \bar{S}})$, a subset $S_1 \subset S$ with $|S_1| = t$ and corresponding t partial decryption K_j, the algorithm computes with the whole set $S' = S_1 \bigcup \bar{S}$ as follows: $K = \prod_{ID_i \in S'} K_i^{\lambda_{i0}} = \frac{1}{e(g_1, g_2)^s}$ and $M = K \cdot C_2$.

Correctness:

In fact, if the ciphertexts $C = (C_0, C_1, C_2)$ is valid, then one can obtain the following equation holds.

$$\frac{e(C_1, d')}{e(d_{i0}\prod_{j=1,j\neq i}^{n} d_{ij}^{ID_j}, C_2)}$$

$$= \frac{e((h\prod_{i=1}^{n} u_i^{ID_i})^s, g^{r_i} R'_{i0})}{e(g_2^{t_i}(h\prod_{i=1}^{n} u_i^{ID_i})^{r_i} R_{i0}(\prod_{j=1,j\neq i}^{n} R_{ij}^{ID_j}), g^s)}$$

$$= \frac{e((h\prod_{i=1}^{s} u_i^{ID_i})^s, g^{r_i}) e((h\prod_{i=1}^{n} u_i^{ID_i})^s, R'_{i0})}{e(g_2^{t_i}, g^s) e((h\prod_{i=1}^{n} u_i^{ID_i})^{r_i}, g^s) e(R_{i0}, g^s) e((\prod_{j=1,j\neq i}^{n} R_{ij}^{ID_j}), g^n)}$$

$$= \frac{1}{e(g_2^{t_i}, g)^s}.$$

Note: In the previous equation, the orthogonality property of $G_{p_1}, G_{p_2}, G_{p_3}$ is used. It is described simply as follows.

Lemma[14] When $h_i \in G_{p_i}, h_j \in G_{p_j}$ for $i \neq j$, $e(h_i, h_j)$ is the identity element in G_1.

By using this lemma, one can obtain

$$e((h\prod_{i=1}^{n} u_i^{ID_i})^s, R'_{i0}) = e(R_{i0}, g^s) = e((\prod_{j=1,j\neq i}^{n} R_{ij}^{ID_j}), g^s) = 1.$$

Efficiency analysis:

Our construction achieves $O(1)$-size ciphertexts. The private key of construction private key is linear in the maximal size of S. In addition, $e(g_1, g_2)$ and $e(T_i, g_2)$ can be precomputed, so there is no pair computations at the phase of *Encryption*. Furthermore, the security of the proposed scheme is reduced to the static assumptions. These assumptions are more natural than those in the existing schemes. However, the size of private keys relies on the number of set S. Based on the proposed scheme, we can give the main construction (II).

Setup, *Encryption* and *Decryption* are similar to the first scheme.

Extract(ID_i) : To generate a private key for a user $ID_i \in Z_p$, the PKG picks randomly $r_i \in Z_p$ and also chooses random elements $R_{i0}, R'_{i0}, R_{i1}, \cdots, R_{i(i-1)}$, $R_{i(i+1)}, \cdots, R_{in} \in G_{p_3}$. Then it computes private keys as follows: $d_{ID_i} = (d_{i0}, d_{i1}) = (g_2^{t_i}(\prod_{j=1,j\neq i}^{n} (hu_i^{ID_i})^{r_i} R_{j0}, g^{r_i} R'_{i0})$.

Partial Decryption: Given the ciphertexts $(C_1, C_2, C_3, \{K_i\}_{ID'_i \in \bar{S}})$, the receiver $ID_i \in S$ with his corresponding private d_{ID_i} computes as follows:

$$K_i = \frac{e(C_3, d_{i1})}{e(d_{i0}, C_1)} = \frac{1}{e(g^{t_i}, g_2)^s}.$$

Correctness can be easily obtained. I

Table 2 give the comparisons of efficiency with our two schemes.

TABLE II THE COMPARISON OF THE EFFICIENCY BETWEEN OUR TWO SCHEMES

Schemes	ASSUMPTION	C-Size	pk Size	Parings
Construction (I)	n+1-wDBDHE	$O(n-t)$	2	0+ t
Construction (II)	Static assumptions	O(1)	2	2

D. Security analysis

In this section, we will prove the security of the proposed scheme. We first define semi-functional keys and semi-functional ciphertexts. Let g_2 denote a generator of G_{p_2}.

Semi-functional keys: At first, a normal key $(\bar{d}_0, \bar{d}', \bar{d}_1, \cdots, \bar{d}_{i-1}, \bar{d}_{i+1}, \cdots, \bar{d}_n)$ is obtained using the *Extract* algorithm. Then some random elements $\gamma_0, \gamma'_0, \gamma_j$ for $j = 1, \cdots, n$ and $j \neq i$ are chosen in Z_N. The semi-functional keys are set as follows.

$d_0 = \bar{d}_0 g_2^{\gamma_0}, d' = \bar{d}' g_2^{\bar{\gamma}}, d_j = \bar{d}_j g_2^{\gamma_j}$, $j = 1, \cdots, n$, $j \neq i$.

Semi-functional ciphertexts: At first, a normal semi-functional ciphertext (C'_0, C'_1, C'_2) is obtained using the *Encrypt* algorithm. Then two random elements λ_1, λ_2 are chosen in Z_N. The semi-functional ciphertexts are set as follows. $C_0 = C'_0$, $C_1 = C'_1 g_2^{\lambda_1 \lambda_2}$, $C_2 = C'_2 g_2^{\lambda_2}$.

We organize our proof as a sequence of games. The first game defined will be the real identity-based encryption game and the last one will be one in which the adversary has no advantage unconditionally. We will show that each game is indistinguishable from the next (under three complexity assumptions). We first define the games as:

Game$_{real}$: This is a real IBTBE security game.

For $0 \leq i \leq q$, the Game$_i$ is defined as follows.

Game$_i$: Let Ω denote the set of private keys which the adversary queries during the games. This game is a real IBTBE security game with the two exceptions: (1) The challenge ciphertext will be a semi-functional ciphertext on the challenge set S^*. (2) The first i keys will be semi-functional private keys. The rest of keys in Ω will be normal.

Note: In game$_0$, the challenge ciphertext is semi-functional. In game$_q$, the challenge ciphertexts and all keys are semi-functional.

Game$_{final}$: This game is same with Game$_q$ except that the challenge ciphertext is a semi-functional encryption

of random group element of G_1.

It can be easily shown that these games are indistinguishable in a set of Lemmas. For the pages limited, we omit them here(It can be obtained from [16-18]). Then we have the following theorem.

Theorem 2 If Assumption 1,2 and 3 hold, then our IBTBE is IND-ID-CPA secure.

ACKNOWLEDGEMENT

This work is supported in part by the Nature Science Foundation of China under grant (60970119, 60803149), the National Basic Research Program of China(973) under grant 2007CB311201 and the Fundamental Research Funds for the Central Universities(Public key broadcast encryption based on new mathematical hardness assumptions).

REFERENCES

[1] A. Fiat, M. Naor. "Broadcast encryption". In: Proceedings of CRYPTO, Berlin: Springer-Verlag, LNCS 773, 1994, pp. 480-491.
[2] H. Ghodosi, J. Pieprzyk and R. Safavi-Naini. "Dynamic threshold cryptosystems: a new scheme in group oriented cryptography". In: Proceedings of Pragocrypt 96, CTU Publishing House, 1996 , pp. 370-379.
[3] A. Shamir, "Identity-based Cryptosystems and Signature Schemes", In: Proceedings of CRYPTO, Berlin: Springer-Verlag, LNCS 196, 1984, pp. 47–53.
[4] D. Boneh and M. Franklin. "Identity-based encryption from the well pairing". In: Proceedings of CRYPTO, Berlin: Springer-Verlag , LNCS 2193, 2001, pp. 213-229.
[5] D. Boneh and X. Boyen. "Efficient selective-id secure identity based encryption without random oracles". In: Proceedings of EuroCryp, Berlin: Springer-Verlag, LNCS 3027, 2004 , pp. 223-238,.
[6] C. Cocks. "An identity based encryption scheme based on quadratic residues". In: Proceedings of Cryptography and coding, Berlin: Springer-Verlag, LNCS 2260, 2001, pp. 360-363.
[7] D. Boneh and J. Katz. "Improved Efficiency for CCA-Secure Cryptosystems Built Using Identity-Based Encryption". In: Proceedings of CT-RSA, Berlin: Springer-Verlag, LNCS 3376, 2005, pp. 87-103.
[8] R. Canetti, S. Halevi, and J. Katz. "Chosen-ciphertext security from identity-based encryption". In: Proceedings of EuroCrypt, Berlin: Springer-Verlag, LNCS 3027, 2004, pp. 207-222.
[9] S. Chattterjee and P. Sarkar. "Generalization of the Selectve-ID Security Model for HIBE Protocols". In: Proceedings of PKC, Berlin: Springer-Verlag, LNCS 3958, 2006, pp. 241-256.
[10] Z. Chai, Z. Cao and Y. Zhou. "Efficient ID-based Broadcast Threshold Decryption in Ad Hoc Network". In: Proceedings of IMSCCS 06, IEEE Computer Society, Volume 2, 2006, pp. 148-154.
[11] V. Daza, J. Herranz and P. Morillo etal. "CCA2-Secure Threshold Broadcast Encryption with Shorter Ciphertexts". In: Proceedings of ProvSec 2007, Berlin: Springer-Verlag, LNCS 4784, 2007, pp. 35-50.
[12] C. Delerablee and D. Pointcheval. "Dynamic Threshold Public-Key Encryption". In: Proceedings of CRYPTO, Berlin: Springer-Verlag, LNCS 5157, 2008, pp. 317-334.
[13] L.Y. Zhang, Y.P. Hu and Qing Wu. "Identity-based threshold broadcast encryption in the standard model". *KSII Trans. on internet and information systems* .Vol. 4, NO. 3, June 2010，pp.400-410.
[14] B. Waters. "Dual system encryption: realizing fully secure ibe and hibe under simple assumptions". Proceeding of Advances in Cryptology-Crypto, LNCS 5677, Berlin: Springer-Verlag press, 2009, pp. 619-636.(The full paper appeared Cryptology ePrint Archive Report 2009/385)
[15] A. Lewko and B. Waters. "New Techniques for Dual System Encryptio n and Fully Secure HIBE with Short Ciphertexts". Proceeding of the 7th Theory of Cryptography Conference 2010, LNCS 5978, Berlin: Springer-Verlag press, 2010, pp. 455–479.
[16] L.Y. Zhang, Y.P. Hu and Q. Wu. "Fully Secure Identity-based Broadcast Encryption in the Subgroups". *China Communications*, 2011, Vol. 8, No. 2, 152-158.
[17] L.Y. Zhang, Y.P. Hu and Q. Wu. "Adaptively Secure Identity-based Broadcast Encryption with constant size private keys and ciphertexts from the Subgroups". *Mathematical and Computer Modelling(In press)*. Online press, 2011, http:// dx.doi.ort/10.1016/j.mcm.2011.01.004.
[18] L.Y. Zhang, Q. Wu and Y.P. Hu. Adaptively Secure Identity-based Encryption in the Anonymous Communications. *ICIC Express Letters*, Vol. 5. No. 9(A), 2011, pp. 3209-3216.

Leyou Zhang: male. He received his M.E. and Ph.D. degrees in Applied Mathematics form Xidian University, Xi'an, China in 2002 and 2009, respectively. Now he is an Associate Professor in the department of Mathematical science of Xidian University. His current research interests include network security, computer security, and cryptography. He has published more than thirty papers in international and domestic journals and conferences.

Qing Wu: female. She received her Ph.D. from the Xidian University in 2009. Now she is an Associate Professor in the school of automation of Xi'an institute of posts and telecommunication. Her current research interests include information security and applied mathematics. She has published more than twenty papers in international and domestic journals and conferences.

Yupu Hu: male. He received his Ph.D. from the Xidian University in 1999. Now he is a Professor in the School of Telecommunications Engineering of Xidian University. His current research interests include information security and cryptography. He has published more than a hundred papers in international and domestic journals and conferences. He is a Member of China Institute of Communications and a Director of Chinese Association for Cryptologic Research.

A Power Allocation Algorithm Based on Cooperative Game Theory in Multi-cell OFDM Systems

Ping Wang[1, 2]
1 Broadband Wireless communications and Multimedia laboratory, Tongji University, Shanghai, China.
2 Shanghai Key Laboratory of Digital Media Processing and Transmission, Shanghai, China
Email: shaffer2008@gmail.com

Jing Han[1], Fuqiang Liu[1], Yang Liu[1], Jing Xu[3]
1 Broadband Wireless communications and Multimedia laboratory, Tongji University, Shanghai, China.
3 Shanghai Research Center for Wireless Communications, Shanghai, China
Email: {han_han0307, liufuqiang, yeyunxuana}@163.com

***Abstract*—A centralized resource allocation algorithm in multi-cell OFDM systems is studied, which aims at improving the performance of wireless communication systems and enhancing user's spectral efficiency on the edge of the cell. The proposed resource allocation algorithm can be divided into two steps. The first step is sub-carrier allocation based on matrix searching in single cell and the second one is joint power allocation based on cooperative game theory in multi-cell. By comparing with traditional resource allocation algorithms in multi-cell scenario, we find that the proposed algorithm has lower computational complexity and good fairness performance.**

***Index Terms*—OFDM, resource allocation, game theory, multi-cell, cooperation**

I. INTRODUCTION

In multi-cell systems, it is a great challenge to use the limited radio resources efficiently. Resource allocation is an important means to improve spectrum efficiency in interference limited wireless networks. In distributed systems, a user usually has no knowledge of other users, so a non-cooperative game model is built. In such model, SIR (signal-to-interference ratio) is used to measure system utility and create a utility function. Each unauthorized user allocates resource independently only to maximize its own utility to reach Nash equilibrium. However, for the whole system, system utility is probably not the best.

Therefore, when non-cooperative game theory is applied in resource allocation, there is always a conflict between individual benefit and system benefit [1]. Though some methods, such as the use of the price function, have been proposed to solve this problem, they are difficult in practice. D. Goodman firstly applied non-cooperative game theory to power allocation in CDMA systems [10-12]. In [13-14], the authors studied multi-cell power control in different aspects. The algorithms first chose the optimal cell and then implemented power control among users in single cell. In [15-16], the authors studied static non-cooperative game. The algorithm in [15] implemented more severe punishment to users with better channel condition, so that it effectively kept good fairness among different users. [17] used dynamic game model. It proposed a distributed power control algorithm based on potential game model. In [18], the power allocation among cells was carried out by non-cooperative game, but it did not give the solving process. All these work are mainly based on non-cooperative game theory, which may not maximize the whole system utility.

In centralized wireless networks, since resource allocation and scheduling are performed by a central base station, a cooperative game theory model can be built for resource allocation. In such model, users can cooperate and consult with each other and the system utility is theoretically optimal [6]. Hence, this paper focuses on centralized resource allocation in multi-cell systems. The best resource allocation scheme can only be obtained by jointly allocating subcarriers and power among cells. But the computational complexity is too high to realize. Most practical resource allocation algorithms generally consist of two steps. The first step is to allocate sub-carriers in single cell. The second one is to jointly allocate power in multi-cell.

This paper proposes an algorithm for multi-cell resource allocation in broadband multi-carrier system, which includes:

(1) Sub-carrier allocation

A new sub-carrier allocation algorithm based on matrix searching in single cell is proposed. Firstly, the initial power allocation is finished based on channel environment and rate ratio constraint of different users. Then sub-carriers are allocated after taking both the maximal sum-rate capacity and user's fairness into account, which guarantees the benefit of users located on the cell edge and makes users with poor channel condition obtain sub-carriers as well. What's more, the complexity is reduced compared with the algorithm

doi:10.4304/jnw.6.11.1610-1617

which allocates sub-carriers first and then exchanges sub-carriers [3].

(2) Power allocation

The main problem in multi-cell systems is co-channel interference among adjacent cells. After using statistical channel state information, this algorithm introduces an idea of cooperative game theory. Aiming at maximize the net utility of system (i.e., QoS-satisfied function based on sum-rate capacity), the proposed algorithm models resource allocation process like a cooperative game among users in different cells, and the Nash bargaining solution (namely the assignment result of sub-carriers and power) can be obtained through a power allocation algorithm whose complexity is controllable. The simulation shows that this algorithm can approximate the maximal sum-rate capacity of a multi-cell system while meet users' QoS fairness as well.

The reminder of this paper is organized as follows. Section 2 gives the system model in single cell and multi-cell. And then the proposed resource allocation algorithm is presented in detail in section 3. At last, the effectiveness and rationality of the proposed algorithm are verified by comparing with other traditional algorithms in section 4. Finally, a conclusion is made in section 5.

II. System Model

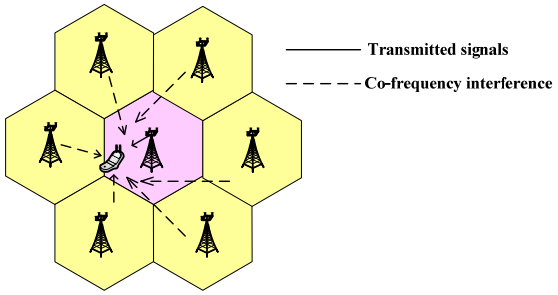


Figure 1. Resource allocation model in multi-cell

The system model is shown in Fig. 1. Assume that the total band of the system is *B* and the number of sub-carriers is *C*. The multi-access mode is orthogonal frequency division multiple access (OFDMA). The degree of fast fading in adjacent sub-carriers is similar so that a group of *S* consecutive sub-carriers with similar fading characteristics can be seen as a sub-channel. Therefore, the number of sub-channels (denoted by *N*) is *C*/*S* and the labels of them are denoted from 1 to *N*. Considering *I* adjacent cells with co-frequency interference, the number of active users in each cell is *K*. Assume that the CSI (channel state information) detected by a mobile station can be fed back to the base station through control channel without error. The base stations among adjacent cells are connected by optical fiber and control information is real time transmission.

A. Single Cell System Model

For a single cell, the downlink resource allocation model is described as follows:

$$
\begin{aligned}
&U = \max_{\alpha_{k,n},p_{k,n}} \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{\alpha_{k,n}}{N}\log_2\left(1+\frac{p_{k,n}\left|h_{k,n}\right|^2}{N_0 B/N}\right)\\
&C_1: \sum_{k=1}^{K}\sum_{n=1}^{N} p_{k,n} \le p_{total}\\
&C_2: p_{k,n} \ge 0,\\
&C_3: \alpha_{k,n} = \{0,1\},\\
&C_4: \sum_{k=1}^{K} \alpha_{k,n} = 1,\\
&C_5: R_1 : R_2 : \cdots R_K = r_1 : r_2 : \cdots r_K
\end{aligned}
\tag{1}
$$

where $\left|h_{k,n}\right|^2$ is channel gain of user *k* on sub-channel *n*, and $p_{k,n}$ is the power assigned to user *k* on sub-channel *n*. Each sub-channel can be considered as an additive white Gaussian noise (AWGN) channel, and N_0 is the power spectral density of such channel. P_{total} presents the total transmission power . $\alpha_{k,n}$ can only be 0 or 1. If it is equal to 1, it means that sub-channel *n* is assigned to user *k*. Otherwise, it is 0. Define the signal-to-noise ratio (SNR) of user *k* on sub-channel *n* as $S_{k,n} = \left|h_{k,n}\right|^2 / (N_0 B/N)$ and the corresponding receiving SNR as $p_{k,n}S_{k,n}$. C_1 restricts that the sum of transmission power of all users does not exceed the maximum transmission power of base station. C_2 restricts that the power assigned to a sub-carrier is not negative. C_3 restricts that a sub-channel only stays in two states, assigned or unassigned. C_4 restricts that a sub-channel can only be assigned to one user. C_5 specifies that the rates the users obtained must meet the requirements of ratio constraints, in which $R_1 : R_2 : \cdots\cdots R_K$ are the rates obtained by users and $r_1 : r_2 : \cdots\cdots r_K$ are the requirements of ratio constraints which should be satisfied.

In this model we assume that users experience independent multipath Rayleigh fading. A base station can obtain the entire CSI. Sub-channel is a basic unit during allocation. Generally, users are located in different places of a cell, so the transmission loss and shadow fading are different. Therefore, the channel gain can be further expressed as follows:

$$h_{k,n} = l_{k,n} s_{k,n} g_{k,n} \tag{2}$$

where $l_{k,n}$, $s_{k,n}$ and $g_{k,n}$ represent transmission loss, shadow fading and multipath fading of user *k* on sub-channel *n*, respectively. The mean gains of them are assumed to be 0dB. If the time scale of resource allocation is a transmission time interval (TTI) and the unit is millisecond, shadow fading and transmission

fading will only depend on user's location. Therefore, the mean SNR can be expressed as:

$$S_k = \left| l_{k,n} s_{k,n} \right|^2 / (N_0 B / N) \tag{3}$$

Define the sum-rate capacity of user *k* as follows:

$$R_k = \sum_{n \in \Omega_k} \frac{1}{N} \log(1 + \frac{p_{k,n} \left| h_{k,n} \right|^2}{N_0 B / N}) \tag{4}$$

where Ω_k is the set of sub-channels which user *k* uses.

B. Multicell System Model

Given *I* adjacent cells with co-frequency interference, the number of active users in each cell is k. $p_{i,k,m}$ and $h_{i,k,m}$ represent the transmission power and channel gain of user k on sub-channel *m* in base station i, respectively. $h^i{}_{j,k,m}$ is the channel gain of this user on sub-channel *m* in co-frequency cell j.

The SIR of this user can be expressed as:

$$\gamma_{i,k,m} = \frac{h_{i,k,m} p_{i,k,m}}{\sum_{j \neq i} \sum_{k=1}^{K} h^i{}_{j,k,m} p_{j,k,m} + \sigma^2} \tag{5}$$

In multi-cell OFDMA systems, sub-channel allocation is assumed to be finished in each cell. $k_{i,m}$ represents the user who is assigned sub-channel *m* in cell i. Then the set of users who need power allocation on co-frequency sub-channel *m* are $M = \{k_{1,m} ... k_{I,m}\}$. Since a sub-channel in one cell can only be assigned to one user during one TTI, the number of users in co-frequency channel equals the number of co-frequency cells. Each co-frequency channel is independent. Therefore, the power allocation in *I* adjacent cells is equivalent to the power allocation among *I* users on co-frequency channels, and the maximizing of system throughput is equivalent to the maximizing of capacity sum on each co-frequency channel in each cell. This can be achieved through cooperation in multi-cell.

III. Resource Allocation Scheme

A. Sub-carrier Allocation

User's channel gain is determined by transmission loss, shadow fading and multipath fading. Besides, transmission loss, shadow fading, together with users' rate ratio are only relative to users. Hence, we can assume that the sub-channels assigned to each user have the same initial power. In a view of average, user's sum-rate capacity also needs to satisfy the requirement of rate ratio constraints. Therefore, how to optimize the assignment of sub-carriers is presented as follows.

$$\begin{aligned}
&U = \max_{x_k, p_k} \sum_{k=1}^{K} \frac{1}{N} x_k \log(1 + p_k S_k) \\
&C_1 : \sum_{k=1}^{K} x_k p_k = P_{total} \\
&C_2 : p_{k,n} \geq 0, \\
&C_3 : \sum_{k=1}^{K} x_k = N \\
&C_4 : 0 \leq x_k \leq N \\
&C_5 : R_1 : R_2 : \cdots R_K = r_1 : r_2 : \cdots r_K
\end{aligned} \tag{6}$$

where x_k is a positive integer, representing the number of sub-channels assigned to user *k*. p_k represents the initial transmission power of user *k* on each sub-channel. R_K represents the sum-rate capacity of user *k* on average.

To solve (6), Lagrange multiplier method is used. The cost function *L* is written as follows:

$$\begin{aligned}
L &= \frac{1}{N} \sum_{k=1}^{K} x_k \log(1 + p_k S_k) \\
&+ \lambda (\sum_{k=1}^{K} x_k p_k - P_{total}) + \mu (\sum_{k=1}^{K} x_k - N) \\
&+ \frac{1}{N} \sum_{k=2}^{K} \beta_k (x_1 \log(1 + p_1 S_1) - \frac{r_1}{r_k} x_k \log(1 + p_k S_k))
\end{aligned} \tag{7}$$

where λ, μ, $\{\beta_k\}_{k=2}^{K}$ are Lagrange multiplier. Find the partial derivative of *L* in relation to x_k and p_k. The following equation is derived after setting both derivatives to 0:

$$\begin{aligned}
&\frac{S_1}{1 + p_1 S_1} \left[\ln(1 + p_k S_k) - \frac{p_k S_k}{1 + p_k S_k} \right] \\
&= \frac{S_k}{1 + p_k S_k} \left[\ln(1 + p_1 S_1) - \frac{p_1 S_1}{1 + p_1 S_1} \right]
\end{aligned} \tag{8}$$

(8) is correct for $k = 2, 3, ..., K$. Combining (8) with (6), the optimal initial power allocation on average can be achieved through Newton-Raphson method [2].

After the initial power is determined, the sub-carrier can be assigned in single cell. The proposed sub-carrier allocation is based on matrix-searching and has three steps. First, the matrix of sum-rate capacity of *K* users on *N* sub-channels is figured out. Thus the question becomes how to find the corresponding user for each sub-channel in a *K*×*N* matrix. Second, sort all the users according to their rate requirements and assign sub-carriers for the first time, ensuring that each user is assigned a sub-channel at least. Define the fairness function δ as:

$$\delta_k = \frac{R_k}{r_k} \tag{9}$$

where R_k represents the obtained sum-rate capacity. r_k represents the required ratio of rate capacity. Finally, finish allocating sub-carriers based on the fairness

function δ. The algorithm is described in detail as shown in Fig. 2:

(1) Initialize $R_k = 0$, $\Omega_k = \varnothing$, $\rho_{k,n} = 0$, $k = 1,2,\ldots K$, $A = \{1,2,\ldots N\}$

(2) Figure out the matrix of rate capacity of k users on n sub-carriers, denoted by $R_{k,n} = \frac{1}{N}\log_2(1 + p_{k,n}S_k)$

(3) Sort users according to their required rate in descending order and repeat K times to find n in row vector of matrix R, which should satisfy $R_{k,n} \geq R_{k,j}$ and $j \in A$. Then set $\rho_{k,n} = 1$, $\Omega_k = \Omega_k \cup \{n\}$, $A = A - \{n\}$, $R_k = R_k + R_{k,n}$

(4) When $A \neq \varnothing$, find k which satisfies $\delta_k \leq \delta_i$ and $1 \leq i \leq K$. If k is obtained, find n which satisfies $R_{k,n} \geq R_{k,j}$ and $j \in A$. Then set $\rho_{k,n} = 1$, $\Omega_k = \Omega_k \cup \{n\}$, $A = A - \{n\}$, $R_k = R_k + R_{k,n}$

Figure 2. The matrix-searching based sub-carrier allocation algorithm in single cell

B. Power Allocation Strategy Based on Game Theory

Power allocation on co-frequency sub-channel is realized via a cooperative game process. To make a user not simply pursue the utility maximization in power allocation, the interference to other users should also be considered. Thus the pricing function of transmission power obtained by each user is introduced into the cooperative game theory, which represents the cost that the user has to pay for using system resource. The system utility can reach the optimal state when each user arrives at a tradeoff between the obtained utility and the produced interference.

In a multi-cell system, let $G = \{P, A, S, I, U\}$ present a gaming process, whose parameters are described below:

(1) P represents participants, who are a set of users experiencing co-frequency interference on the same sub-channel of each cell.

(2) A represents strategy set, which include $A = \{p_1, ..., p_I\}$, where $p_i = [p_{i,1,1}, ..., p_{i,k,n}]$.

(3) S represents the gaming order, whose default value is conducting strategy choosing at the same time.

(4) I represents information. Every participant in game knows the strategy choices of all participants.

(5) U represents the income, that is, utility function. Let vector P represent the set of transmission power obtained by all users after the game, and $u_k^c(p_k, p_{-k})$ represent the net utility obtained by user k in the end, which can be expressed as follows:

$$u_k^c(p_k, p_{-k}) = u_k(p_k, p_{-k}) - c_k(p_k, p_{-k}) \quad (10)$$

where p_k and p_{-k} are the transmission power chosen by user k and other users after the game, respectively. $u_k(p_k, p_{-k})$ is the utility function of user k ignoring pricing factor. $c_k(p_k, p_{-k})$ is the pricing function of user k. In a cooperative game, the goal is to maximize system utility, that is:

$$\max U = \max \sum_{i=1}^{I}\sum_{k=1}^{K} u_k^c(p_k, p_{-k}) \quad (11)$$

In this paper, the sum-rate capacity is used as the utility function, and on any co-frequency sub-channel, the utility function of user k is:

$$u_k(p_k, p_{-k}) = \log(1 + \frac{p_{i,k,m}h_{i,k,m}}{\sum_{j \neq i} h_{j,k,m}^i p_{j,k,m} + \sigma^2}) \quad (12)$$

where $p_k = p_{i,k,m}$ and the pricing function increases linearly with the transmission power as follows:

$$c_k(p_k, p_{-k}) = \lambda_k p_k \quad (13)$$

where λ_k represents the pricing factor, which gives the price of power per unit. If the priorities of users are identical, so is the pricing factor. The function of net utility is derived below:

$$u_k^c(p_k, p_{-k}) = \log(1 + \frac{p_{i,k,m}h_{i,k,m}}{\sum_{j \neq i} h_{j,k,m}^i p_{j,k,m} + \sigma^2}) - \lambda_k p_k \quad (14)$$

After sub-carriers are finished allocation in single cell, multi-cell power allocation based on cooperative game theory are conducted, which is shown in Fig. 3. The gaming goal in a multi-cell system is to obtain Nash bargaining solution, which maximizes the net utility of system. To guarantee the fairness among users, the system strategy set is updated until the following conditions are met: the number of cells with power gain is larger than that with power loss and the whole system must obtain power gain. If the net utility does not increase after the strategy set is changed, Nash bargaining solution is obtained.

(1) Allocate sub-carriers in every cell. The result is taken as the initial value of multi-cell power allocation.
(2) For each co-frequency sub-channel, calculate the net utility function of each user in single cell without or with the cooperation, which are denoted by $u_k(p_k, p_{-k})$ and $u_k^c(p_k, p_{-k})$, respectively. Let $\Delta u_k = u_k^c(p_k, p_{-k}) - u_k(p_k, p_{-k})$.
(3) Starting from the user with the smallest Δu_k, change the choice of strategy set and re-calculate the net utility function. If the net utility increases and the number of cells with power gain is larger than that with power loss, update the strategy set. Otherwise, return to step (2).
(4) Loop all the co-frequency users in sequence till the net utility of system does not increase.
(5) Loop all the sub-channels in sequence till the system strategy set does not change. To this point, the strategy set is the result of Nash bargaining solution.

Figure 3. The multi-cell power allocation algorithm based on cooperative game theory

Ⅳ. Simulation Analysis

A. Simulation environment

The power allocation algorithm in multi-cell OFDM systems is simulated by MATLAB. Frequency selective channels contain six independent Rayleigh multipath and the maximum delay spread is 5us. Other system parameters are shown in Table I. This simulation uses discrete event-driven mechanism for dynamic simulation. In order to obtain stable and reliable performance, the results are obtained from the average of 10000 implementations on random channel.

TABLE I.
Simulation Parameters

Parameters	Value
Number of co-frequency cells	2
Number of sub-carriers	1024
Total transmission power	1W
Total system bandwidth	1M
Number of users	2-10
AWGN power spectral density	-80dBw/Hz
Average channel gain	0-30dB

The proposed algorithm is compared with other three traditional resource allocation algorithms, namely:

Algorithm 1: represents the proposed algorithm, which includes sub-carrier allocation based on matrix searching and power allocation based on cooperative game theory.

Algorithm 2: consists of direct sub-carrier allocation [7-9] and equal power allocation.

Algorithm 3: consists of direct sub-carrier allocation [7-9] and Water-filling power allocation [4][5].

Algorithm 4: consists of the proposed sub-carrier allocation in the paper and equal power allocation.

B. Results and discussion

Fig. 4 and Fig. 5 compare the four resources allocation algorithms from the perspective of system capacity, which plot the normalized system throughput. The users' rate ratios in Fig. 4 and Fig. 5 are 4：2：1：…:1 and equal, respectively. As shown in Fig. 4 and 5, algorithm 3 has the highest throughput and the proposed algorithm has better throughput than algorithm 2 and 4. This is because in algorithm 3, sub-carrier allocation and power allocation are based on throughput, which can achieve the highest system throughput. Although sub-carrier allocation based on matrix searching in algorithm 1 will lose part of system throughput for its paying attention to fairness, the result of total resource allocation in algorithm 1 is still better than those in algorithm 2 and 4. This is because algorithm 1 allocates power based on cooperative game theory, which is to maximize the net utility of system and thus can approximate the maximal sum-rate capacity in multi-cell systems. Furthermore, its superiority is enhanced as the number of users is increased and the system capacity is more approximate to that of algorithm 3.

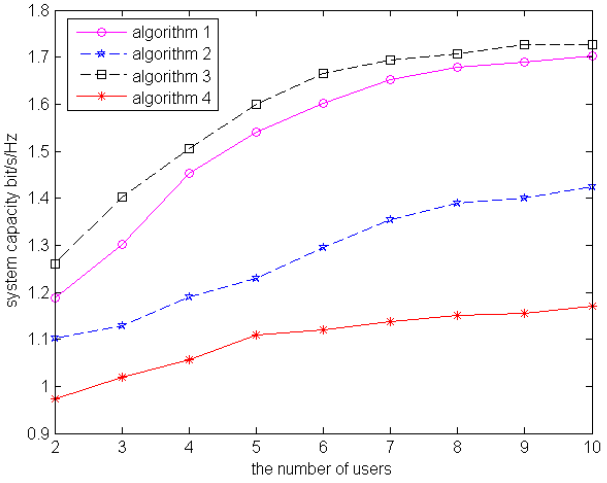


Figure 4. System throughput when the requirements of rates are unequal

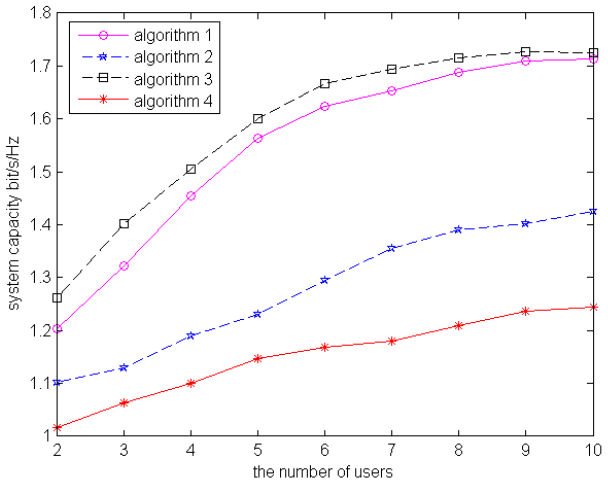


Figure 5. System throughput when the requirements of rates are equal

Next the proposed algorithm is compared with algorithm 3 and 4 from the perspective of system fairness. Figs. 6, 7 and 8 plot the sum-rate capacity of all users in a cell with users' rate ratio of 4:2:1: ...: 1 when the number of users in each cell is 3, 5 and 8, respectively. We can see that the proposed algorithm is the best in terms of fairness and it can most approximate the requirement of users' rate ratio. Algorithm 4 is better than algorithm 3 in terms of fairness because it adopts the proposed subcarrier allocation algorithm in this paper. However, algorithm 4 has smaller sum-rate capacity than algorithm 1 and its fairness is also slightly weaker. This is because equal power allocation does not differentiate channel gain for users in different locations. Though algorithm 3 maximizes the sum-rate capacity, it does not nearly meet any requirements of users' rate ratio. Algorithm 2 is not taken into consideration in comparison because of its worst fairness. Therefore, the proposed algorithm improves system fairness greatly, which can improve the spectral efficiency on the cell edge and guarantee the rate demands of all users.

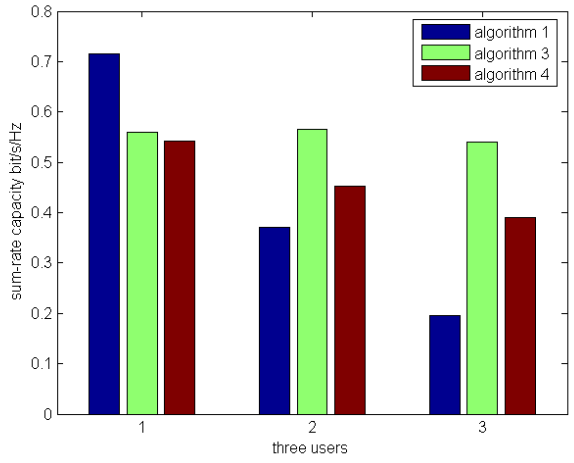


Figure 6. System fairness when the number of users is 3

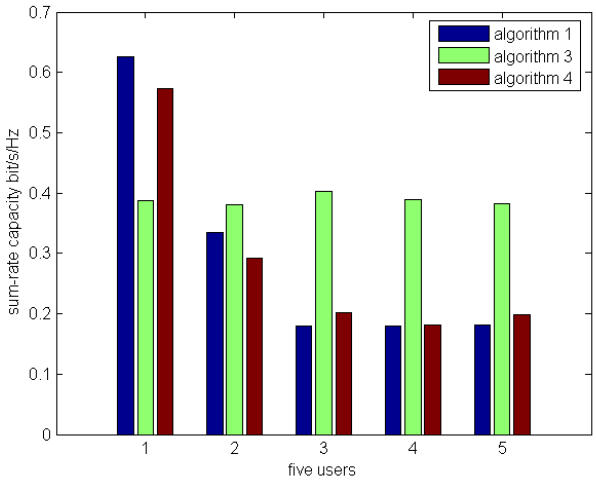


Figure 7. System fairness when the number of users is 5

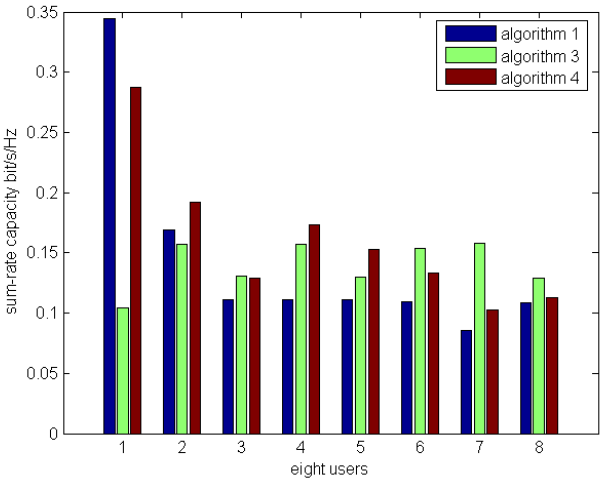


Figure 8. System fairness when the number of users is 8

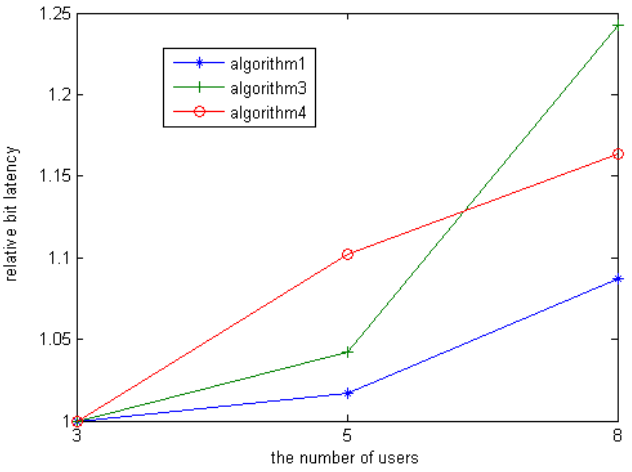


Figure 9. Relative bit latency when the number of users is 3, 5, 8

From Fig. 9, we can see that when the number of users increases, the proposed algorithm has the slowest growth in relative bit latency. This is because our algorithm considers the fairness among users and gives

each user a relatively fair chance to obtain the time-frequency resource. On the contrary, we can find that algorithm 3 and algorithm 4 only pursue high sum rates, and give all the resources to the users with good channel condition. Thus, they fail to meet the requirements of other users.

From the perspective of computational complexity, the proposed algorithm requires numerical iteration. During initial power allocation before subcarrier allocation, there is a numerical iteration. During multi-cell power allocation based on cooperative game theory, there is also a numerical iteration. Thus it has higher algorithm complexity compared with equal power allocation.

As sub-carrier allocation is considered, compared with the algorithm which allocates sub-carriers first and then exchanges sub-carriers, the proposed matrix-searching algorithm reduces complexity actually. However, the initial power allocation requires numerical iteration, which is more complex than sub-carrier allocation itself. Thus compared with the sub-carrier allocation algorithm with equal initial power, the proposed algorithm is a little more complex. But its performance is improved greatly since users in different location of a cell have different channel gain.

As power allocation is considered, although equal power allocation is the simplest, it is not used in reality for its inability to meet the requirements of users' rate. Compared with algorithm 3, the proposed algorithm indicates that sub-carrier allocation in single cell has approximated the system optimal solution to some extent. And in multi-cell cooperation, the strategy set of game theory needs to be changed only when the system utility is increased and the number of cells with power gain is bigger than that with power loss. Thus, the complexity caused by the proposed resource allocation is lower than that by Water-filling power allocation in algorithm 3.

In general, the initial power allocation in sub-carrier allocation and the multi-cell power allocation based on cooperative game theory both have the linear complexity $O(k)$ (where k is the number of users). Although this algorithm's complexity is higher when compared with that of ideal resource allocation algorithms such as equal power allocation, it is actually decreased when compared with some currently applied algorithms such as Water-filling. Furthermore, the proposed algorithm is only slightly worse than algorithm 3 in terms of system capacity. And it is the best in terms of system fairness.

Ⅴ. Conclusion

This paper proposes a resource allocation algorithm based on game theory in a centralized multi-cell OFDM system, including the matrix-searching based subcarrier allocation algorithm in single cell and the joint power allocation algorithm in multi-cell based on cooperative game theory.

The proposed algorithm is compared with other three traditional resource allocation algorithms from the perspective of system capacity, fairness and complexity. The results show that the proposed algorithm achieves a good tradeoff between system throughput and fairness. And its complexity is reduced compared with the multi-cell water-filling algorithm which achieves highest throughput. Furthermore, it can nearly satisfy the requirements of users' rate ratio and the users on the cell edge can get a significant spectral efficiency gain.

Acknowledgment

This work was supported by the National Science and Technology Major Project of China under Grant 2010ZX03002-007, Sino-Finland International Cooperation Project under Grant 2010DFB10410, Shanghai Science and Technology Committee under Grant 09511501100, the Opening Project of Shanghai Key Laboratory of Digital Media Processing and Transmission and National Natural Science Foundation of China under Grant 61073153.

References

[1] Han Tao. "Spectrum Allocation Technology Based on Game Theory in Cognitive Radio Networks," *Doctor thesis*: Beijing University of Post and Telecommunications, 2009, pp. 48-67.

[2] Wong C. Y., Tsui C. Y., Cheng R. S., et al. "A real-time subcarrier allocation scheme for multiple access downlink OFDM transmission," *Proceedings of IEEE VTC*. Amsterdam, Netherlands, 1999: 1124～1128.

[3] Kim Keunyoung, Kim Hoon and Han Youngnam, et a1. "Iterative and greedy resource allocation in an uplink OFDMA system," *Proceedings of IEEE PIMRC*. Barcelona, Spain, 2004: 2377～2381.

[4] Kim Keunyoung, Kim Hoon and Han Youngnam. "Subcarrier and power allocation in OFDMA systems," *Proceedings of IEEE VTC 2004*. Los Angeles, California, USA, 2004: 1058～1062.

[5] Choe Kwang Don, Lim Yeon Ju and Park Sang Kyu. "Subcarrier allocation with low complexity in multiuser OFDM systems," *Proceedings of IEEE MILCOM 2008*. Monterey, CA, United States, 2004: 822～826.

[6] Zhang Guopeng. "Research on Resource Allocation and Cooperative Mechanism in Wireless Networks Based on Game Theory," *Doctor thesis*: Xidian University, 2009, pp. 62-84.

[7] Xu Wenjun. "Resource Allocation Strategies Study in Broadband Wireless Communication System," *Doctor thesis*: Beijing University of Post and Telecommunications, 2008, pp. 54-90.

[8] Shen Zukang, Andrews J. G., Evens B. L.. "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Transactions on Wireless Communications*, 2005, 4(6): 2726~2737.

[9] Hu Yahui. "Researches on Radio Resource Management in MIMO OFDM Systems," *Doctor thesis*: Beijing University of Post and Telecommunications, 2009, pp. 45-78.

[10] Goodman D., Mandayam N. "Power control for wireless data," *IEEE Wireless Communications*, 2000, 7(2): 48-54.

[11] Saraydar C U, Mandayam N B, Goodman D J. "Efficient power control via pricing in wireless data networks," *IEEE Transactions on Communications*, 2002, 50(2):291-303.

[12] Saraydar C U, Mandayam N B, Goodman D J. "Pricing and power control in a multicell wireless data network,"

IEEE Journal on Selected Areas in Communications, 2001, 19(10): 1883-1892.
[13] Alpcan T, Basar T, Dey S. "A power control game based on outage probabilities for multicell wireless data networks," *IEEE Transactions on Wireless Communications*, 2006, 5(4):890-899.
[14] Sarma G, Paganini F. "Game theoretic approach to power control in cellular CDMA," *IEEE VTC*. ORlando, FL, United States, 2003. pp. 6-9.
[15] Zhong Chong-xian, Li Chun-guo, Yang Lu-xi. "Dynamic Resource Allocation Algorithm for Multi-cell OFDMA Systems Based on Noncooperative Game Theory," *Journal of Electronics & Information Technology*, 2009, 8(31):1935-1940.
[16] H. Kwon, B. G. LEE. "Distributed resource allocation through noncooperative game approach in multi-cell OFDMA systems," *IEEE ICC 2006*, Istanbul, June 2006.
[17] Qiu Jing, Zhou Zheng. "Distributed Rescource Allocation Based on Game Theroy in Multi-cell OFDMA Systems," *International Journal of Wireless Information Networks*, 2009, 1(16):44-50.
[18] Wang L, Xue Y S, Schulz E. "Resource allocation in multicell OFDM systems based on noncooperative game," *The 17th Annual IEEE International Symposium on Personal Indoor and Mobile Radio Communications*. 2006.1-5

Ping Wang, born in China, 1978-2-28. He graduated from the department of computer science and engineering at Shanghai Jiaotong University, China and received Ph. D. degree in 2007. His major field of study is wireless communication. He joined the college of electronic and information engineering at Tongji University in 2007 and now is a lecturer. His current and previous interests include routing algorithms and resource management in wireless networks, vehicular ad hoc network and streaming media transmission.

Jing Han, born in China, 1987-3-7. She graduated from the department of information and communication engineering at Tongji University and received B.S. degree in 2009. Her major field of study is wireless communication. Now she is a graduate in the department of information and communication engineering at Tongji University. Her main research interests are in enhanced MIMO and radio resource management for the next generation mobile communications.

Fuqiang Liu, born in China, 1963-3-7. He graduated from the department of automation at China University of Mining and received Ph. D. degree in 1996. His major field of study is signal processing. Now he is a professor in the department of information and communication engineering at Tongji University. His main research interests are in routing algorithms in wireless broadband access and image manipulation.

Yang Liu, born in China, 1987-8-10. He graduated from the department of information and communication engineering at Tongji University and received B.S. degree in 2010. His major field of study is wireless communication. Now he is a graduate in the department of information and communication engineering at Tongji University. His main research interests are in relay technologies and radio resource management for LTE systems.

Jing Xu, born in China, 1975-5-6. Now he is a researcher in the Shanghai Research Center for Wireless Communications. His main research interests are in system architecture, networking and resource allocation in B3G/4G systems.

Expectation Value Calculation of Grid QoS Parameters Based on Algorithm Prim

Kaijian Liang
School of Application & Technology, Hunan Institute of Engineering, Xiangtan, China
Email: liangkaijian@sina.com

Linfeng Bai
School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China

Xilong Qu
School of computer and communication, Hunan Institute of Engineering, Xiangtan, China

***Abstract*—From the perspective of selecting service by QoS attributes, a computation method of QoS expectation value, which is based on Algorithm Prim, was presented to provide support for selection of service. On the basis of the ability of service providers, by Algorithm Prim, this method succeeded in calculating a set of balanced expectation values of QoS. Selection of service based on these QoS values would be beneficial to optimization of system resources and protection of the users of those services. An example with analysis has been provided to demonstrate the feasibility and effectiveness of the method.**

***Index Terms*—grid computing, service matchmaking, QoS parameters, algorithm prim**

I INTRODUCTION

To network technique as the core, new generation network computing environment is a hotspot and leading edge domain of current international research. The goal of network system construction combining with network technique is integrating computing facility; storage device; service facility and instrument from different place, building large computing and data processing shoring of foundation construction and achieving wide range sharing; effective aggregating and full releasing abased on computing resource; data resource and service resource on the internet. Traditional internet achieved the connection of computer hardware, Web achieved the connection of web page, when grid try to achieve the connection of all resource, including computing resource; storage resource; communication resource; software resource; information resource and knowledge resource. In Service-Oriented network environment, discovering and selection is a very important link, in the process, we need not only guarantee the veracity of service location, but also consider the need of user, so foundation and selection abased on QOS service arise.

Service foundation is a process which could meet the need of specific service of user in the network, and achieve automation and intellectualization. There is not strict divide between service foundation and service selection, in some research work, service foundation includes service selection. Generally speaking, service foundation emphasizes the process in founding candidate service collection, namely the way on gaining candidate service, but service selection emphasizes selecting a suitable service for user from candidate service collection. In this sense, service foundation is the preorder step of service selection, as a roughing process, the result collection is the object of service selection operating. The size of result collection; the way of gaining; and veracity have direct effect on service selection strategy. If it adopts a very strict standard for the need of all users in service foundation, service selection has to do nothing, and vice versa.

From the view of user, they always want to find an optimal service, but they can't, owing to price; service; times; factors. Practical application, users take two sides into consideration: one side, meeting the need of QOS with a better cost performance; on the other side, different user has different attention on the attributive of QOS except satisfy basic QOS needs, some have specific requests on service price, others may pay attention to the response time of service or creditworthiness. So you say, when taking the two side into consideration, aiming at specific user, it is user to measure the satisfaction of QOS needs at last. User has different attention on the attributive of QOS; it is the QOS needs predilection of user. But in practical application, it is very important to study the QOS needs predilection to satisfy QOS needs of specific user.

It has important significance to realize the potential value of gridding service resource. In microeconomics, price is the effective lever of adjusting supply-demand relations between consumer and commodity. In foreseeable gridding technology application domain, there are many similarities between behaviors of gridding service with user and consuming behavior of commodity in market with consumer. Combining microeconomics theory, it can introduce market mechanism into "paid service" of gridding service. In fact, "paid service" itself reflect an effective mechanism that configures "gridding service" resource, and it is beneficial for the whole operation order of gridding environment. Because lever of price can have an effect guiding impact on

doi:10.4304/jnw.6.11.1618-1624

network using behavior in different user, as a adjusting user incentive mechanism in using “gridding service” resource, despite “paid service” cannot distribute resource and restrain it’s behavior explicitly when facing limitless condition at present network environment. Therefore, this paid use in gridding service has “benefit” property in commercial activity.

Because of basic mechanism sustained by QOS attribute, it can configure, discover, select, distributes resource on the basic of QOS attribute. In current many system, not only grid system, but also distributed system and Peer-to-Peer system, all its introduce SLA mechanism, which can describe QOS information resource and bind specific application. Some researcher introduces Service data into grid service, which can be used to describe a kind of grid service information including QOS information. G-QOSM base on OGSA, provides a QOS management model facing service, and expand grid service description on the foundation of service data. It sustains resource based on QOS attribute and service foundation, also the latest GGF standard, and match OGSA’ latest standard. QGS in G-QOSM frame exist in every domain, keeping in touch with user application program, and catch service request constrained by QOS parameter. According to the given parameter, it can find the best matching service and consult SLA; Base on foundation sign a contract to guarantee user service quality.

The discovery and selection of service based on QoS attributes can facilitate the optimization of system resources and guarantee the quality of customer service, which has been a hot research topic in grid computing. Moreover, it is also an issue to be sorted out for the application and commercialization of grid computing. In the commercialized environment of service-oriented grid application, the users will consider their own benefit and efficiency while using the service. Whereas among a number of candidate services, the way users determine the equilibrium requirements of QoS appears critical as equilibrium requirements of QoS have a direct impact on QoS matchmaking parameters and the selection of services. Therefore, it is essential for users to present the expected value of QoS parameters and method of computing [1].

Up to date, similar researches have focused on models of service selection and algorithm, esp. establishing effective and applicable models combining closely the system structure so as to improve the system efficiency [2,3]. As for Algorithm, the main interest lies in how to improve the precision and accuracy of algorithm and stress the effectiveness of computing [4, 5]. About the estimation of parameter, some researchers have been carried out in related fields. For example, for the estimation of network performance, the reference literature no.6 [6] as listed has proposed a method of computing which can be used to estimate path capacity, on the basis of algorithm which can deal with the estimation of capacity of end-to-end single congestion path and available bandwidth. In literature no.7 as listed [7] the estimation method of discrete wavelet transform is applied in the research of synthesized business flow of high-speed Internet. The related work in parameter estimation has received adequate attention and plays a great role in the corresponding fields. These achievements have provided useful theoretical basis and method for selection of service and protection of quality of user service, although the function of QoS requirements was ignored and no specific result was achieved. For the optimization of system resources and protection of the user service efficiency, this paper will study how to calculate the value of expected QoS parameters.

Source reservation technology in service-oriented computing environment is a basic technology for service quality control, but there are still great difficulties. On one hand, factors such as network environment heterogeneity, the breadth of distributed independent nodes, node management and complexity of security strategies, etc, increase the difficulty of resource reservation; on the other hand, reservation has a lot of key issues to be resolved, including reservation technical fault tolerance, the validity of reserved resources, resource sharing to be faced by reservation, etc, which require valid, reliable and robust reservation technology; while not increase too much system expense and ensure not affect the overall system performance.

With the development of computer science, graph theory progress at an alarming rate, and it is a major embranchment in applied mathematics. On one hand, computer science provides computing equipment for graph theory; on the other hand, it needs graph theory to describe and solve many problems in modern sciences practical application. Graph theory was applied to many domains as a method or tool in describing the relation of affairs at present, such as computer science, physics, chemical, operational research, information theory, cybernetics, network communication, social science, economic management, military, national defense, and agriculture and industry production. Prim is an important method to solve the weighted graph shortest or the optimal path problem in graph theory, and then it can be used to project decision described by graph theory.

II. QoS Parameters of Grid Services and Service Matchmaking

A Efficiency Type Qos Parameters

"Efficiency" is a term used in the field of economic management and means "income", "interest" originally [8].Network application should also follow the principles of "market economy" and commercialized "efficiency" also exists. As economic grid environment is concerned, owing to the existing "commodity market”, economic laws also function. Users of service expect not only the basic function but also others such as the most convenient and safe service at the minimum cost. Consequently, the users’ requirement of QoS is also accompanied by pursuit of "efficiency" and the QoS attributes of service also include the consideration of "benefit”. Both parties of supply and demand of the service follow the rules of market economy for QoS matching parameters. On the users’ side, economic benefits constitute the prior consideration, of which service the price, response time may be included in the cost efficiency type QoS parameters which the smaller, the better. But, for other

QoS parameters such as credit and reliability which can be listed in economic and social efficiency type parameters, the bigger value is preferred..

Classification of grid QoS parameters on the basis of efficiency has it practical value in application. Based on efficiency, users can carry out their calculation of QoS parameters matchmaking by means of certain effective algorithm when they implement resources discovery and selection of service so as to decide the most appropriate service for themselves and get the best efficiency and provide groundings for the specific service finally. On the other hand, it also helps keep the balance between supply and consumption of the resources and improves the level of optimization of the system and operational effectiveness of resources.

B Service Matchmaking based on Qos Attributes

In the service-oriented computation, a unified port can be abstracted from service for designated access to various resources including computation, storage and network. In practical implementation a unified service port can be formulated hidden to users. For example, a computation service can be done by a single or multiple processing machines, of which details need not directly be expressed in the service contract. In other words, the granularity of service function is changeable and the function can be realized by a single host or distributed system [9]. It thus provides a possibility that the QoS attributes are made as a part of the port so that the system can select among services based on QoS attributes, which makes it easier to ensure the QoS requirements of users.

To make service discovery and selection based on the attributes of QoS, it is required to establish the QoS attribute set for each service and determine the corresponding QoS parameters. When the user applies for service, firstly they are supposed to declare their QoS requirements, then the system can make matchmaking calculations according to the candidate QoS service attributes to discover the service to satisfy the requirements. To be specific, it is to match the QoS parameters of the service with the required parameter of the user. In this way the quality of customer service is ensured [9].

One of the ways to establish QoS attributes is to extend WSDL&UDDI. The purpose of extending WSDL is to better describe service, add QoS attributes to the description of WSDL and expand the service attributes. For instance, a new genre of service QoS can be added to WSDL [10] to describe the various QoS attributes of service. Meanwhile, corresponding extend is also supposed to be implemented in UDDI so that when the service in the UDDI is published, users can discover and select service according to QoS attributes. With the support of the service discovery and QoS attributes, we guarantee the demand of users for QoS more closely.

There are three functions in pretreatment of data named standardization: firstly, comparing size by different type' attribute value. If QOS attribute data is different, weight comparison would not express easily. Secondly, the not dimension, if the QOS attribute dimension is different, attribute would not common measure. Even the same attribute, it may use different prickle, then the different numerical value. In various kinds multiple target assessment method, assessing require remove the effect of dimensional selection on assessment result, this is the not dimension. It tries to eliminate dimension, reflecting the good or bad of attribute value with only the size of numeric. Thirdly, the normalization, different type attribute value numerical value size is different in the primary attribute value table, putting it into the interval between 0 to1.

Besides, it also could solve the incomplete compensatory by nonlinearity transformation or other methods in pretreatment of data. There are many data preprocessing method, including linear transformation, standard 0-1 transformation, vector standardization, and so on. This text adopts the following method to dispose.

III ESTIMATION OF QOS PARAMETER EXPECTATION VALUE BASED ON PRIM ALGORITHM

A The Prim Algorithm of Minimum Spanning Tree (MST)

The minimum spanning tree of the graph can be obtained by means of prim algorithm in an undirected connected graph. This algorithm, like Kruscal algorithm, is also widely used in multitudinous domains such as network, civil engineering and so on to solve many practical problems [11].

Kruscal is a very mature arithmetic in graph theory, it can evaluate shortest path tree in a weighted undirected connected graph. According to limbic weight number compositor from a small beginning into a force, it investigates each side of graph G side collection T. If the been investigated two peaks belong to two different connected component, then putting this side into the selected side collection TE, meanwhile, connecting two connected component to one connection component; else rounding this side. And so on, if the connected component number of T is 1, this connected component is one of G' minimum spanning tree.

Prim Algorithm suppose an undirected connected graph is $G=(V,E)$, the two tuples represents the set of points and edge set respectively, then the minimum spanning tree of G is $T=(U,TE)$. The basic idea of prim algorithm is: the initial status is $U=\{v_0\}$, $TE=\{\}$ and then repeat execution of the following operations: among all sides of $u\in U$, $v\in V-U$, find a side of minimum cost (u,v) and merge it into the collection TE and at the same time merge v into U until $U=V$. Then in TE there must be $n-1$ sides. T makes the Minimum Spanning Tree. The specific algorithm pseudo-code is described as :

1. Initialization: $U=\{v_0\}$ (v_0 means any vertex in V); $TE=\{\}$;

2. The cycle stops until $U=V$

(1)Among all sides of $u\in U$, $v\in V-U$ find a side of minimum cost (u,v);

(2) $TE=TE+\{(u,v)\}$;

(3) $U = U + \{v\}$ 。

Obviously, the key in Prim algorithm is to find the shortest side to connect U and $V-U$ to expand the spanning tree T. The spanning tree selected in this way bears the minimum overall weights. With regard to the efficiency type QoS parameters of users in grid computing environment, by means of proper method of modeling, these parameters can be converted to the form of undirected connected graph, by prim algorithm, a spanning tree of minimum overall weights can be produced. Namely, a QoS parameter expectation value which keeps equilibrium between both parties of demand and supply can be obtained.

B Efficiency Type QoS Parameter Modeling

The description of candidate services shows the ability of the QoS grid services to provide users service. For the modeling of QoS parameters by graph structure, the QoS parameters of the values in same or different types should be unified by a common measure standard. Owing to the difference in the capacity of each candidate service, there can be a huge discrepancy in the values so that standardization is necessary for the unification of the measure.

Definition 1.The standardization of numeric QoS parameters. It includes QoS Parameters such as service price, response time etc. If a numeric QoS parameter is $q_i, i \in Z$, and the corresponding stanardized one is $q_i^{'}$ then

$$q_i^{'} = q_i \Big/ \sum_{i=1}^{n} q_i \ , i, n \in Z$$

Definition 2. Standardization of ratio type QoS parameters. It includes QoS Parameters such as reliability, credit etc. Suppose a ratio type QoS parameter is $q_i, i \in Z$ and the corresponding stanardized one is $q_i^{'}$, then

$$q_i^{'} = (1 - q_i \Big/ \sum_{i=1}^{n} q_i) \ , i, n \in Z$$

The following theorem will show many a side will require at least n vertexes to form a single connected complete graph or dense graph.

Theorem 1. Suppose there are e sides and n vertexes are required to construct a single connected complete graph or dense graph. And n satisfies:

$$n = \left| \sqrt{1+8e}/2 + 1 \right|$$

Proof: Mathematical induction is applied. When $e = 1, 2$ and by calculation with the theorem, we get $n = 2$, $n = 3$ Obviously it is true. Suppose $e = k$ and at least

$$n = \left| \sqrt{1+8k}/2 + 1 \right|$$

vertexes is required; when

$$e = k + 1 \ (k \in Z),$$

$$n^{'} = \left| \sqrt{1+8(k+1)}/2 + 1 \right|,$$

Obviously, there is:

$$\left| \sqrt{1+8(k+1)}/2 + 1 \right| \geq \left| \sqrt{1+8k}/2 + 1 \right|$$

within, and

$$\left| \sqrt{1+8(k+1)} - \sqrt{1+8k} \right| = \left| \sqrt{8k+9} - \sqrt{8k+1} \right| \leq 2$$

when the equality establishes, a single noncomplete connected graph is formed, $n^{'} = n$, only one side is to be added to the original graph; if the equality doesn't establish, then definitely $n^{'} = n + 1$, when one side and one vertex are added to the original graph, a single connected dense graph is formed. The above two situations conform to the reality. Q.E.D

Definition 3. Weighted-edge of QoS. It represents QoS parameter and the weight of the side is the standardized value of such QoS parameters.

Definition 4. Single connected graph of QoS attributes. The single connected graph made with certain type of QoS $G_{QoSType} = (V, E)$, E stands for the collection of QoS weighted edges, v stands for the collection of vertexes related to the QoS weighted side. If $|E| = e, e \in Z$, then

$$|V| = \left| \sqrt{1+8e}/2 + 1 \right|.$$

Among the multiple candidate services, definition 4 establishes an association model for the related QoS attributes and each QoS parameters are closely related to each other via the model which thus provided basis and groundings for examining the relationships between those QoS parameters.

C Estimation of QoS parameter expectation value

By establishing the QoS attributes single connected graph of each candidate service by Definition 4, with Prim algorithm we can get a spanning tree of minimum dissipation value, namely, the QoS parameter expectation value of minimum dissipation value, which can be used for next stage of selection of service .

Suppose the candidate service collection is

$$S = \{ s_i \mid i \in Z^+ \}$$

and the corresponding QoS attributes collection is:

$$QoS_s = \{ q_{ij} \mid i \in [1, n], \ j \in [1, m] \}$$

in which j actually stands for the type of QoS attributes. Here follows the actual method of calculation:

While $j \leq m$

do

{

1. Establish the QoS attributes single connected graph G_j of $|E| = n$;

2. Initialize $G_j : U = \{v_0\}, TE = \{\}$;

3. The cycle goes on until $U = V$

(1)Among all sides of $u \in U$, $v \in V - U$ find a side at minimum cost (u, v);

(2) $TE = TE + \{(u, v)\}$;

(3) $U = U + \{v\}$;

4. To get the expectation value of q_j ,

$$E(q_j) = \sum_{i=1}^{n} q_{ij} \times \left(\sum_{k=1}^{|TE|} e_k \Big/ |TE| \right)$$

Or

$$E(q_j) = 1 - \sum_{k=1}^{|TE|} e_k \Big/ |TE| ,$$

in which $e_k \in TE$.

}

Then the targe QoS parameter expectation value is:

$$E(QoS_s) = (E(q_j) \mid j \in [1, m]) \text{ 。}$$

As far as the density of probability is concerned, a random collection of candidate services can reflect the equilibrium distribution of the service provider. Consequently, with such capacity of service provision, users can estimate the reasonable QoS expectation parameter value by means of the above mentioned method and use it as the groundings for discovery and selection of service so that then they can maximize the efficiency while using the paid service.

IV THE EXAMPLE AND ANALYSIS

Hereby let's demonstrate the process of the specific method of calculation and compare it with Kruscal algorithm [11] used for solving the same problem. Suppose

$$S = \{s_i \mid i \in [1,5]\} ,$$

$$QoS_s = \{q_{ij} \mid i \in [1,5], j \in [1,3]\} ,$$

in which q_{i1} 、 q_{i2} 、 q_{i3} stand for service price (currency unit), response time (millisecond) and reliability (percentage). Specific value will be offered in the following matrix:

$$QoS_s = \begin{pmatrix} 100 & 10 & 82 \\ 120 & 20 & 90 \\ 150 & 16 & 80 \\ 80 & 30 & 92 \\ 200 & 6 & 88 \end{pmatrix}$$

According to the above-mentioned method of QoS parameter expectation value, we firstly process the data by standard and get the following matrix of numbers:

$$QoS_s' = \begin{pmatrix} 0.15 & 0.12 & 0.26 \\ 0.18 & 0.24 & 0.15 \\ 0.23 & 0.20 & 0.29 \\ 0.12 & 0.37 & 0.12 \\ 0.31 & 0.07 & 0.18 \end{pmatrix}$$

Following that the QoS attributes single connected graph $G_1 = (V_1, E_1)$ of q_{i1} and weighted edge sets $E_1 = \{0.15, 0.18, 0.23, 0.12, 0.31\}$ are established. By Theorem 1 we can get that

$$|V_1| = \left| \sqrt{1 + 8|E_1|} / 2 + 1 \right| = 4 .$$

If the initial status is $TE_1 = \{\}$ then the status of the selected set of points U and edge set TE_1 are

$$U = \{v_0, v_3\}, TE_1 = \{0.12\} ;$$

$$U = \{v_0, v_3, v_1\}, TE_1 = \{0.12, 0.15\} ;$$

$$U = \{v_0, v_3, v_1, v_2\} ,$$

$$TE_1 = \{0.12, 0.15, 0.18\} .$$

Finally the first QoS expectation value $E(q_{i1}) = \sum_{i=1}^{n} q_{ij} \times \left(\sum_{k=1}^{|TE_1|} e_k \Big/ |TE_1| \right) \approx 98$ is obtained. Similarly, we can get

$$E(q_{i2}) = \sum_{i=1}^{n} q_{ij} \times \left(\sum_{k=1}^{|TE_2|} e_k \Big/ |TE_2| \right) \approx 11 ,$$

$$E(q_{i3}) = 1 - \sum_{k=1}^{|TE_3|} e_k \Big/ |TE_3| \approx 0.85 .$$

Thus we can get QoS expectation value

$$E(QoS_s) = (98, 11, 0.85)$$

From the above example, it is obvious that the calculation method has disposed of some QoS parameters and provided equilibrium combination value of low dissipation. It actually has reduced the range of service selection, the load of calculation in matchmaking and clarified the target of matchmaking, which makes the process of service selection more precise and accurate. The time complexity of Prim algorithm is $O(n^2)$, whereas that of Kruscal algorithm is

$O(e(\log_2 e))$. The time complexity of the former seems worse than the latter but each has its own advantage. Although the calculation of QoS parameter based on Prim algorithm has no advantage in time complexity, it is more appropriate for the calculation of minimum spanning tree of dense graph, which is very similar to the QoS parameter model. In contrast, the calculation of QoS expectation parameter value based on Kruscal algorithm is more suitable for spares graph. In fact, the results of the two methods of calculation are very close to each other, although it has strengths in time complexity.

V CONCLUSION

From the perspective of selecting service by QoS attributes, a computation method of QoS expectation parameter value based on Algorithm Prim is presented in order to provide support for selection of service, which is beneficial to the optimization of resource consumption and the protection of customers' efficiency in use. The achievement expressed this article provides a useful perspective and method for selction of service and QoS gurantee and therefore bears significant value in both theory and practice.At the next stage our research will concentrate on combining the method of computation of QoS expectation parameters value with the effective selection of service, testing and assessing its efficiency correctly.

Service foundation is a process which could meet the need of specific service of user in the network, and achieve automation and intellectualization. There is not strict divide between service foundation and service selection, in some research work, service foundation includes service selection. Generally speaking, service foundation emphasizes the process in founding candidate service collection, namely the way on gaining candidate service, but service selection emphasizes selecting a suitable service for user from candidate service collection. In this sense, service foundation is the preorder step of service selection, as a roughing process, the result collection is the object of service selection operating. The size of result collection; the way of gaining; and veracity have direct effect on service selection strategy. If it adopts a very strict standard for the need of all users in service foundation, service selection has to do nothing, and vice versa.

Because of basic mechanism sustained by QOS attribute, it can configure, discover, select, distributes resource on the basic of QOS attribute. In current many system, not only grid system, but also distributed system and Peer-to-Peer system, all its introduce SLA mechanism, which can describe QOS information resource and bind specific application. Some researcher introduces Service data into grid service, which can be used to describe a kind of grid service information including QOS information. G-QOSM base on OGSA, provides a QOS management model facing service, and expand grid service description on the foundation of service data. It sustains resource based on QOS attribute and service foundation, also the latest GGF standard, and match OGSA' latest standard. QGS in G-QOSM frame exist in every domain, keeping in touch with user application program, and catch service request constrained by QOS parameter. According to the given parameter, it can find the best matching service and consult SLA; Base on foundation sign a contract to guarantee user service quality.

With the development of computer science, graph theory progress at an alarming rate, and it is a major embranchment in applied mathematics. On one hand, computer science provides computing equipment for graph theory; on the other hand, it needs graph theory to describe and solve many problems in modern sciences practical application. Graph theory was applied to many domains as a method or tool in describing the relation of affairs at present, such as computer science, physics, chemical, operational research, information theory, cybernetics, network communication, social science, economic management, military, national defense, and agriculture and industry production. Prim is an important method to solve the weighted graph shortest or the optimal path problem in graph theory, and then it can be used to project decision described by graph theory.

The discovery and selection of service based on QoS attributes can facilitate the optimization of system resources and guarantee the quality of customer service, which has been a hot research topic in grid computing. Moreover, it is also an issue to be sorted out for the application and commercialization of grid computing. In the commercialized environment of service-oriented grid application, the users will consider their own benefit and efficiency while using the service. Whereas among a number of candidate services, the way users determine the equilibrium requirements of QoS appears critical as equilibrium requirements of QoS have a direct impact on QoS matchmaking parameters and the selection of services. Therefore, it is essential for users to present the expected value of QoS parameters and method of computing.

ACKNOWLEDGEMENT

Authors gratefully acknowledge the Projects Supported by Scientific Research Fund of Hunan Provincial Education Department(09C271, 08A009 and 08B015) for supporting this research.

Project supported by Provincial Natural Science Foundation of Hunan(10JJ6099)supports the research.

Project supported by Provincial Science & Technology plan project of Hunan (2010GK3048) supports the research.

This research is supported by the construct program of the key discipline in Hunan province.

This work was supported by the National Natural Science Foundation of China (51075138)

REFERENCES

[1] LIANG Quan YANG Yang LIANG Kai-jian. Guarantee and control of quality of service on grid system: A survey. Control and Decision, 2007, 22(2): 121~126.

[2] Charles Kubicek. Applying a Stochastic Model to a Dynamic, QoS Enabled Web Services Hosting Environment. Electronic Notes in Theoretical Computer Science, 2006, 151(3): 77-95.

[3] Junseok Hwang, Martin B.H. Weiss. Service differentiation economic models and analysis of market-based QoS interconnections. Telematics and Informatics, 2008, 25(4): 262-279.

[4] Sanya Tangpongprasit, Takahiro Katagiri, Kenji Kise, et al. A time-to-live based reservation algorithm on fully decentralized resource discovery in Grid computing. Parallel Computing, 2005, 31(6): 529-543.
[5] Haibin Cai, Xiaohui Hu, Qingchong Lü, et al. A novel intelligent service selection algorithm and application for ubiquitous web services environment. Expert Systems with Applications, 2009, 36(2): 2200-2212.
[6] Liu Shi-dong Zhang Shun-yi Qiu Gong-an Sun Yan-fei, An Improved Path Performance Parameter Estimation Technique Based on End-to-End Measurements, Journal of Electronics and Information Technology, 2007, 29(7): 1618-1621.
[7] WAN Jun DOU Wen-Hua LUO Jian-Shu CHEN Ying-Wu, Discrete Wavelet Spectrum's Characherization and Its Parameter Estimating for Multifractal Network Traffic, Chinese Journa of Computers , 2007, 30(1): 18-26.
[8] ZHU Jianming, WANG Yuhong, SUN Baowen, Development Requirement and Efficiency of E-Government, an Analysis, Economic Science Press, Beijing, 2009.
[9] Rashid J. Al-Ali, Omer F. Rana, David W. Walker. G-QoSM: Grid Service Discovery Using QoS Properties. Computing and Informatics Journal, 2002, 21(4): 363–382.
[10] Shaikh Ali A, Rana O, Al-Ali R, et al. UDDIe: An extended registry for web services. Proceedings of Workshop on Service-Oriented Computing: Models, Architectures and Applications SAINT2003, Olando, USA: IEEE CS Press, 2003:1623~1632.
[11] WANG Xiaodong, Design and Analysis of Algorithm. Qing Hua University Press, Beijing, 2008
[12] Al-Ali R, Laszewski G, Amin K, et al. QoS support for high-performance scientific applications. In: Proceedings of the IEEE/ACM 4th International Symposium on Cluster Computing and the Grid, Chicago IL, 2004. Los Alamitos: IEEE Computer Society Press, 2004, pp. 134~143.
[13] von Borstel F D, Gordillo J L. Model-based development of virtual laboratories for robotics over the inter-net. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2010, 40 (3) :623-634 .
[14] J. MacLaren. Advance reservation: State of art. GGF GRAAP-WG, See Web Site at:http://www.fz-juelich.de/zam/RD/coop/ggf/fraap/graap-wg.html,Last visited: Feb 2005.
[15] Al-Ali R, Hafid A, Rana O, et al.A n Approach for QoS Adaptation in Service-Oriented Grids [J].Concurrency and Computation: Practice and Experience Journal,2004,16(5):401~412.
[16] Liang Quan. Study of Service-Oriented Grid Models, Strategies and Methods with QoS Guarantee. Beijing: Beijing University of Technology, 2008
[17] Bruno R. Preiss. Translated by Hu Guangbin, Wang Song, Hui Min, etc. Data Structures and Algorithms. Beijing: Electronic Industry Press, 2003.
[18] Awad M K,Xuemin Shen. OFDMAbased two-hop coopera-tive relay network resources allocation. IEEE ICC 08. USA: Institute of Electrical and Electronics Engineers, 2008, pp. 4414-4418.
[19] Shi Jinfa, Jiao Hejun, Sun Jianhui. Research on Collaborative Design System of small and medium-sized enterprises forNetworked Manufacturing. Proc.38th International Conference on Computers and Industrial Engineering. Beijing, China: 2008, pp. 2146-2153.
[20] F.Y. Zhu. Fractal description——A new analysis technique for information system. Journal of East China University of Science and Technology, Vol.14, pp.101~103, 1988.
[21] S.X. Qu. The Relation between Fractal Dimenasion and Entropy. Chinese Journal of High Pressure Physics, Vol.7 No.2, pp.127~132, 2011.
[22] Z.L. Yan, W.H. Qiu, Z.Q. Chen. Evaluation of System Order Degree as Viewed from Entropy. Systems Engineering——Theroy & Practice, No. 6, pp. 46~49, 1997.
[23] N. Cheng. Fractal and MIS. Modern Information, No.2, pp.37~39, 2003.
[24] B. Cheng, H. H. Hu, Z. Wu. Fractal Knowledge Chain Research in Knowledge Management. Modern Management Science, No. 9, pp. 58~60, 2005.
[25] J. Wu, S.F. Liu.Entropy Model of Enterprise Knowledge Metastasis. Statistics & Decision, No. 2, pp.141~143, 2007.
[26] X. B. Li. Entropy-Information Theory and an Analysis of the Effectiveness of Systems Engineering's Methodology. Systems Engineering Theory & Practice, No. 2, pp. 38~44, 2010.

Liang Kaijian, Male, Born in August 1965, in Dongkou, Hunan, PhD, the professor of Hunan Insitute of Engineering. research direction: intelligent technology, funded by Hunan Natural Science Joint Fund leader and current vice president of application technology, manufacturing information in Xiangtan City Group experts. Research Interests: knowledge discovery and intelligent technology. In recent years, chaired the participating countries from the provincial education department Corky gold key projects and scientific research 4; published more than 20 academic papers, which were retrieved included six three.

Linfeng Bai School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, China

Xilong Qu (1978-), PhD., the associate professor of Hunan Insitute of Engineering. He graduated from Southwest Jiaotong University in 2006 and earned the Doctor degree. His research interesting are networked manufacturing, agile supply chain, and papers with high quality, and more than 20 papers are indexed by ISTP and EI.

Web Page Classification using an Ensemble of Support Vector Machine Classifiers

Shaobo Zhong*
College of Elementary Education, Chongqing Normal University, Chongqing 400700, China
Email: zshaob@163.com

Dongsheng Zou
College of Computer Science, Chongqing University, Chongqing, 400044, China
Email: dszou@cqu.edu.cn

***Abstract*-Web Page Classification (WPC) is both an important and challenging topic in data mining. The knowledge of WPC can help users to obtain useable information from the huge internet dataset automatically and efficiently. Many efforts have been made to WPC. However, there is still room for improvement of current approaches. One particular challenge in training classifiers comes from the fact that the available dataset is usually unbalanced. Standard machine learning algorithms tend to be overwhelmed by the major class and ignore the minor one and thus lead to high false negative rate. In this paper, a novel approach for Web page classification was proposed to address this problem by using an ensemble of support vector machine classifiers to perform this work. Principal Component Analysis (PCA) is used for feature reduction and Independent Component Analysis (ICA) for feature selection. The experimental results indicate that the proposed approach outperforms other existing classifiers widely used in WPC.**

***Index-Terms*— Web Page Classification, Support Vector Machine, Ensemble Classifier.**

I. INTRODUCTION

With the rapid development of the World Wide Web, the mass of online text data has grown at very fast speed in recent years. Information retrieval is facing great challenge due to the explosion of the network scales. How to obtain useable information from the huge internet raw data automatically and efficiently becomes more and more important than any time before. Researchers have been actively studying on web mining with various data in the World Wide Web. They study various fields such as focused crawler, information extraction, opining mining, usage mining, information integration, social network analysis and so on. Search engines and Web directories are the essential attempts. Actually in each field, classification is one of the methods that organize the subject. Classification is a supervised method of grouping data in a way, that more similar elements come together in the same group, but clustering is an unsupervised method that can find hidden relations among data, which can be used to divide members of a class to even more related clusters. Usually classification is done according to some rules such as latent or obvious analogies among things which are studied. Finding existent pattern is a complicated procedure because these patterns are usually hidden and can not be seen obviously. Therefore, machine learning algorithms are needed for classification. This makes many researchers focus on the issue of WPC technology. WPC can deal with the unorganized data on the web. The purpose of WPC is to classify the Internet web pages into a certain number of pre-defined categories.

During the past two decades, many methods have been proposed for WPC, such as Naive Bayes (NB) classifier [1], self-organization neural networks [2], Support Vector Machine [3], etc. Recently some methods attempt to use some hybrid approach for WPC. For example, Weimin and Aixin [4] used body, title, heading and meta text as feature by using SVM and Naive Bayesian classifier. The result shows that combination of these features with SVM classifier gives higher efficiency for web page classification system. Xin Jin et al. [5] used ReliefF, Information Gain, Gain ratio and Chi Square as feature selection technique for improving the web page classification performance. Rung-Ching and Chung-Hsun [6] proposed a web page classification method by using two types of features as inputs to SVM classification. The output of two SVM is used as inputs of voting schema to determine the category of the web page. The voting improves the performance when compares with the traditional methods. Fang et al. [7] proposed a web page classification by using five classification methods. The output of these SVMs is used as inputs of voting

Manuscript received July 1, 2010; revised January 1, 2011; accepted January 22, 2011.

This work was funded by the Key Project of Chinese Ministry of Science and Technology (No. 2008ZX07315-001), Major scientific and technological special project of Chongqing (No.2008AB5038).

*corresponding author: Shaobo Zhong.

doi:10.4304/jnw.6.11.1625-1630

method and picks the class with the most votes as the final classification result. This method improves the performance when compared with the individual classifiers. Zhang et al. [8] presented a web page categorization based on a least square support vector machine (LS-SVM) with latent semantic analysis (LSA). LSA uses Singular Value Decom-postion (SVD) to obtain latent semantic structure of original term-document matrix solving the polysemous and synonymous keywords problem. LS-SVM is an effective method for learning the classification knowledge from massive data, especially on condition of high cost in getting labeled classical examples. The F-value is 98.2% by using LS-SVM method. Moayed et al. [9] used a swarm intelligence algorithm in the filed of WPC by focusing on Persian web pages. Ant Miner II is the used algorithm. The highest accuracy for News site 1 is 89%. Hossaini et al. [10] used Genetic Algorithm (GA) for classification and clustering. The algorithm works on variable size vectors. At the GA part they combined standard crossover and mutation operators with K-means algorithm for improving diversity and correctness of results. By means of this method they achieved more accurate classes and defined subclasses as clusters. Their method shows more accurate results than fixed size methods. The accuracy rate is about 90.7% and also overload of unnecessary elements in vectors is bypassed.

He et al. [11] used an approach using Naive Bayes (NB) classifier based on Independent Component Analysis (ICA) for WPC. Some other researchers also addressed this problem [13-22]. However, there is significant room for improvement of current approaches.

One particular challenge in training classifiers comes from the fact that the dataset used for WPC is unbalanced [12] to some extent. The number of one kind of web pages can be much smaller or greater than another. Standard machine learning algorithms without considering class-imbalance tend to be overwhelmed by the major class and ignore the minor one and lead to high false negative rate by predicting the positive point as the negative one[23]. However, the accurate classification of web page from the minority class is equivalently important as others. In order to overcome this disadvantage, a common approach is to change the distribution of positive and negative sites during training by randomly selecting a subset of the training data for the majority class. But this approach fails to utilize all of the information available in the training data extracted from the original web pages.

In this paper, a novel approach for WPC is proposed. Our approach uses an ensemble classifier to deal with WPC. The novel approach implements an ensemble of SVM classifiers trained on the ‘‘natural’’ distribution of the data extracted from the original web pages. The ensemble classifier can reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. In addition, PCA algorithm is used for feature reduction and ICA algorithm for feature selection. The experimental results indicate that the proposed approach was indeed providing satisfactory accuracy in web page classification.

This paper is organized as follows. Section II focuses on the method. Section III describes the experiments. The conclusion and future work are discussed in Section IV.

II. METHODS

The process of WPC consists of web page retrieval processing, stemming, stop-word filtering, the weight of regular words calculating, feature reduction and selection, and finally the document classification using ensemble classifier. In web page retrieval phase, we will also retrieval the latest news web pages category form the Yahoo.com, and store them in our local databases according to Ref. [4]. In this way out research work can be compared with previous efforts.

A. Web Page Representation

It is difficult to carry on the WPC directly because the words in web documents are huge and complex. In this paper, we extract character words constitutes eigenvector with Vector Space Model (VSM), which is considered as one of most popular model for representing the feature of text contents. In this model, each document is tokenized with a stop-word remover and Porter stemming [24] in order to get feature words used as Eigen values. Finally the documents are projected to an eigenvector, as follow:

$$V(d) = (t_1, w_1(d); t_2, w_2(d); \cdots, t_n, w_n(d)), \quad (1)$$

Where t_i denotes the i-th keyword and $w_i(d)$ is the weight of t_i in document d.

B. Weight calculation

One obvious feature that appears in HTML documents but not in plain text documents is HTML tags. The information derived from different tags bear different importance. For example, a word present in the TITLE element is generally more representative of the document’s content than a word present in the BODY element. So, according to the HTML tags in which the terms are included in, we defined a new method of weight calculation as follows:

$$W_j(d) = \frac{1}{2}\left[\left(W_j(t, \tilde{d})\right) + \left(\Psi(t_j, d_i)\right)\right] \quad (2)$$

where $W(t, \tilde{d})$ is the weight of t in document $\tilde{d}$ according to frequency of words appeared in the HTML documents.

$$W_j(t,\tilde{d}) = \frac{tf(t,\tilde{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{N \in d} \left[tf(t,\tilde{d}) \times \log(N/n_t + 0.01) \right]^2}} \quad (3)$$

where, $tf(t,\tilde{d})$ is the frequency of t in document $\tilde{d}$. N is the number of total documents. And n_t is the number as documents in which i -th keyword appears. $\Psi(t_j, d_i)$ is the location of the words appeared in the HMTL document as following functions

$$\Psi(t_j, d_i) = \sum_{e_k} \left(\partial(e_k) \cdot TF(t_j, e_k, d_i) \right) \quad (4)$$

Where e_k is an HTML element, $\partial(e_k)$ denotes the weight assigned to the element e_k and $TF(t_j, e_k, d_i)$ denotes the number of times term t_j is present in the element e_k of HTML page d_i . We define the function $\partial(e_k)$ as:

$$\partial(e) = \begin{cases} \alpha, & if\ e\ is\ META or\ TITLE \\ 1. & elsewhere \end{cases} \quad (5)$$

where, $\alpha = 2,3,4,5,6$ were tested and compared with standard $TF(t_j, d_i)$ .The experimental results showed that using ensemble classifier can obtain the best results while the value of α equals 6.

C. Feature reduction

The method presenting feature words will generally create multidimensional datasets. PCA is certainly the most widely used method for multivariate statistical analysis. It reduces data dimensionality by performing a covariance analysis between factors. As such, it is suitable for datasets in multiple dimensions. The efficiency of the filter approach of PCA is relatively high. According to the different processing manners, PCA can be divided into data method and matrix method. We choose matrix method, and represent the training sample in the form of document-lemma matrix $\Re = (w_{ij})_{m \times n}$, where covariance is the weight of terms existing in the set of documents. All data which calculated the variance and covariance are represented in matrix. Then, get the eigenvectors of the covariance matrix, which are corresponding to the main component of the original data. We selected the first-used eigenvectors $\xi \le m$, the ξ herein, as eigenvectors is 100, 200, 400, etc. The principal components set is $n \times \xi$ matrix $M = (\ell_{ij})_{n \times \xi}$, where ℓ_{ij} is the eigenvectors being extracted out of the reduced state from original data size $m \times n$ to data size $n \times \xi$. The complete analysis of the PCA method used in this paper is given in Ref. 25 and Ref. 26.

D. Feature selection

Independent Component Analysis (ICA) [27] is a novel statistical signal and data analysis method. The purpose of ICA is to linearly transform the original data into components which are as much as statistically independent [28]. The task of ICA is to find Separation matrix W to make $y = Wx$ where $y = (y_1, y_2, \cdots, y_N)^T$ is called output variable, and $x = (x_1, x_2, \cdots, x_N)^T$ is an observed random variable. If y_i is mutually independent, then y_i is the estimated value of an independent random variable $s = (s_1, s_2, \cdots, s_N)^T$. It can be seen as an extension of PCA towards higher order dependencies.

E. An ensemble of SVM classifiers

Support vector machine

Support vector machine (SVM)classifier, motivated by results of statistical learning theory[29][30], is one of the most effective machine learning algorithms for many complex binary classification problems .Given the training set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l) \in (X \times Y)^l\}$ when the penalty factor C and kernel function $K(.,.)$ are selected properly, we can construct a function

$$g(x) = \sum_{i:x \in X_+} \alpha_i K(x, x_i) - \sum_{i:x \in X_-} \alpha_i K(x, x_i) + b, \quad (6)$$

where the non-negative weights α_i and b are computed during training by solving a convex quadratic programming. In order to estimate the probability of an unlabeled input x belonging to the positive class, $P(y = 1 \mid x)$, we map the value $g(x)$ to the probability by (Platt, 1999)

$$\Pr(y=1 \mid x) = P_{A,B}(g(x)) = 1/[1 + \exp(A * g(x) + B)] \quad (7)$$

Where A and B are then obtained by solving the optimization problem

$$\min_{z=(A,B)} F(z) = -\sum_{i=1}^{l} (t_i \log(p_i) + (1-t_i)\log(1-p_i))$$
$$s.t. \quad ti = \begin{cases} (N_+ + 1)/(N_+ + 2) & if\ y_i = +1, \\ 1/(N_- + 2) & if\ y_i = -1, \end{cases} \qquad (8)$$
$$pi = P_{A,B}(g(xi)),\ i = 1,2,\cdots l$$

Where N_+ and N_-, respectively, represent the number of positive and negative points in training set. Then the label of the new input x is assigned to be positive if the posterior probability is greater than a threshold, otherwise negative, i.e.

$$f(x) = \begin{cases} 1, & if\ \Pr(y=1|x) > threshold \\ -1. & otherwise \end{cases} \qquad (9)$$

where 1 corresponds to positive class, whereas -1 corresponds to negative class.

An ensemble of SVM classifiers

An ensemble of SVM classifiers is a collection of SVM classifiers, each trained on a subset of the training set (obtained by sampling from the entire training points) in order to get better results [31]. The prediction of the ensemble of SVMs is computed from the prediction of the individual SVM classifier, that is, during classification, for a new unlabeled input x_{test}, the j-th SVM classifier in the collection returns a probability $P_j(y=1|x_{test})$ of x_{test} belonging to the positive class, where $j = 1,2,\cdots m$ and m is the number of SVM classifiers in the collection. The ensemble estimated probability, $P_{Ens}(y=1|x_{test})$, is obtained by

$$P_{Ens}(y=1|x_{test}) = (1/m) \times \sum_{j=1}^{j=m} P_j(y=1|x_{test}) \quad (10)$$

Fig.1 shows the architecture of the ensemble of SVM classifiers.

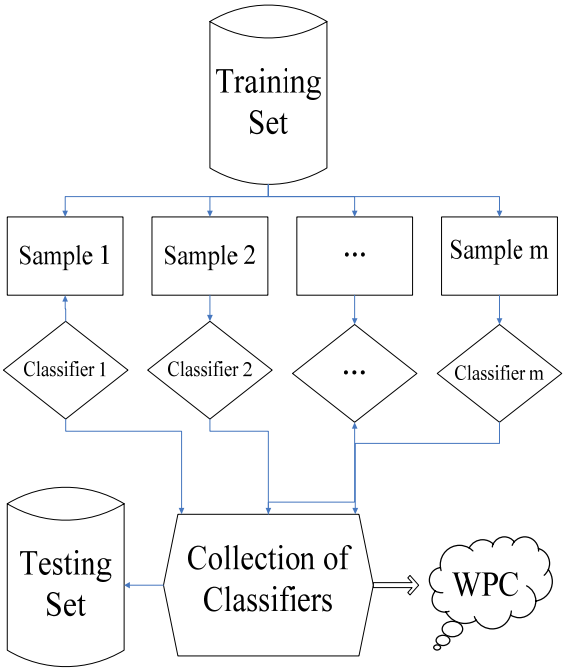


Figure 1. Architecture of the ensemble classifier fusing m SVM classifiers. Each one is trained on a balanced subsample of the training data.

III. EXPERIMENTAL RESULTS AND DISCUSSION

For experimental purpose, we build the dataset in the similar way as He et al. [11]. We choose the web page dataset from the Yahoo sports news. The dataset includes six categories of web pages. They are Soccer, NBA, Golf, Tennis, Boxing and NFL. The whole set include 3,160 web pages, i.e.880 documents of Soccer, 560 documents of NBA, 320 documents of Golf, 640 documents of Tennis,280 documents of Boxing, 480 documents of NFL. Among the dataset, 2500 documents (about 80%) selected randomly from different classes were used for training data, and the remaining other document for test data.

As for performance measure, the standard information retrieval measures, such as recall (r), precision (p), and F1 ($F1 = 2rp/(r+p)$) are used to estimate the performance of our method. To compare with other approaches, we have done the classification on the same dataset by using TFIDF, NB classifier and He's improved NB (denoted as NBICA)[11] .

The experimental results of WPC on our dataset are shown in Table 1. For the category of Soccer, NBA, Golf, Tennis, Boxing and NFL, the value of F1 are 91.55%, 92.97%, 94.40%, 92.50%, 94.55% and 93.87%, respectively. Meanwhile, the overall average of F1 measure is 93.31%. Comparing with NBICA, the overall F1 value is increased modestly from 92.13 to 93.31% by

using our approach. In addition, the F1 value for each category is relatively stable with our approach. However, the lowest F1 value is 75.85% for Soccer category while the highest one is 98.81% for rugby category with NBICA. The F1 value varies evidently because the sizes for each category of web pages are unbalanced with NBICA. As observed from Table1, we can summarize that this problem is solved with our approach by using an ensemble of SVM classifiers.

For comparison we used some other methods, such as TFIDF [32], NB and NBICA for WPC on the same dataset. The experimental results of WPC are shown in Table 2. By using TFIDF, NB and NBICA methods, the overall average F1 value are 81.78, 84.04 and 89.63%, respectively. Our method of ensemble classifier improves F1 by 3-11%. These results indicate the superior performance of our approach over that of some existing methods for WPC.

TABLE 1.

EXPERIMENTAL RESULT USING ENSEMBLE CLASSIFIER.

Class No.	Recall (%)	Precision (%)	F1 (%)
1.Soccer	90.36	92.78	91.55
2.NBA	95.66	90.42	92.97
3.Golf	96.27	92.6	94.40
4.Tennis	94.50	90.58	92.50
5.Boxing	95.68	93.45	94.55
6.NFL	95.45	92.35	93.87
Average	94.65	92.0	93.31

TABLE 2.

F1 VALUE BY USING DIFFERENT APPROACHES

Class No.	TFIDF (%)	NB (%)	NBICA (%)	Ensemble classifier (%)
1.Soccer	84.32	85.85	90.25	91.55
2.NBA	83.44	93.56	93.68	92.97
3.Golf	74.37	76.30	84.56	94.40
4.Tennis	85.60	85.81	93.56	92.50
5.Boxing	80.16	78.69	83.40	94.55
6.NFL	82.76	84.05	92.30	93.87
Average	81.78	84.04	89.63	93.31

IV. CONCLUSION

Automated web pages classification, which is a challenging research direction in text mining, plays an important role to establish the semantic web. Many efforts have been made for WPC. However, there is significant room for improvement of current approaches. One particular challenge in training classifiers comes from the fact that the dataset used for WPC is unbalanced to some extent. Consequently, the F1 value of most existing methods is unstable. In this article, we have studied the problem of unbalanced dataset in WPC. We proposed a novel approach using an ensemble of SVM classifiers to address this problem. The comparison of performance among four methods, namely TFIDF, NB, NBICA and our ensemble classifier, has been presented in this paper. The experimental results indicate that the proposed approach could solve the problem well. Moreover, the F1 value is increased modestly with our approach.

In future research, we should address to increase the number of categories to a large extent to observe the F1 value with our approach. Moreover, combined with some existing algorithms, such as Genetic algorithm, our method of ensemble classifier can be further improved.

ACKNOWLEDGMENTS

The authors thank the editor and referees for their careful review and valuable critical comments. We also thank Prof. He for valuable suggestions and comments. This work is supported by the Key Project of Chinese Ministry of Science and Technology (No. 2008ZX07315-001), Major scientific and technological special project of Chongqing (No.2008AB5038) , Education project of Chongqing Normal University(080201), The Chongqing Key Research Base of Humanities and Social Sciences: the Financial Support from Chongqing Research Center of Elementary Teacher Education. The authors are grateful for reviewers who made constructive comments.

REFERENCES

[1] Fan Y., Zheng C., Wang Q. Y., Cai Q. S., Liu J. Web Page Classification Based on Naive Bayes Method (In Chinese), Journal of Software, 2001, pp. 1386-1392.

[2] Zhang Y. Z. The Automatic Classification of Web Pages Based on Neural Networks. Neural information processing, ICONIP2001 Proceedings, Shanghai, China, 14-18 November 2001, Vol.2, pp. 570- 575.

[3] Xue W. M., Bao H., Huang W. T., Lu Y. C. Web Page Classification Based on SVM. Intelligent Control and Automation, 21-23 June 2006, vol.2, pp. 6111- 6114.

[4] W. Xue, H. Bao, W. Huan, and Y. Lu, "Web Page Classification Based on SVM," 6th World Congress on Intelligent Control and Automation, Dalian, China, 2006, pp. 6111-6114,.

[5] J. Xin, L. Rongyan, S. Xian, and B. Rongfang, "Automatic Web Pages Categorization with ReliefF and Hidden Naive Bayes," Proceedings of the 2007 ACM symposium on Applied computing, Seoul, Korea, 2007, pp. 617-621.

[6]Chen R., Hsieh C., and Chen H. Web Page Classification

Based On A Support Vector Machine Using A Weighted Vote Schema. Expert Systems with Applications, 2006, vol. 31, pp. 427-435.
[7] Rui F., Alexander M., and Babis T. A Voting Method for the classification of Web Pages. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006, pp. 610-613.
[8]Zhang Y., Fan B., Xiao L. B. Web Page Classification Based-on A Least Square Support Vector Machine with Latent Semantic Analysis. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 528-532.
[9] Moayed M. J.; Sabery, A. H.; Khanteymoory, A. Ant Colony algorithm for Web Page Classification. 2008 International Symposium on information technology Kuala Lumpur, Malaysia, 26-29 August 2008, pp. 8-13.
[10]. Hossaini, Z Rahmani, A. M. Setayeshi, S. Web pages classification and clustering by means of genetic algorithm: a variable size page representing approach. 2008 International conference on Computational Intelligence for Modeling Control & Automation (CIMCA 2008), 10-12 December 2008, pp. 436-440.
[11] He Z. L., Liu Z. J. A Novel Approach to Naïve Bayes Web Page Automatic Classification. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 361-365.
[12] Japkowicz N. The class imbalance problem: significance and strategies. In: IC-AI'2000, Special Track on Inductive Learning Las Vegas, Nevada, 2000.
[13] Xu S.M., Wu B.,Ma C.. Efflcient SVM Chinese Web page classifier based on pre-classification. Computer Engineering and Applications, 2010, pp. 125-128.
[14] Araujo L.,Martinez R.J. Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models. IEEE Transaction on information forensics and security, 2010, Vol. 5 (3), pp. 581-590.
[15] Chen T.C., Dick, S., Miller, J. Detecting Visually Similar Web Pages: Application to Phasing Detection. ACM transaction on Internet techonology. 2010, Vol.10 (2), pp. 5
[16] Ofuonye E., Beatty P., Dick S.. Prevalence and classification of web page defects. Online Information Review, 2010, Vol. 34 (1), pp.160-174.
[17] Golub K., Lykke M. Automated classification of web pages in hierarchical browsing. Journal of documentation, 2009, Vol. 65 (6), pp. 901-925.
[18] Hou C.Q., Jiao L.C. Graph based Co-training algorithm for web page classification. Acta Electronica Sinica, 2009, pp.2173-80.
[19] Farhoodi, M., Yari A., Mahmoudi M. A Persian Web Page Classifier Applying a Combination of Content-Based and Context-Based Features.International Journal of Information Studies, 2009,pp.263-71.
[20] Selamat A., Subroto I.M.I., Choon C. Arabic script web page language identification using hybrid-KNN method.International Journal of Computational Intelligence and Applications,2009,pp.315-43.
[21] Zhu Z.G., Deng C.S., Kong L.P. Algorithm research on classifying Web users navigation patterns based on N-gram .Journal of the China Society for Scientific and Technical Information,2009,pp.389-394.
[22] Peng X.G., Ming Z., Wang H.T. WordNet based Web page classification system with category expansion. Journal of Shenzhen University Science & Engineering, 2009, pp.118-122.
[23] Liu, X. Y., Zhou, Z. H.. The influence of class imbalance on cost-sensitive learning: an empirical study. In: Sixth IEEE International Conference on Data Mining (ICDM'06), Hong Kong, 2006.
[24] The Porter Stemming algorithm, http://www. tartarus.org/~martin/PorterStemmer.
[25] Calvo R. A., Partridge M., Jabri M.. A comparative study of principal components analysis techniques. In Proceedings 9th Australian Conference on Neural Networks, Brisbane, QLD1998, pp. 276-281.
[26] Selamat, A., Omatu, S. Neural Networks for Web News Classification Based on PCA. Proceedings of the International Joint Conference, 20-24 July 2003, vol. 3, pp. 1792 - 1797.
[27] Hyvarinen A., Karhunen J., and Oja E., 2001. Independent Component Analysis, Wiley-Interscience, New York.
[28] Nacim F. C., Bernard R., Nathalie A.G. A Comparison of Dimensionality Reduction Techniques for Web Structure Mining. IEEE/WIC/ACM International Conference on, 2-5 Nov. 2007, pp. 116 – 119.
[29] Vapnik V., 1995. The Nature of Statistical Learning Theory. Springer, New York.
[30] Vapnik V., 1998. Statistical Learning Theory. Wiley, New York.
[31] Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Lecture Notes in Computer Science, vol. 1857, pp. 1–15.
[32] Yang J. P., Honavar V., Miler L. Mobile intelligent agents for document classification and retrieval: a machine learning approach. Proceeding of the Eurpoean Symposium on cybemetics and Systems Research, Vienna, Austria, 1998, pp.707-712.

Shaobo Zhong was born in Sichuan, P.R. China, in January 24, 1973. He obtained the bachelor's the master's degree in Mathematics and Computer Science of the Chongqing Normal University, China in 1998, and the doctor's degree in College of Computer Science of the Chongqing University, China in 2008.His research interest includes machine learning, data mining and web page classification.

Integration of Unascertained Method with Neural Networks and Its Application

Huawang Shi
Hebei University of Engineering, Handan, P. R. China
stone21st@163.com

***Abstract*—This paper presents the adoption of artificial neural network (ANN) model and Unascertained system to assist decision-makers in forecasting the early warning of financial in China. Artificial neural network (ANN) has outstanding characteristics in machine learning, fault, tolerant, parallel reasoning and processing nonlinear problem abilities. Unascertained system that imitates the human brain's thinking logical is a kind of mathematical tools used to deal with imprecise and uncertain knowledge. Integrating unascertained method with neural network technology, the reasoning process of network coding can be tracked, and the output of the network can be given a physical explanation. Application case shows that combines unascertained systems with feedforward artificial neural networks can obtain more reasonable and more advantage of nonlinear mapping that can handle more complete type of data.**

***Index Terms*—artificial neural network, unascertained system, financial early warning**

I. INTRODUCTION

Unascertained system that imitates the human brain's thinking logical is a kind of mathematical tools used to deal with imprecise and uncertain knowledge. Artificial neural network that imitates the function of human neurons may function as a general estimator, mapping the relationship between input and output. Combination of these two methods can take into account the effect of complementary effect of each other? Our theoretical analyses are the following aspects: First, the artificial neural network is a nonlinear mapping from input to output; it does not rely on any mathematical model. Unascertained system also as a nonlinear mapping is to convert input signals x in domain U into signal y in domain V as output. Second, artificial neural networks can only deal with explicit data classification, and not suitable for the expression of a rule-based knowledge. However unascertained systems can handle abnormal, incomplete and uncertain data. Third, the artificial neural network's knowledge representation and treatment are simple in form, and hard to the introduction of heuristic knowledge, and the lower efficiency of the network. Unascertained system can make use of expertise knowledge, thus be easy to introduce of heuristic knowledge that making the reasoning process more reasonable. Finally, artificial neural network's greatest strength are memory, learning and inductive functions; Unascertained system does not have the learning function.

So, in theory, combining unascertained systems with feed forward artificial neural networks can obtain more reasonable and more advantage of nonlinear mapping that can handle more complete and comprehensive type of data.

The rest of this paper is organized as follows: Unascertained Number and Algorithm are described in Section2. Section3 describes Unascertained BP Neural Networks in detail and gives Network Learning Process. The experimental results on Unascertained BP Neural Networks and some discussions are presented in Section4. Finally, Section5 provides the conclusion.

II. MATERIALS AND METHODS

A. Introduction to Unascertained Number:

1) Definition of Unascertained number:

Unascertained mathematics, proposed by Want [1], is a tool to describe subjective uncertainty quantitatively. It deals mainly with unascertained information, which differs from stochastic information, fuzzy information, and grey information. Unascertained information refers to the information demanded by decision-making over which the message itself has no uncertainty but, because of situation constraints, the decision-make cannot grasp the whole information needed. Hence, all systems containing the behavior factors, such as the problem of clustering have unascertained property.

Definition 1: Suppose a is arbitrary real number, $0 < \alpha \le 1$, then definite $[[a,a],\varphi(x)]$ is first-order unascertained number, where

$$\varphi(x)=\begin{cases}\alpha, x=a\\ 0, x\neq a \cup x\in R\end{cases} \tag{1}$$

Note that $[a,a]$ express the interval of value, and $\varphi(x)=\alpha$ express belief degree of a. When α =1, belief degree of a is 1. Where α =0, belief degree of a is zero.

Definition 1: Suppose $[a,b]$ is arbitrary closed interval, $a=x_1<x_2<\cdots<x_n=b$, if

$$\varphi(x)=\begin{cases}\alpha_i, x=x_i (i=1,2,\cdots,n)\\ 0, other\end{cases} \tag{2}$$

and $\sum_{i=1}^{n}\alpha_i=\alpha$, $0<\alpha\le 1$, then $[a,b]$ and $\varphi(x)$ compose a n -order unascertained number, as follow $[[a,b],\varphi(x)]$,

doi:10.4304/jnw.6.11.1631-1638

where α is total degree belief, $[a,b]$ is the interval of value, is $\varphi(x)$the density function.

Definition 1: Suppose unascertained number is $A=[[x_1,x_2],\varphi(x)]$, where

$$\varphi(x)=\begin{cases}\alpha_i, x=x_i(i=1,2,\cdots,k)\\ 0, other\end{cases} \tag{3}$$

$0<\alpha_i<1, i=1,2,\ldots,k$, $\alpha=\sum_{i=1}^{k}\alpha_i\le i$.

Then first-order unascertained number :

$$E(A)=\left[\left[\frac{1}{\alpha}\sum_{i=1}^{k}x_i\alpha_i,\frac{1}{\alpha}\sum_{i=1}^{k}x_i\alpha_i\right],\varphi(x)\right],$$

$$\varphi(x)=\begin{cases}\alpha, x=\frac{1}{\alpha}\sum_{i=1}^{k}x_i\alpha_i\\ 0, other\end{cases} \tag{4}$$

It is expected value of unascertained number A. When $\alpha=1$, as $E(A)$, unascertained number A is discrete type random variable. When $\alpha<1$, $E(A)$ is first-order unascertained number. Where $\frac{1}{\alpha}\sum_{i=1}^{k}x_i\alpha_i$ as expected value of A that belief degree is α.

2) Algorithm of unascertained number:

Each unascertained number includes two parts of probable value and belief degree. So, unascertained number algorithm also includes two parts. Suppose unascertained numbers are A and B.Where

$$A=f(x)=\begin{cases}\alpha_i, x=x_i(i=1,2,\cdots,m)\\ 0, other\end{cases},$$

$$B=g(x)=\begin{cases}\beta_i, y=y_i(i=1,2,\cdots,n)\\ 0, other\end{cases} \tag{5}$$

$C=A\times B$ also is unascertained number. Probable value and belief degree of C is calculated as follows.

(1)Constituted multiply matrix of probable value of unascertained number A and B, where individual is probable value number series x_1, x_2,…, x_k and y_1, y_2,…, y_m as A and B, permute from little to big.

(2)Constituted multiply matrix of belief degree of unascertained number A and B, where individual is belief degree number series α_1, α_2,…, α_m and β_1, β_2,…, β_n are A, B.Suppose a_{ij} and b_{ij} individual is element of multiply matrix of probable value of A and B, here i is line of matrix, j is array of matrix. We called a_{ij} and b_{ij} as relevant position element.

(3) $\bar{x}_1$, $\bar{x}_2$,…, $\bar{x}_k$ result from multiply matrix of probable value of unascertained number A and B, which permute from little to big. And an equal element is one element of relevant position element in multiply matrix of belief degree. Suppose $\bar{r}_1$, $\bar{r}_2$,…, $\bar{r}_k$ is relevant position element permutation. Where

$$C=\varphi(x)=\begin{cases}\bar{r}_i, x=\bar{x}_i(i=1,2,\cdots k)\\ 0, other\end{cases} \tag{6}$$

Suppose $C=\varphi(x)$ is arithmetic product of unascertained number A and B. Where

$$C=A\times B=f(x)\times g(x)=\begin{cases}\bar{r}_i, x=\bar{x}_i(i=1,2,\cdots k)\\ 0, other\end{cases} \tag{7}$$

3) Unascertained membership:

Using Unascertained to describe "uncertain" or "unclear boundary" phenomenon, the key problem is that a reasonable Unascertained membership function. Despite the clear definition rules of the construction unascertained measure, the definition is non-structural in nature, and did not give a specific construction method. It still needs to be in accordance with the background knowledge in specific areas, known to the measured data and personal experience of decision-makers, etc.

Under normal circumstances, decision-makers do not know exactly state of membership function. At this point, the simplest and most reasonable method is by fitting line shape of membership function. A standard membership functions is in Figure 1.

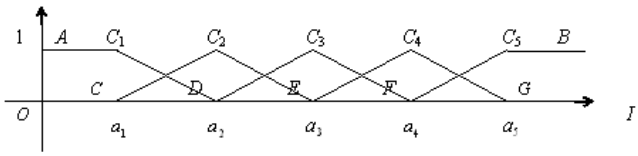


Figure 1. A standard membership functions curve

The class I_1 membership function $\varphi_1(x)$ was expressed by the broken line AC_1DI; The class I_2 membership function $\varphi_2(x)$ expressed by the broken line OCC_2EI; The class I_3 membership function $\varphi_3(x)$ was expressed by the broken line ODC_3FI; The class I_4 of membership function $\varphi_4(x)$ was expressed by the broken line OEC_4GI; The class I_5 of membership function $\varphi_5(x)$ was expressed by the broken line OFC_5B

B. Introduction to ANN

Artificial Neural Networks (ANNs) are composed of simple elements that imitate the biological nervous systems. In the last few decades, significant research has been reported in the field of ANNs and the proposed ANN architectures have proven the inefficiency in various applications in the field of engineering. The structure of a neural network of most commonly used type is schematically shown in Fig.1. It consists of several layers of processing units (also termed neurons, nodes). The input values are processed within the individual neurons of the input layer and then the output values of these neurons are forwarded to the neurons in the hidden layer. Each connection has an associated

parameter indicating the strength of this connection, these called weight.

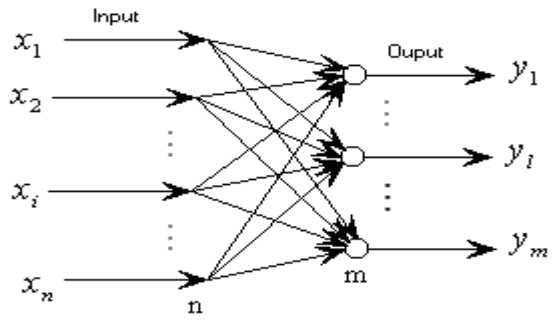


Figure 1. The single layer of feedforward networks.

The NN model frequently used is multilayer perceptron learning with error back-propagation. In the present research work, the sequence with which the input vectors occur for the ANN straining is not taken into account, thus they are static networks that propagate the values to the layers in a feed-forward way. The training of the neural networks is performed through a back-propagation algorithm. In General, the back-propagation algorithm is a gradient-descent algorithm in which the network weights are moved along the negative of the gradient of the performance function.

Artificial Neural Network (ANN) is basically as implied model of the biological neuron and uses an approach similar to human brain to make decisions and to arrive at conclusions[7]. Every neuron model consists of a processing element with synaptic input connections and a single output. The structure of a neural network of most commonly used type is schematically shown in figure 1.

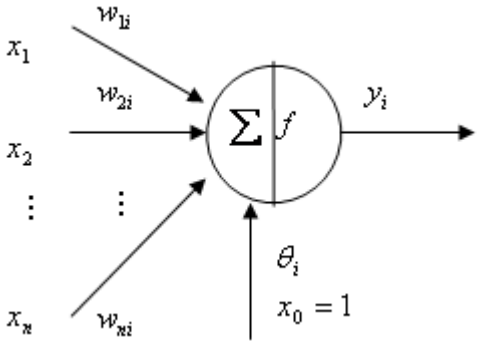


Figure. 2 Neural model.

The neuron can be defined as

$$y = f(W \times X + \theta_j) = f(\sum_{i=1}^{n} w_{ij} x_i - \theta_j)$$

where, x is input signals, ω_{ij} is synaptic weights of neuron, f is the activation function and y is the output signal of neuron. The architecture of multi-layered feedforward neural network is shown in Fig. 2.

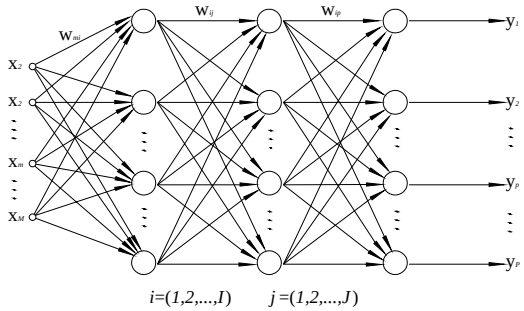


Figure 3. The model of BP net

It consists of one input layer, one output layer and hidden layer. It may have one or more hidden layers. All layers are fully connected and of the feedforward type. The outputs are nonlinear function of inputs, and are controlled by weights that are computed during learning process.

At present, the BP neural network is one of the most matures, wide spread artificial neural network. Its basic network is three-layer feed-forward neural network such as input layer, hidden layer, and output layer. The input signals must firstly disseminate for- ward into the hidden node. The output information of the concealment node transmits into output node Via- function action. Finally the output variable result is obtained. The BP network can realize complex non-linear mapping relations will fully from input to output and has good exuding ability, which can complete the duty of complex pattem recognition.

ANN has outstanding characteristics in machine learning, fault, tolerant, parallel reasoning and processing nonlinear problem abilities. It offers significant support in terms of organizing, classifying, and summarizing data. It also helps to discern patterns among input data, requires few ones, and achieves a high degree of prediction accuracy. These characteristics make neural network technology a potentially promising alternative tool for recognition, classification, and forecasting in the area of construction, in terms of accuracy, adaptability, robustness, effectiveness, and efficiency. Therefore, cost application areas that require prediction could be implemented by ANN.

C. Unascertained BP Neural Network

1) Description of unascertained BP network:

Assuming there is N known samples, divided into K categories, X^k represents the k th sample space with the sample size for N_k, apparently: $\sum_{k=1}^{K} N_k = N$.

x_i^k represents i th sample ($1 \le i \le N_k$), so $X^k = \left\{x_1^{(k)}, \cdots, x_i^{(k)}, \cdots, x_{N_k}^{(k)}\right\}^T$. Suppose that each sample x_i^k has J characteristics (or indicators), the j th feature (or indicators) is $I_j, 1 \le j \le J$.

x_{ij}^k represents the observation value of sample x_i^k with reference to the j th characteristic (or indicator).

2) Unascertained BP neural network structure:

Unascertained BP neural network structure and the structure of BP neural network is basically the same seen in Fig.1

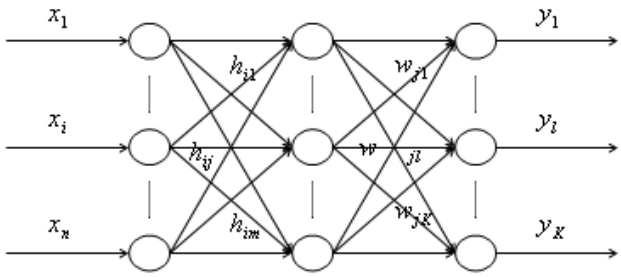


Figure 4. The unascertained BP network structure.

The first layer is input layer, the number of nodes is the same of feature space dimension. The second layer is hidden layer. The third layer is output layer; the layer number of nodes is equal to the classification number.

3) The desired membership calculating method:

In the usual artificial neural network, training samples were divided into specific categories, that is, a sample is determined belonging to a category. Therefore, training in the network, its corresponding output node of the desired output as "1", and the rest of the output node of the desired output for the "0". However, in practice, data are often sick, and its classification border is not very specific, and samples are belonging to categories in certain degree of membership. Therefore, the desired output is not simply a two-valued logic, need to calculate exactly, which leads to uncertainty in the network.

As the input data may be numerical value, also possible be the degree of membership, the corresponding desired output, there are differences in the calculation. The following discussion is made under numerical value input:

Supposing there are $N_k \left(\sum_{k=1}^{K} N_k = N \right)$ samples in k th category and category center is O_k:

$$O_k = (O_1^k, \cdots, O_j^k, \cdots, O_J^k)^T \ (1 \le j \le J, 1 \le k \le K) \tag{8}$$

Where, $O_j^k = \dfrac{1}{N_k} \sum_{i=1}^{N_k} x_{ij}^k$

$$\overline{O_j} = \frac{1}{K} \sum_{k=1}^{K} O_j^k = \left(\overline{O_1}, \overline{O_2}, \cdots, \overline{O_J} \right)^T \tag{9}$$

$$\sigma_j^2 = \frac{1}{K} \sum_{k=1}^{K} \left(O_j^k - \overline{O_j} \right)^2 \tag{10}$$

$$w_j = \frac{\sigma_j^2}{\sum_{j=1}^{J} \sigma_j^2} \tag{11}$$

Obviously, $0 \le w_j \le 1$,and $\sum_{j=1}^{J} w_j = 1$. Therefore, w_j is the indicator j 'classification weight of given classification.

Set $x_i = (x_{i1}, \cdots, x_{ij}, \cdots, x_{iJ})$ $(1 \le i \le N)$ as any training samples.

When the larger $\|x_i - O_k\|$, the farther sample x_i away from the center of k th category and its membership belonging to the k th category be smaller. On the other hand, when $\|x_i - O_k\|$ the smaller, the nearer sample x_i away from the center of k th category and its membership belonging to the k th category be larger.

When the larger w_j , the greater the contribution to classification of indicator I_j , that is, the more important to classification of indicator I_j On the other hand, when the smaller w_j , it shows that the smaller the contribution to classification of indicator I_j , that is, the less important for classification of indicators I_j . From the above, we can define the weighted distance of sample x_i to the k th class center O_k:

$$\gamma_{ik} = \sum_{j=1}^{J} w_j (x_{ij} - O_j^k)^2 \tag{12}$$

$$\mu_k(x_i) = \frac{1}{\gamma_{ik} + \varepsilon} \Big/ \sum_{k=1}^{K} \frac{1}{\gamma_{ik} + \varepsilon} \tag{13}$$

Obviously, $0 \le \mu_k(x_i) \le 1$, $\sum_{k=1}^{K} \mu_k(x_i) = 1$. Therefore, as $\mu_k(x_i)$ is unascertained membership of sample belonging to the k th category, that is, it is the expectations output of membership degree that we have to calculate: $d_k = \mu_k(x_i)$

4) Mathematical derivation of amendment w_{ij} :

Supposing that O_j represent output of the j th node, O_i express the output of i th node of the relative former layer and O_k express output of the k th node of the relative behind layer w_{ij} express connection weights of the upper layer node i to this node j :

$$net_j = \sum_i w_{ij} \cdot O_i$$

$$O_j = f(net_j) = \frac{1}{1+e^{-net_j}} \quad (14)$$

Where, net_j express the net input of nodes j.

- When O_j is the output of output layer nodes, the actual output $y_j = O_j$, Set d_j is the desired output of node j: $d_j = \mu^k(x_i)$, then squares sum error of output are as follows:

$$E_i = \frac{1}{2}\sum_j (y_j - d_j)^2 = \frac{1}{2}\sum_j (O_j - d_j)^2 \quad (15)$$

The total output error is:

$$E = \sum_i E_i \quad (16)$$

Considering amendments to the weights:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial O_j} \cdot \frac{\partial O_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}} \quad (17)$$

$$= (O_j - d_j) \cdot [O_j \cdot (1 - O_j)] \cdot O_i$$

$$\delta_j = (O_j - d_j) \cdot O_j \cdot (1 - O_j) \quad (18)$$

$$\frac{\partial E}{\partial w_{ij}} = \delta_j \cdot O_i$$

- When O_j is the output of hidden layer nodes, O_j affects each node of the lower classes.

Output square error:

$$E = \frac{1}{2}\sum_k (y_k - d_k)^2$$

The actual output of k th node of the output layer:

$$y_k = f(net_k) = \frac{1}{1+e^{-net_k}}$$

$$net_k = \sum_j w_{jk} \cdot O_j$$

$$O_j = f(net_j) = \frac{1}{1+e^{-net_j}}$$

$$net_j = \sum_i w_{ij} \cdot O_i$$

Considering amendments to the weights:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial net_k} \cdot \frac{\partial net_k}{\partial O_j} \cdot \frac{\partial O_j}{\partial net_j} \cdot \frac{\partial net_j}{\partial w_{ij}}$$

$$= \sum_k (y_k - d_k) \cdot y_k \cdot (1 - y_k) \cdot w_{jk} \cdot O_j \cdot (1 - O_j) \cdot O_i$$

Set $\delta_k = (y_k - d_k) \cdot y_k \cdot (1 - y_k)$

Then $$\frac{\partial E}{\partial w_{ij}} = \sum_k \delta_k \cdot w_{jk} \cdot O_j \cdot (1 - O_j) O_i$$

Set $$\delta_j = \sum_k \delta_k \cdot w_{jk} \cdot O_j \cdot (1 - O_j)$$

Then $$\frac{\partial E}{\partial w_{ij}} = \delta_j \cdot O_i .$$

5) Network learning process:

Set counter t, and $t = 0$, randomly generated initial values of weights $w_{ij}(t)$, set learning rate η, the system error ε as well as the impulse factor α, set the maximum number of iterations T.

Enter the study samples X, and calculate the desired output membership d_k of sample

Calculate the input value $net_j(t)$ of each node and output value $O_j(t)$:

$$net_j(t) = \sum_i w_{ij}(t) \cdot O_i(t)$$

,

$$O_j(t) = f(net_j(t)) = \frac{1}{1+e^{-net_j(t)}}$$

Calculate error $E(t)$,

$$E(t) = \frac{1}{2}\sum_k (y_k - d_k)^2$$

Stop criteria: If $E(t) \le \varepsilon$ or $t > T$, then stop. Otherwise, turn to (6);

Calculate the Adjustment value $\delta_j(t)$ of calculation errors.

Output layer:

$$\delta_j(t) = (O_j(t) - d_j) \cdot O_j(t) \cdot (1 - O_j(t))$$

Hidden layer:

$$\delta_j(t) = \sum_k \delta_k(t) \cdot w_{jk}(t) \cdot O_j(t) \cdot (1 - O_j(t))$$

Where, k express the lower node number to node j

Calculate the Adjustment value of weights

$$\Delta w_{ij}(t) = \alpha \Delta w_{ij}(t-1) + \eta \delta_j(t) \cdot O_i(t)$$

Revise weights:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t)$$

$t = t + 1$, Turn to (3).

6) Network identification:

Suppose x is the sample to be recognized. Input x into the trained network. Suppose the greatest output is of the k_0 output node, x belongs to the k_0 th category is determined.

$$k_0 = \max_k \{\mu^k(x) \mid k = 1,2,\cdots,K\} \quad (19)$$

D. Unascertained RBF Neural Networks

1) Structure of unascertained RBF network:

Unascertained RBF network consists of three layers such as input layer, hidden layer and output layer, which neurons in same layers has no connection, and between the adjacent two-layers has fully connected. Number of input layer neurons is the sample dimension; hidden layer

and output layer neuron number are the classification number of samples. Unascertained RBF network is characterized by only one hidden layer; hidden layer neurons nodes select the Gaussian function to have a non-linear mapping of input and output, layer neurons are linear combine node. Its structure is shown in Fig.2.

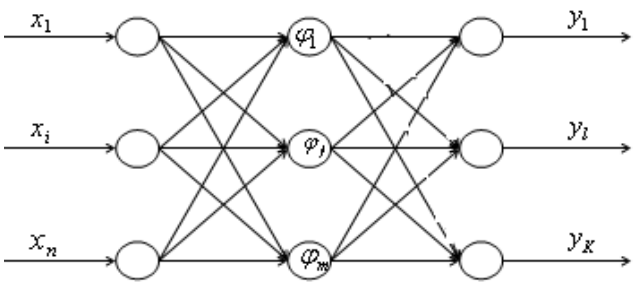


Figure 5. The unascertained RBF network structure.

Suppose the input sample was x, then the output of the i th hidden layer nodes was as follows:

$$\varphi_i = \exp\left(-\frac{\|x - m_i\|^2}{\nu_i^2}\right) \tag{20}$$

Where $\|*\|$ is European norm, m_i and ν_i were the centers and width of the i th hidden layer units of RBF.

The j th neuron actual output of output layer is:

$$y_j = \sum_{i=1}^{K} w_{ij} \cdot \varphi_i \tag{21}$$

$$\varphi_0 = 1$$

Compared with BP neural network, RBF network many has quicker convergence speed, because o close to m_i has a larger output value, far away from m_i, and its output decreases rapidly.

2) The desired membership calculating method:

Unascertained RBF neural network in the desired output method of calculating degree of membership.

Given known n samples, each of the known samples x_i are point of d dimensional feature space, that is: $x_i = (x_{i1,} x_{i2}, \cdots, x_{id})^T$

The n samples are divided into K categories: $C_1, C_2, \cdots, C_K$, m_k is the category center vector of C_k $(k = 1,2,\cdots,K)$. Considering the same type of sample point should be in-dimensional feature space with each other more "close" is reasonable. We have assumed that the "close" is the Euclidean distance proximity.

Supposing the i th training sample is $x_i = (x_{i1,} x_{i2}, \cdots, x_{id})^T$, the j th $(1 \le j \le d)$ data is x_{ij} that is the nominal quantity of data. Supposing m_k classified Center Vector of C_k:

$$m_k = (m_{k1}, m_{k2}, \cdots, m_{kd})^T \tag{22}$$

Unascertained classification in accordance with the point of view, give a classification, first of all concerned are in a given category, the characteristics of the classification of all make a little contribution, and contribution to the value of quantitative calculation. Hereinafter referred to as "normalized" after the classification of the characteristics of the contribution value of the characteristics regarding the classification of the classification weights. And, in the calculation of the sample about when various types of membership, in essence, to use a variety of characteristics of the weight classification.

In order to quantitatively describe the contributions of d characteristics to the initial classification.

Let $$\overline{m} = \frac{1}{C}\sum_{k=1}^{C} m_k \tag{23}$$

$$\overline{m} = (\overline{m_1}, \overline{m_2}, \cdots, \overline{m_d})$$

Let $\sigma_j^2 = \frac{1}{C}\sum_{k=1}^{C}(m_{kj} - \overline{m_j})^2, \quad 1 \le j \le d,$ (24)

The size of variance σ_j^2 reflects the extent of discrete the type of K centers as $m_1, m_2, \cdots, m_K$ in the first feature on values.

Let $$w_j = \sigma_j^2 \Big/ \sum_{j=1}^{d} \sigma_j^2 \tag{25}$$

Obviously, w_j satisfied : $0 \le w_j \le 1$ and $\sum_{j=1}^{d} w_j = 1$.

Then, w_j is called the classification weights of j characteristics under a given classification conditions.

Let $$\rho_{ik} = \sum_{j=1}^{d} w_j (x_{ij} - m_{kj})^2 \tag{26}$$

Where: δ is non-negative real number, usually taken as $\delta = 0.001 \sim 0.01$.

In (26), if $w_j = 0$, it is illustrated that the characteristic j has no contribution of distinction between K categories, so j should not appear in the calculation of the weight in the distance.

Thus, we can calculate the possibility of some measure that the sample x_i belonging to the k th category as follows:

$$\mu_{ik} = \mu(x_i \in C_k) = \frac{1}{\rho_{ik} + \varepsilon} \Big/ \sum_{k=1}^{K} \frac{1}{\rho_{ik} + \varepsilon} \tag{27}$$

Where, $\varepsilon = 0.01 \sim 0.001$,

Obviously, $0 \le \mu_{ik} \le 1$ and $\sum_{k=1}^{K} \mu_{ik} = 1$, therefore, μ_{ik} known as the unascertained measure of samples x_i belonging to the k th category, that is, we want to calculate the expectations output membership degree d_k^i, then $d_k^i = \mu_{ik}$

3) Unascertained RBF neural Network Learning Process:

RBF network has been used for the study, it is to classify the *N* known samples in *K* categories, and to determine the classification of unknown samples x. Unascertained RBF network is not only stress the desired output 1 or 0, but also required specific calculation, and the rest are same with BP networks. Unascertained RBF network parameters need to learn there are three: the center of basis function and the variance as well as the weights of hidden layer to output layer connection. The learning steps are as follows:

Center adjust: Unascertained-means clustering algorithm.

Given classification number *K* and the system accuracy ε_1, set counter $t = 0$;

Give the initial classification of n samples, get *K* cluster center vector $m_k(t), (k = 1,2,\cdots,n)$

Calculating the unascertained measure $\mu_{ik}(t), i = 1 \sim n, k = 1 \sim K$ of samples x_i belonged to the k th category

Determine a new type of center vector $m_k(t+1)$ from μ_{ik} as follows:

$$m_k(t+1) = \sum_{i=1}^{n} \mu_{ik} \cdot x_i \Big/ \sum_{i=1}^{n} \mu_{ik} \tag{28}$$

Calculate $err = \sum_{k=1}^{K} \left\| m_k(t+1) - m_k(t) \right\|$, and if $err \le \varepsilon_1$, so, stop iteration and turn to f); Otherwise, let $t = t+1$,turn to c)

Recalculate unascertained measure of the sample x_i belonging to the *k* th category,

Determining the varianc: In the center adjustment process, the variance v_k is determined by (28).

The study of connection weights

Supposing $\varphi_i, (i = 1 \sim K)$ is the output of the i th neuron months in hidden layer, $y_j, (j = 1 \sim K)$ is the actual output of the j neurons in output layer, d_j is the expectations corresponding output. Then, $y_j = \sum_{i=1}^{K} w_{ij} \cdot \varphi_i, (j = 1 \sim K)$.

The output layer error is:

$$E = \frac{1}{2} \sum_{j=1}^{K} (d_j - y_j)^2 \tag{29}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial w_{ij}} = -(d_j - y_j)\varphi_i \tag{30}$$

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = \eta (d_j - y_j)\varphi_i \tag{31}$$

Weight correction formula is as follows:

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \tag{32}$$

4) Unascertained RBF neural network learning process:

To identify samples x, input x to the trained network, supposing the greatest output is of the k_0 th output node, then x belongs to the k_0 th category.

Recognition Criteria:

$$k_0 = \max_k \left\{ \mu^k(x) \mid k = 1,2,\cdots,K \right\} \tag{33}$$

It's said that the sample x belongs to the k_0 th category.

E. Application case

In a market economy, enterprises are faced with a wide variety of risks. Therefore the establishment of a sound and effective financial risk early warning system is of great necessity to the monitoring and control of financial risk [11, 12]. We put the 45 selected sample data divided into training samples and test samples (30 as training samples, 15 as test samples) into unascertained neural network system. There were 15 nodes which value affect financial risk[11,12] is to input into neural network , 13 nodes in the hidden layer, and 1 node that indict the output value ('1' represents safety and '0' represents unsaved) of the risk in the output layer.

The learning rate was 0.01, and expectative error was 0.001.Then the neural network was programmed by software Matlab7.1. The training results are shown in Table1. The network structure is 15x13x1. The average variance EMS was 2.343 11×10-5,a nd training time was 54 second. Trained 2386 times, reaching the goal, training completed, the network convergence, when the total error is 0.000996. Re-enter the training samples to the best network training network detection, error rate to 0, and the network fitting fitting rate of 99.8%. Samples will be entered into the prediction network prediction, prediction results were shown in table 1.

F. Conclusions

Comparing Table.1 with the sample data, there is only one sample of mistake. Therefore the misjudgment rate is 6.67%, that is the correct identification rate is 93.33%.

From this example, we can see unascertained neural network for classification has a high application value. So, not only in theory but also in practice, combining unascertained systems with feedforward artificial neural

TABLE I.
PREDICTIVE RESULT TABLE OF PREDICTIVE SAMPLE

No.	**1**	**2**	**3**	**4**	**5**
Unsafe	0.6662	0.8934	0.2795	0.7171	0.7472
Safe	0.3338	0.1066	0.7205	0.2829	0.2528
No.	**6**	**7**	**8**	**9**	**10**
Unsafe	0.4332	0.6769	0.7105	0.5683	0.6029
Safe	0.5668	0.3231	0.2895	0.4317	0.3971
No.	**11**	**12**	**13**	**14**	**15**
Unsafe	0.0412	0.178	0.9355	0.8348	0.0642
Safe	0.9588	0.822	0.0645	0.1652	0.9358

networks can obtain more reasonable and more advantage of nonlinear mapping that can handle more complete type of and comprehensive data.

Comparing Table.2 with the sample data, there were no samples of mistake. Therefore, the misjudgment rate is 0, that is, the correct identification rate is 100%.

From this example, we can see unascertained RBF neural network for classification has a high application value. So, not only in theory but also in practice, combining unascertained systems with RBF artificial neural networks can obtain more reasonable and more advantage of nonlinear mapping that can handle more complete type of and comprehensive data.

REFERENCES

[1] Wang Guangyuan. Unascertained information and unascertained process. Journal of Harbin University of engineering, 1990(4):1-9. (in Chinese).

[2] Liu K D, Wu H Q, Pang Y J. Process and Application of Uncertain Information. Beijing: Science Press, 1999 (in Chinese).

[3] LIU Ya-jing, MAO Shan-jun, LI Mei, YAO Jiming. Study of a Comprehensive Assessment Method for Coal M ine Safety Based on a Hierarchical Grey Analysis. J China Univ Mining Technol 2007, 17(1):0006—0010.

[4] H.W.Shi, W.Q. Li, W.Q. Meng. A New Approach to Construction Project Risk Assessment Based on Rough Set and Information Entropy. 2008 International Conference on Information Management, Innovation Management and Industrial Engineering. Dec 2008:187-190.

[5] Lee,H.S.(2005).A fuzzy multi-criteria decision making model for the selection of the distribution center. Lecture notes in articial intelligence, 3612, 1290-1299.

[6] Liu Jun'e,Wang Haikuai,Zhang Likun. Application of Evaluating Model of Unascertained Measure in Bid & Tender of Construction Supervision. Hong Kong, China: Proceedings of 2004 International Conference on Construction & Real Estate Management, 2004, 337-340.

[7] Li Wan-qing, Ma Li-hua, Meng Wen-qing. Based on Unascertained Number Estimating Method of Project's Duration. Statistic and Decision, 2006, (5):131-133.(in Chinese)

[8] SHI Huawang. The Risk Early-warning of Hidden Danger in Coal Mine Based on Rough Set-neural network. Proceeding of the 2nd International Conference on Risk Management and Engineering Management. November 4-6,2008.pp314-317

[9] B. Irie, S. Miyake, Capability of three-layered perceptions, Proceedings of IEEE International Conference on Neural Networks, San Diego, USA, July 1988, pp. 641–648.

[10] Salchenberger, L.M., Cinar,E.M., Lash,N.A. Neural networks: a new tool for predicting thrift failures. Decision Sciences, 1992, 23, 899-916.

[11] Bose NK, Liang P. Neural network fundamental with graphs, algorithms and applications. McGraw-Hill International Editions; 1992.

[12] MATLAB The Mathworks Inc., version 7.0.1.24704.

Researches on Grid Security Authentication Algorithm in Cloud Computing

Keshou Wu *
Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361005, China
Email: kollzok@yahoo.com.cn

Lizhao Liu
Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361005, China
Email: 493107149@qq.com

Jian Liu
College of Information Sciences and Technology, The Pennsylvania State University, PA, USA
Email: 33095944@qq.com

Weifeng Li, Gang Xie, Xiaona Tong and Yun Lin
KOLLZOK Intelligent Technology Co., Ltd, Xiamen, 361024, China
Email: kollzok@yahoo.com

***Abstract*[1]—Focusing on multi-machine distributed computing security problems in cloud computing, the paper has proposed a grid distributed parallel authentication model based on trusted computing, which can realize simultaneous verification of grid authentication and grid behavior on upper layer of SSL and TLS protocols. Adaptive grid authentication method is established applying adaptive stream cipher framework; an adaptive stream cipher heuristic code generator and k-means heuristic behavior trust query function is proposed and acted as authentication kernel. Through comparison of the test results of TLS and SSL authentication protocol and the new grid authentication method, the effectiveness of the new grid authentication method has been explained.**

***Index Terms*—distributed computing; trusted computing; cloud computing; grid behavior; grid authentication; TLS; SSL**

I. INTRODUCTION

Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are cryptographic protocols that provide communications security over the Internet[1][2]. TLS and SSL encrypt the segments of network connections above the Transport Layer, using symmetric cryptography for privacy and a keyed message authentication code for message reliability. Several versions of the protocols are in widespread use in applications such as web browsing, electronic mail[3][4], Internet faxing, instant messaging and voice-over-IP (VoIP).TLS is an IETF standards track protocol, last updated in RFC 5246 and is based on the earlier SSL specifications developed by Netscape Corporation[5][6][7]. The TLS protocol allows client/server applications to communicate across a network in a way designed to prevent eavesdropping and tampering.A TLS client and server negotiate a stateful connection by using a handshaking procedure. During this handshake, the client and server agree on various parameters used to establish the connection's security[8][9][10].

Cloud computing refers to the provision of computational resources on demand via a computer network. In the traditional model of computing, both data and software are fully contained on the user's computer; in cloud computing, the user's computer may contain almost no software or data (perhaps a minimal operating system and web browser only), serving as little more than a display terminal for processes occurring on a network of computers far away[11][12]. A common shorthand for a provider's cloud computing service (or even an aggregation of all existing cloud services) is "The Cloud". The most common analogy to explain cloud computing is that of public utilities such as electricity, gas, and water. Just as centralized and standardized utilities free individuals from the vagaries of generating their own electricity or pumping their own water, cloud computing frees the user from having to deal with the physical, hardware aspects of a computer or the more mundane software maintenance tasks of possessing a physical computer in their home or office. Instead they use a share of a vast network of computers, reaping economies of scale [13][14].

Grid computing is a term referring to the combination of computer resources from multiple administrative domains to reach a common goal. The grid can be thought of as a distributed system with non-interactive workloads that involve a large number of files. What

*the corresponding author.
The work is supported by: The national natural science Foundation (60903203)

doi:10.4304/jnw.6.11.1639-1646

distinguishes grid computing from conventional high performance computing systems such as cluster computing is that grids tend to be more loosely coupled, heterogeneous, and geographically dispersed. Although a grid can be dedicated to a specialized application, it is more common that a single grid will be used for a variety of different purposes. Grids are often constructed with the aid of general-purpose grid software libraries known as middle ware[15][16].

Trusted Computing (TC) is a technology developed and promoted by the Trusted Computing Group.The term is taken from the field of trusted systems and has a specialized meaning. With Trusted Computing, the computer will consistently behave in expected ways, and those behaviors will be enforced by hardware and software. In practice, Trusted Computing uses cryptography to help enforce a selected behavior. The main functionality of TC is to allow someone else to verify that only authorized code runs on a system. This authorization covers initial booting and kernel and may also cover applications and various scripts. Just by itself TC does not protect against attacks that exploit security vulnerabilities introduced by programming bugs[17][18].

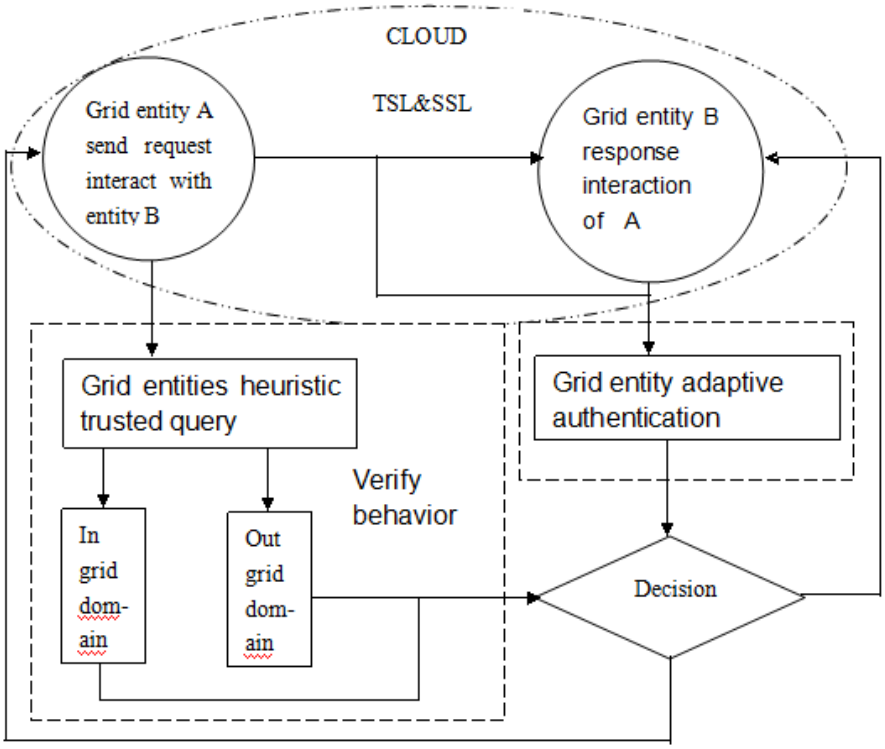


Figure 1. Grid distributed parallel authentication model

II. GRID DISTRIBUTED PARALLEL AUTHENTICATION MODEL

If the grid entity A want to intact with grid entity B in a cloud ,the grid entity A will first go into grid entities heuristic trusted query, this process need to calculate trusted value in grid domain and out of grid domain; at the other hand it need to compute the grid entity adaptive authentication. If the verify behavior reach its gate value the information will be sent to the decision module besides the information of grid entity adaptive authentication, then the decision module will give the comprehensive information of the trusted value of grid B for A. During the process grid B will interact with grid entity adaptive authentication module to give sufficient information or else it will be rejected.

III. ADAPTIVE GRID AUTHENTICATION VERIFY FRAME

Adaptive grid authentication verifies can realize signal self- detection and self-adjusting. [19] The adaptive generator initialization within the production of continuous or intermittent output with automatic recognition and adjustment function of the generate signal, through the design of reference models or self-tuning controller module can be achieved on the output or received signal real-time adjustment and dynamic match. Adaptive encryption control principle is as Figure 2.First initiative the clock module and the clock stimulus module as a self-reference model, since the self-reference model will reconstruct when the detective signal received from the self-detection module does not match, and the re-constructed reference model is not dependent on external stimulation, which depends only on the initial algorithm^{E}. This means that as long as both encryption and decryption have the same reference model, after the same initialization, they can always get synchronous control signal. For example, the use of the two CMOS unit can keep output synchronism at $k \cdot 10^{8} / S$ .Take the output signal from the self-reference model as the first stage parameter of chaos cascade module, the output signal of the first stage of chaotic module as the input

signal of self-tuning module and at the same time, as the input signal of the second stage Logistic generator; the output signal of the second stage Logistic generator as the input signal of the third stage or input signal of key generator; the output signal of the third stage and key k together as the initial key of key generator.The self-detection module consists of two detectors, which are responsible for adding state value w to the plaintext and testing encrypted state value w_{1c} and channel feedback state value w'_{1c} . When the encrypted value and the channel feedback state values are abnormal, feedback should be done to the reference model, then the self-reference model will update with a new reference model according to the current state value of initial algorithm, thus to update all the output signals of chaotic module, key generators will also update new key stream without changing of key k.The receiver uses the decoder to return w'_{1c} to the sender first to confirm the acceptance of the cipher, at the same time offers w'_{1c} evaluation to the sender for the channel safety index testing thus to determine whether there is need to resend the cipher or renew the reference model.

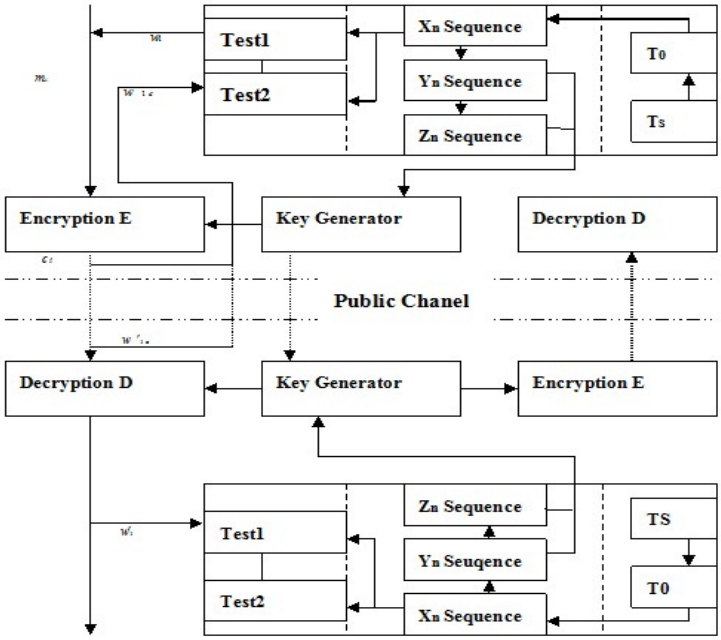


Figure 2. Adaptive grid authentication method

IV. HEURISTIC CODE GENERATOR AND HEURISTIC BEHAVIOR TRUST QUERY FUNCTION

A. Structure of heuristic code generator

Supposed that the grid entity have some shape volume and weight[20], the length of grid entity is L, the width W and the height H, the direction of the layers vary from each other due to different position of grid entity, and the x, y, z mentioned above correspondingly represents different value[21]. If the grid entity is located behind of grid domain, then x equals to W, y equals to H, and z equals to L, which means the direction of the layer is along with the length of entity. If the grid entity is located in the side of compartment, then x equals to L, y equals to H, z equals to W, and the direction of the layer is along with the width of compartment. If there are two entitiess set respectively behind and side of the grid, then x equals to W, y equals to L, z equals to H, and the direction of the layer is along with the height of grid. Call the best layer loading program and the plane optimal layout program to solve these and the specific steps are as follows:

(1) enter the grid entity size L, W, H, the gridgo box size s_1, s_2, and s_3 (make $s_1 \geqq s_2 \geqq s_3$). Assign x, y, z with L, W, H according to the position of the grid entity;

(2) call the best layer loading program to compute a, b, c;

(3) Calculate the optimal layout of each layer:

① If $a \neq 0$, make $s_m = s_2$, $s_n = s_3$, call the plane optimal layout program genenator (s_2, s_3);

② If $b \neq 0$, make $s_m = s_1$, $s_n = s_3$, call the plane optimal layout program genenator (s_1, s_3);

③ If $c \neq 0$, make $s_m = s_1$, $s_n = s_2$, call the plane optimal layout program genenator (s_1, s_2);

(4) Calculate the total number of the loading container Sum= a*genenator (s_2, s_3) + b*genenator (s_1, s_3) + c*genenator (s_1, s_2);

(5) Output Sum, a, b, c, and the corresponding values of the parameters of the layout.

Given a grid grid entityset C, C={1,2,…,n}, the set $C_0=\{0\}$ denotes the grid; d_{ij} is the trusted value between

arbitrary two nodes i, j in C and C_0; q_i (i=1,…,n) is the demand of the grid i; w is the maximum of the trusted capability of, R is the number of the grid entity that needs to finish the vertity, which is

$$R = \left\lceil \sum_{i=1}^{n} q_i / w \right\rceil, \tag{1}$$

"$\lceil\ \rceil$"is the rounded up function, such as $\lceil 6.2 \rceil = 7$; x_{ij}^r (r=1,…, R, i and j=0,…, n, and where i not equals to j) is the decision variables, $x_{ij}^r = 1$ if and only if the r routine pass the arc(i, j), otherwise $x_{ij}^r = 0$; y_i^r ((r=1,…, R, i =1,…, n) is the demand of the i grid which meets by the r routine; S^r denotes the grid set served by the r routine, $|S|^r$ denotes the number of grid included in S. There are some assumptions of the model:

(1) the trusted values between two nodes is symmetric, $d_{ij} = d_{ji}$;

(2) the trusted values of the nodes satisfy the triangular inequality, which is $d_{ik} + d_{kj} > d_{ij}$;

(3) all the grid entity start from the grid and back to grid after each delivery;

(4) every grid 's needs must be satisfied and can be done by one or more grid entity.

The objective of this problem is to arrange the routine to minimize the cost of delivery. The cost is represented by the total travelling trusted value. As the description above, the problem can be modeled as:

$$\min \sum_{r=1}^{R} \sum_{i=0}^{n} \sum_{j=0}^{n} d_{ij} x_{ij}^r \tag{2}$$

$$\sum_{i=0}^{n} x_{ik}^r = \sum_{j=0}^{n} x_{kj}^r \quad k = 0,\dots,n; r = 1,\dots,R \tag{3}$$

$$\sum_{r=1}^{R} \sum_{i=0}^{n} x_{ij}^r \geq 1 \quad j = 0,\dots,n \tag{4}$$

$$\sum_{r=1}^{R} y_{ri} = q_i \quad i = 1,\dots,n \tag{5}$$

$$\sum_{i\in s} \sum_{j\in s} x_{ij}^r = |S| - 1 \quad r = 1,\dots,R; S \subseteq C - \{0\} \tag{6}$$

$$\sum_{i=1}^{n} y_{ri} \leq w \quad r = 1,\dots,R \tag{7}$$

$$\sum_{j=0}^{n} x_{ij}^r q_i \geq y_{ri} \quad r = 1,\dots,R; i = 1,\dots,n \tag{8}$$

$$x_{ij}^r \in \{0,1\} \quad i, j = 1,\dots,n; r = 1,\dots,R \tag{9}$$

$$q_i \geq y_{ri} \geq 0 \quad i = 1,\dots,n; r = 1,\dots,R \tag{10}$$

The constraint (2) is to minimize the total travelling trusted value; constraint (3) means the flow conservation, that is, the number of grid entity is equal between entering and exiting of a node; Constraint (4) and (5) ensure that each node is visited at least one time and the requirement is satisfied; (6) shows that the edges between served grid s equals to the number of served grid s minus 1 in each route, (7) shows thetruseted capability of grid; (8) shows that the grid is served only the grid pass.

Compute 1: input x, y, s_m, s_n

Compute 2: determine whether the grid A and B can be trusted in its domain, return 0 if not.

Compute 3: for $(y_1 = \lfloor y/2s_n \rfloor; y_1 \leq \lfloor y/s_n \rfloor; y_1 ++)$

{ for $(x_1 = \lfloor x/2s_m \rfloor; x_1 \leq \lfloor x/s_m \rfloor; x_1 ++)$

{initialize $x_5 = y_5 = 0$ and compute x_2, y_4 using $\sum_{r=1}^{R} \sum_{i=0}^{n} \sum_{j=0}^{n} d_{ij} x_{ij}^r$ and

$\sum_{i=0}^{n} x_{ik}^r = \sum_{j=0}^{n} x_{kj}^r \quad k = 0,\dots,n; r = 1,\dots,R$;

If $(\lceil s_n y_1 / s_m \rceil \geq \lfloor y/s_m \rfloor)$

{ $y_2 = \lfloor y/s_m \rfloor$; $x_3 = y_3 = 0$; Compute x_4 using $\sum_{r=1}^{R} y_{ri} = q_i \quad i = 1,\dots,n$ }

else

{ $y_2 = \lceil s_n y_1 / s_m \rceil$; Compute y_3 using $\sum_{i\in s} \sum_{j\in s} x_{ij}^r = |S| - 1 \quad r = 1,\dots,R; S \subseteq C - \{0\}$;

If $(s_n x_2 - s_m \lfloor s_n x_2 / s_m \rfloor > s_m / 2)$ { compute x_3, x_4 using $q_i \geq y_{ri} \geq 0 \quad i = 1,\dots,n; r = 1,\dots,R$.}

Else { compute x_3, x_4 using $\sum_{i=1}^{n} y_{ri} \leq w \quad r = 1,\dots,R$ and $\sum_{r=1}^{R} y_{ri} = q_i \quad i = 1,\dots,n$;

If$(x - s_m x_3 - s_n x_4 > s_n)$ { compute x_5, y_5 using $\sum_{j=0}^{n} x_{ij}^r q_i \geq y_{ri} \quad r = 1,\dots,R; i = 1,\dots,n$ and $x_{ij}^r \in \{0,1\} \quad i, j = 1,\dots,n; r = 1,\dots,R$ }

}

}

Compute the total number of generator= $x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 + x_5 y_5$.

Compare and record the generator number and the agreement method.

}

};

Compute 4：Output the optimized result.

We can also have $x_5 = \lfloor (x - s_m x_3 - s_n x_4) / s_n \rfloor$

$y_5 = \lfloor (y - s_m y_2) / s_m \rfloor$.

The objective function is generator= $x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4 + x_5 y_5$.

When $s_n x_2 - s_m \lfloor s_n x_2 / s_m \rfloor > s_m / 2$,

$x_3 = \lfloor s_n x_2 / s_m \rfloor + 1 \quad x_4 = \lfloor (x - s_m x_3) / s_n \rfloor$,

The objective function is

generator= $x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$.

So in such verify model, the parameters can be solved as long as x_1 and y_1 are known. The range of x_1, y_1 are $0 \le x_1 \le \lfloor x/s_m \rfloor$ and $0 \le y_1 \le \lfloor y/s_n \rfloor$, which becomes to $\lfloor x/2s_m \rfloor \le x_1 \le \lfloor x/s_m \rfloor$ and $\lfloor y/2s_n \rfloor \le y_1 \le \lfloor y/s_n \rfloor$ on considering the symmetry of verify model 4. The objective optimization value can be found after the traversal of all the combinations of x_1, y_1 .

B. k-means heuristic behavior trust query function

In many practical applications, the k-means clustering algorithm (k-means algorithm) which is based on partition clustering has been proven to be effective and generate good results[22][23]. The steps of a general k-means algorithm are:

Select k couples of initial cluster center;

Assign sample x which need to be classified to some cluster center one by one according to the minimum trusted value principle;

Calculate new value of every cluster center. Generally the new cluster center is the mean vector of the sample contained in the cluster field. The mean vector of the sample in k couples of cluster need to be calculated respectively.

Reclassify the sample and repeat iteration. The algorithm converges when every cluster center no longer moves, then calculation finishes.

The principle of k-means algorithm is to find k couples of partition with a least square error and make the generated result as compact and separate as possible. The k-mean algorithm is relatively scalable and efficient dealing with large data sets and the complexity is $O(nkt)$, in which n means the number of objects, k is the number of cluster, and t is the number of iterations. The case mainly discussed in this paper is that the demand of grid may be greater than the maximum trusted capacity of grid entity. Hence, it is prior to meet each grid wholly, and then merge the remaining part to other grid to meet.

Next the principle discussed is used to cluster the grid entities and determine the grid s served by the same grid. However, the SDVRP is a constraint clustering problem, the calculation may not converge, so the number of iterations N needs to be set to terminate forcibly and set the clustering evaluation criteria to select better clustering results. The clustering evaluation function used in this paper is:

$$Min(sumD) = \sum_{j=1}^{R} \sum_{i \in C_j} d_{ij} ,$$

C_j represents cluster j. The formula above calculates the sum of trusted value between every grid entity and the center in the cluster. Select the minimum sum as the best clustering result. The concrete steps are below:

Step 1: Find the grid entity whose demand is greater than or equal to the trusted capacity of grid. Split the demand q_i to two parts q_i^s and q_i^c , and

$q_i^s = w \lfloor q_i / w \rfloor$

$q_i^c = q_i - w \lfloor q_i / w \rfloor$

"$\lfloor \quad \rfloor$" means to round down, for example $\lfloor 6.6 \rfloor = 6$.

The demand of q_i^s is individually met and the remained demand q_i^c and the other entity are merged to some other circuit to meet. Modify the demand of the grid i to be q_i^c ;

Step 2: Randomly select R couples of initial cluster center $1^1, \cdots, R^1$ from the grid set $C = 1, 2, \cdots, n$, and mark as set $P^1 = 1^1, \cdots, R^1$. Initialize every cluster set $C_i = \Phi(i = 1, \cdots, R)$, and set the value of the maximum number of iterations N;

Step 3: Cluster the grid s. Calculate the trusted value d_{ij} between every grid entity and every cluster center, and find the nearest cluster center of every grid entity. The nearer the trusted value is, the higher priority the grid entity has to join the center. If the cluster wanted to join is full loaded, then choose the second nearest. When there is still remaining demand in the cluster, and if the adding of the demand make the total demand of cluster C_j exceed W ($QC_j > W$), compute the unmet demand of grid i, which is denoted by S, and transmit the unmet demand to other grid of C_j. The transmission principle is: firstly find the grid entity (include grid i) whose demand is not less than S in cluster C_j, then find the cluster whose residual demand $SuQ_z = W - QC_z \ge S(z \in P^1 - j)$. Compute the trusted value between these grid entitys and these clusters and choose the grid entity with smallest trusted value to split. Guess the grid entity k and its corresponding cluster center p, add k to cluster C_p and the unmet demand S is met by this route. If the residual demand of all clusters $SuQ_z < S$, then select the cluster with largest residual demand to join until S is fully met. Repeat this step until all the grid s' demands are met.

Step 4: Calculate the sum of the trusted value between every clustering grid entity and its cluster center *SumD*;

Step 5: Use the following way to adjust the cluster center and get the new $1^2, \cdots, R^2$. The coordinate position of the cluster center $j^2 (j = 1, \cdots, R)$ is

$$x_{j^2} = \frac{1}{n_i} \sum_{x_i \in C_i} x_i, y_{j^2} = \frac{1}{n_i} \sum_{y_i \in C_i} y_i ,$$

Where n_i is the number of grid entity in C_i;

Step 6: Repeat Step 3-5 until reach the maximum iteration number N. Output the clustering results corresponding to the minimum value of *sumD*;

Step 7: Optimize the result of step 6 by simulated annealing algorithm. The cool way

is $T(t+1)=k\times T(t)$. In the formula k is a positive constant slightly less than 1.00 and t is the times of cooling.

In step 1, the situation that the grid demand is greater than the trusted capacity of grid is considered. In step 2 to 6, cluster the grid s need to, and find the optimal clustering solution. In step 7, the route optimization is done for solving TSP problem.

The clustering process is:

(1) Random determine R (obtained from formula (1)) couples of cluster center;

(2) Calculate the trusted value between every grid entity and every cluster center $d_{ij}(i=1,\cdots,n;\ j=1,\cdots,R)$. Sort $d_{ij}(j=1,\cdots,R)$ from small to large and find the smallest trusted value from every grid entity to the cluster center.

(3) If the smallest trusted value d_{kp} is found, then the corresponding grid k is added to cluster p, and add the grid corresponding to the second smallest value to the corresponding cluster, compute the residual demand *SuQ* (that is, the capacity of gridriage minus the amount of grid mounted) of the cluster and turn down. When the residual demand of cluster is less than the demand the grid s want to add, the split entities are selected to split in cluster. The principle of split entity selection will be discussed later.

(4) When the total demand of the cluster that the grid s want to join has reached the maximumtruseted capacity of grid entity, the second nearest cluster will be considered. Turn down until all grid s are added into a cluster.

In order to ensure the load factor and the least requirement of grid entity, the grid's need is allowed to split, so the principle of grid choice splitting should be considered. If grid i is added into a cluster p which is not fully loaded, which makes the total demand of the cluster exceeds the maximum trusted capacity of grid entity, the demand needs to be split to meet. If the second nearest cluster center is far away from the grid, the traffic trusted values increase greatly. The unmet demand will be allowed to transmit to a entity whose demand is greater than the unmet demand of grid i in cluster p and which is relatively close to the other cluster whose residual demand should be greater than the unmet demand of grid i, to make the demand of this entity split meet. The demand of grid i is totally met by cluster p. If the residual demand of all clusters is lower than the unmet demand of grid i, then choose the one with the maximum residual demand to join to avoid being split too many times.

V. COMPARISON OF THE TEST RESULTS OF TLS AND SSL AUTHENTICATION PROTOCOL AND THE NEW GRID AUTHENTICATION METHOD

We set the clouds as a pool with hundreds of computers and there are many grid entities that cannot be trusted or should be limited for intact, then we set some entity to send the request to other grids to compute or calculate some information together, so every grid in the clouds will go into the TLS\SSL model and our new model using distributed parallel authentication model based on trusted computing, then we reminder the accuracy and lead time of all the model.

From the table1 and table2 we can see that the accuracy rate of SSL&TLS authentication is lower than distributed parallel authentication model, the lead time of SSL&TLS authentication is longer than distributed parallel authentication model. In table3 we will show the detail comprehensive improvement for different clouds and different internet environment.

TABLE I.
ACCURACY RATE OF SSL&TLS AUTHENTICATION AND THE LEAD TIME

Experiment index	Accuracy rate(km)	TLS TIME(ms)	SSL TIME(ms)
1	63.1	163	537
2	69.2	175	805
3	67.9	170	966
4	66.4	166	881
5	65.0	169	946
Aver.	66.32	168.6	827

TABLE II.
THE ACCURACY RATE AND LEAD TIME FOR USING DISTRIBUTED PARALLEL AUTHENTICATION MODEL

Experiment index	Accuracy rate	Computation time(s)
1	94.49	5.515
2	92.12	5.170
3	94.57	4.911
4	99.45	5.069
5	94.74	5.010
6	91.12	5.053
7	94.49	5.586
8	92.12	5.174
9	94.57	4.938
10	99.45	5.068
Aver.	92.712	5.1494

From the table3 we can see the comprehensive evaluation of distributed parallel authentication model is much better than SSL&TLS that the distributed parallel authentication model use less computing operation and computing times but with 30 higher correct accuracy percent and 35.7 equal total percent.

TABLE III.
ACCURACY RATE OF SSL&TLS AUTHENTICATION AND THE LEAD TIME
AUTHENTICATION MODEL AND SSL&TLS

Authentication	Comprehensive evaluation of our algorithm		Comprehensive evaluation of SSL&TLS		Improve-ment (%)
	Correct times	Computation time(s)	Correct times	Computation time(s)	
A01	4675585	4.6	5307907.00	17	31.9
A02	8158990	7.9	8542757.00	64	34.5
A03	8149102	9.5	8413577.00	60	33.1
A04	10696819	15.2	10708613.00	440	30.1
A05	13682582	22.5	13403505.00	1900	32.1
A10	13929231	22.3	13403505	40	33.9
A11	9972833	10.7	10569587.00	86	35.6
B01	4228238	5.6	4629056.00	27	38.7
B02	5966489	9.2	6239394.00	78	34.4
B03	7618932	13.6	7714649.00	122	31.2
B04	9915250	24.3	9471386.00	545	34.7
B05	12895562	36.1	11482700	1224	32.3
B10	12575805	35.8	11482700	516	39.5
B11	12074965	17.9	10552825.00	85	34.4
C01	6894370	7.0	7653121.00	56	3.9
C02	9809075	11.7	11340760.00	71	33.5
C03	14331431	17.6	15151732.00	206	35.4
C04	21142701	31.8	21018042.00	564	30.6
C05	27019924	47.4	25858494.00	3811	34.5
C10	27011949	46.2	25858494	259	34.5
C11	32827882	23.8	30604668	188	37.3
D01	9747314	9.5	10391059.00	34	36.2
D02	15100849	16.6	15566936.00	311	33.0
D03	19555495	25.2	20541296.00	412	34.8
D04	31193493	48.7	29916416.00	1822	34.3
D05	41245476	72.4	36242004	2598	33.8
D10	41002814	70.6	36242004	1037	33.1
D11	50395453	35.7	45026152	523	31.9
EQUAL	9238523	34..9	20958379	581	35.7

VI. CONCLUSION

From the above analysis, take trusted computing as the basis, in a cloud computing, grid distributed parallel authentication method which is realized by grid authentication and grid behavior simultaneous authentication, established on the upper layer of SSL and TLS protocols, by adaptive stream cipher heuristic code generator and heuristic behavior trust query function, plays well in authentication. However, on the trust issue

of grid behavior, further standardization is needed on entities quantitative trust level within a domain, while the core of the heuristic algorithm needs to quantify the grid entities with the shape, weight, size and other physical indicators as a physical entity, this quantitative method still needs to be further improved, so as to promote adaptive stream cipher authentication framework and improve the upper trusted computing platform.

REFERENCES

[1] http://tools.ietf.org/html/rfc5246
[2] The SSL Protocol: Version 3.0Netscape's final SSL 3.0 draft (November 18, 1996)
[3] "SSL/TLS in Detail". Microsoft TechNet. Updated July 31, 2003.
[4] Thomas Y. C. Woo, Raghuram Bindignavle, Shaowen Su and Simon S. Lam, SNP: An interface for secure network programming Proceedings USENIX Summer Technical Conference, June 1994
[5] Dierks, T. and E. Rescorla. "The Transport Layer Security (TLS) Protocol Version 1.1, RFC 4346". http://tools.ietf.org/html/rfc5246#ref-TLS1.1.
[6] National Institute of Standards and Technology. "Implementation Guidance for FIPS PUB 140-2 and the Cryptographic Module Validation Program". http://csrc.nist.gov/groups/STM/cmvp/documents/fips140-2/FIPS1402IG.pdf.
[7] Eric Rescorla (2009-11-05). "Understanding the TLS Renegotiation Attack". Educated Guesswork. http://www.educatedguesswork.org/2009/11/understanding_the_TLS_renegoti.html.
[8] McMillan, Robert (2009-11-20). "Security Pro Says New SSL Attack Can Hit Many Sites". PC World.
[9] "SSL_CTX_set_options SECURE_RENEGOTIATION". OpenSSL Docs. 2010-02-25.
[10] Various (2002-08-10). "IE SSL Vulnerability". Educated Guesswork.
[11] Sean Marstona; Zhi Lia; Subhajyoti Bandyopadhyaya; Juheng Zhanga; Anand Ghalsasib. "Cloud computing — The business perspective". Decision Support Systems.
[12] M. Armbrust; A. Fox; R. Griffith; A.D. Joseph; R.H. Katz; A. Konwinski; G. Lee; D.A. Patterson; A. Rabkin;

I. Stoica and M. Zaharia. "Above the Clouds: A Berkeley View of cloud computing". University of California at Berkeley. 10 April 2011.
[13] "NIST.gov – Computer Security Division – Computer Security Resource Center". Csrc.nist.gov.
[14] "Gartner Says Cloud Computing Will Be As Influential As E-business". Gartner.com. 2010-08-22.
[15] a b "What is the Grid? A Three entity Checklist". http://dlib.cs.odu.edu/WhatIsTheGrid.pdf.
[16] Diuf.unifr.ch. "Pervasive and Artificial Intelligence Group: publications [Pervasive and Artificial Intelligence Research Group]". May 18, 2009.
[17] Chris Mitchell, Trusted Computing, Institution of Electrical Engineers, 2005.
[18] Ross Anderson, "Cryptography and Competition Policy - Issues with 'Trusted Computing' ", in Economics of Information Security, from series Advances in Information Security, Vol. 12, April 11, 2006.
[19] Liu Lizhao, A New Adaptive SSC and SSSC Stream Cipher Model Design and Implementation [J]. Advanced Materials Research Journal: 2011, 1(143), 298 -303
[20] Yang, H., J. Shi.: A Hybrid CD/VND Algorithm for three-dimensional bin packing [C]. The 2nd International Conference on Computer Modeling and Simulation. IEEE Press, Sanya(2010)
[21] Almeida A. d., Figueiredo M.B.: A particular approach for the three-dimensional packing problem with additional constraints [J]. Computers & Operations Research. 37(11), 1968-1976(2010)
[22] C. Archetti, A. Hertz, M.G. Speranza. A Tabu search algorithm for the split delivery vehicle routing problem[J].Transportation Science, 40, 64-73(2006)
[23] C. Archetti, M.W.P. Savelsbergh, M.G. Speranza. An optimization-based heuristic for the split delivery vehicle routing problem [J]. Transportation Science, 42, 22-31(2008)

Keshou Wu (1975.3-), Xiamen city, Fujian Province, China, PhD of Huazhong university of science and technology, majored in software engineering. Research field: System Engineering, Information System, Data Mining, GIS.

Lizhao Liu (1983.3-), Xiamen city, Fujian province, China. PhD candidate of Xiamen university, majored in automation, system engineering, Information Science and Technology Department. Research field: chaotic modeling and control of unmanned airplane vehicle and information system, feature attraction and detection, scale space and multiscale technology.

He has done the China national 985 engineering process of unmanned airplane vehicle for the UAV\UAIS chaotic phenomenon analysis, UAV\UAIS chaotic modeling and control. He made the paper such as The Chaotic Characters and New Control Strategy of Unmanned Airplane Information System 2008 ISCID and Error! Reference source not found. The Chaotic Disturbance of UAV System's Communication And Coping Strategy 2008 ICCCAS. He also has done the work of grid behavior trust model and has the paper such as The Quantitative Assignment of The Grid Behavior Trust Model Based on Trusted Computing 2010 Wuhan university journal.Now he is doing the work of scale space and multiscale technology for the image analysis especially for the feature describtion definition detection and matching.

Non-line-of-sight Error Mitigation in Wireless Communication Systems

Chien-Sheng Chen
Tainan University of Technology / Department of Information Management, Tainan, Taiwan

Yi-Jen Chiu
Taiwan Shoufu University / Department of Digital Entertainment and Game Design, Tainan, Taiwan
E-Mail: cyj@tsu.edu.tw

Ho-Nien Shou
Air Force Institute of Technology / Department of Aviation and Communication Electronics, Kaohsiung, Taiwan

Ching-Lung Chi
Shu-Te University / Department of Computer and Communication, Kaohsiung, Taiwan

***Abstract*—The need for determining the position of a mobile station (MS) is increasing rapidly in wireless communications systems. When there is non-line-of-sight (NLOS) path between the MS and base stations (BSs), it is possible to integrate many kinds of measurements to achieve more accurate measurements of the MS location. This paper proposed hybrid methods that utilize time of arrival (TOA) at five BSs and angle of arrival (AOA) information at the serving BS to determine the MS location in NLOS environments. The methods mitigate the NLOS effect simply by the weighted sum of the intersections between five TOA circles and the AOA line without requiring priori knowledge of NLOS error statistics. Simulation results show that the proposed methods always give superior performance than Taylor series algorithm (TSA) and the hybrid lines of position algorithm (HLOP).**

***Index Terms*—Time of arrival (TOA), Angle of arrival (AOA), Non-line-of-sight (NLOS)**

I. INTRODUCTION

The problem of position determination of a mobile user in a wireless network has been studied extensively in recent year. It is always desirable to achieve the highest possible accuracy in location applications. However, the requirements in different applications may differ due to various reasons such as the cost and the technology. There are various techniques for wireless location, which can be broadly classified into two categories --handset-based techniques and network-based techniques. From the technical aspect, the handset-based techniques are easy to implement and accurate to determine the mobile station (MS). Global positioning system (GPS) requires installation of a receiver and transmitting the received GPS data to the base station (BS) for further processing and position determination. The drawbacks of this technique include the high cost for developing a suitable low-power and economical integrated technology for use in the handsets. Moreover, a GPS receiver needs to have at least four satellites constantly visible. Therefore, the GPS-based solution is a feasible option for outdoor positioning but not for indoor positioning within urban environments. The existing wireless communications infrastructure without supplementary technology has been utilized in MS location estimation. One of the goals of the location solution is to allow carriers to locate current users by existing network without expensive modifications and be adaptable to complement satellite handset-based techniques.

And the primary network-based techniques of wireless communication systems include signal strength [1], angle of arrival (AOA) [2], time of arrival (TOA) [3], and time difference of arrival (TDOA) [4] techniques. Signal strength is a location method that uses a known mathematical model describing the relation between the path loss attenuation and distance. If the angle in which the signal MS arrives to the BS can be measured, an AOA line can be drawn. By measuring AOA angles at least two BSs, the intersection of two lines can be obtained where the MS would be located. TOA location scheme measures the propagation time it took for the signal to travel between the MS and the BS. The TDOA is to determine the relative position of the MS by examining the difference in arrival-time measurements at multiple BSs, rather than absolute arrival time.

The accuracy of mobile location estimation strongly depends on the propagation conditions of the wireless channels. The radio signals are usually corrupted by additive noise, multipath propagation, and non-line-of-sight (NLOS) propagation in wireless location system [5]. To enhance the precision of the location estimation, appropriate steps must be taken to mitigate these impairments. The additive noise is relatively easy to control comparing to other wireless channel impairment. Usually, it is modeled as zero mean

doi:10.4304/jnw.6.11.1647-1653

Gaussian noise with variance determined by the signal to noise ratio (SNR), measurement resolution, and other factors. In wireless locating systems, the transmitted signals are frequently corrupted with multipath propagations; the line-of-sight (LOS) is blockaged. Many procedures are necessary to reduce the effects of impairments. First of all, we define multipath as the presence of multiple signal paths between the MS and BS. In general, multipath propagation is already a problem for most positioning techniques even if the LOS path is existent. Severe multipath propagation can reduce the positioning accuracy significantly. The reflections and diffractions from buildings in urban areas or mountains in rural terrains can cause significant path blockages and multipath time dispersions [6]. In the case of AOA schemes, multipath in the propagation channels would significantly degrade the performance of the estimation of direction-of-arrival [7]. In particular, the multipath propagations may also cause serious problems in the signal strength measurements [5]. Even when the LOS propagation exists, the multipath propagations can induce errors in the timing estimations of the time-based location systems [5].

The NLOS condition is even more critical, because the LOS path is blocked additionally. A common requirement for high location accuracy is the presence of a LOS path between the MS and each participating BS. In practice, LOS paths are not always readily available. The NLOS propagation occurs usually in urban or suburban areas. Due to the reflection or diffraction of the signals between the MS and the BSs, NLOS propagation results in significant errors in the time and angle measurements. When a *priori* knowledge of the NLOS error is available, different NLOS identification and correction algorithms for determining MS location are proposed [8]. The standard deviation of the range measurements for NLOS propagation is much higher than LOS propagation [9].

To improve the accuracy of MS location, it is reasonable to combine two or more schemes give location estimation of the MS. Hybrid techniques can be used to take the advantage of various positioning schemes. A hybrid TDOA/AOA algorithm that can offer more accurate location estimation for wideband CDMA cellular systems was proposed in [10]. To achieve high location accuracy, the scheme uses TDOA information from all BSs and the AOA information at the serving BS in small error conditions. A hybrid range/range difference algorithm was used to estimate the MS location in a GSM system when only two base transceiver stations (BTS's) are available and the MS is located at the mass center of the serving cell [11]. The positioning performance improvement of TDOA schemes using the AOA measurement from the serving BS over pure TDOA schemes is evaluated in [12]. We have proposed hybrid geometrical positioning schemes to estimate MS location under the condition that the MS can be heard by only two BSs in [13].

In this paper, we apply the hybrid geometrical positioning schemes to locate MS when five BSs are available for location purposes. We present a mobile positioning system that adopts TOA-aided AOA information at five BSs to estimate the location of an MS. By acquiring the intersections of five TOA circles and AOA line, it is possible to locate the desired MS in wireless communication systems. The proposed positioning methods are based on the weighted sum of the intersections of five TOA circles and the AOA line. Simulation results show that the proposed methods always achieve better location accuracy than Taylor series algorithm (TSA) [14] [15] and the hybrid lines of position algorithm (HLOP) [16].

The remainder of this paper is organized as follows. The system model is given in Section II. Section III presents the commonly used positioning methods TSA and HLOP. Section IV describes various approaches using the intersections of the five TOA circles and the AOA line to estimate the position of MS. Simulation results are presented in Section V. Conclusion is given in Section VI.

II. System Model

TOA measurements from five BSs and the AOA information at the serving BS can be employed to give a location estimate of the MS, as shown in Fig. 1 [17]. Let t_i denote the propagation time from the MS to BS i, and the coordinates for BS i are given by (X_i, Y_i), $i = 1, 2, ...5$. The distances between BS i and the MS can be expressed as

$$r_i = c \cdot t_i = \sqrt{(x - X_i)^2 + (y - Y_i)^2} \tag{1}$$

where (x, y) is the MS location and c is the propagation speed of the signals. We assume that BS1 is the serving BS, and denote by θ as the angle between MS and its serving BS.

$$\theta = \tan^{-1}\left(\frac{y - Y_1}{x - X_1}\right) \tag{2}$$

III. Taylor Series Algorithm (TSA) and Hybrid Lines of Position Algorithm (HLOP)

To determine the MS location, TSA [14] [15] and HLOP [16] are the most used schemes.

A. Taylor Series Algorithm (TSA)

TOA and AOA measurements are inputs to the Taylor series position estimator. Let (x, y) be the true position and (x_v, y_v) be the initially estimated position. Assume that $x = x_v + \delta_x$, $y = y_v + \delta_y$. By linearizing the TOA and AOA equations through the use of a Taylor series expansion and retaining second-order terms, we have

$$A\delta \cong z \tag{3}$$

where $A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{51} & a_{52} \\ b_{11} & b_{12} \end{bmatrix}$, $\delta=\begin{bmatrix}\delta_x \\ \delta_y\end{bmatrix}$, $z=\begin{bmatrix} r_1-r_{v1} \\ r_2-r_{v2} \\ \vdots \\ r_5-r_{v5} \\ \theta-\theta_v \end{bmatrix}$,

and $a_{i1}=\left.\frac{\partial r_i}{\partial x}\right|_{x_v,y_v}$, $a_{i2}=\left.\frac{\partial r_i}{\partial y}\right|_{x_v,y_v}$,

$r_{vi}=\sqrt{(x_v-X_i)^2+(y_v-Y_i)^2}$, $i=1,2,...5$,

$b_{11}=\left.\frac{\partial \theta}{\partial x}\right|_{x_v,y_v}$, $b_{12}=\left.\frac{\partial \theta}{\partial y}\right|_{x_v,y_v}$, $\theta_v=\tan^{-1}(\frac{y_v-Y_1}{x_v-X_1})$.

Then, the least-square (LS) estimation can be solved by

$$\delta=(A^TA)^{-1}A^Tz \tag{4}$$

The process starts with an initial guess for the MS location and can achieve high accuracy. This method is recursive but tends to be computationally intensive. TSA may suffer from the convergence problem if the initial guess is not accurate enough [14] [15].

B. Hybrid Lines of Position Algorithm (HLOP)

The method uses linear lines of position (LLOP) to replace the circular LOP for estimating the MS location. The detail algorithm of the linear LOP approach can be acquired by using the TOA measurements as in [18], and the hybrid inear LOP and AOA measurement (HLOP) in [16]. The line which passes through the intersections of the two circular LOPs can be found by squaring and subtracting the distances obtained by Eq. (1) for $i=1,2$ and can be expressed as

$$2(X_1-X_2)x+2(Y_1-Y_2)y=(r_2^2-r_1^2+X_1^2-X_2^2+Y_1^2-Y_2^2). \tag{5}$$

Given the linear LOPs and AOA line, the equations that describe all the lines can be written in matrix form as

$$Gl=h \tag{6}$$

where $l=\begin{bmatrix}x \\ y\end{bmatrix}$ denotes the MS location,

$G=\begin{bmatrix} X_1-X_2 & Y_1-Y_2 \\ \vdots & \vdots \\ X_1-X_5 & Y_1-Y_5 \\ \tan\theta & -1 \end{bmatrix}$ and

$h=\frac{1}{2}\begin{bmatrix} r_2^2-r_1^2+(X_1^2+Y_1^2)-(X_2^2+Y_2^2) \\ \vdots \\ r_5^2-r_1^2+(X_1^2+Y_1^2)-(X_5^2+Y_5^2) \\ 2(X_1\cdot\tan\theta-Y_1) \end{bmatrix}$.

According to the LS, the solution to Eq. (5) is given by

$$l=(G^TG)^{-1}G^Th \tag{7}$$

VI. Proposed Hybrid TOA/AOA Geometrical Schemes

In the TOA schemes, it is necessary to measure the propagation time it took for the signal traveling between the MS and all BSs. This time is multiplied with the speed of light to calculate the MS-BS distance. The distance can be used to form a circle and the MS lie on a circle centered at the BS. A single AOA measurement constitutes the MS along a line. The equations of the five TOA circles and the AOA line can be expressed as

$$\text{Circle 1-5: } (x-X_i)^2+(y-Y_i)^2=r_i^2, \quad i=1,2,...5 \tag{8}$$

$$\text{Line 1: } \tan\theta\cdot x-y=0 \tag{9}$$

Under the assumption of LOS propagation and there exists no measurement error, the circles intersect one single common point. However, it is very often that the LOS does not exist for propagation of signals between an MS and some fixed BSs. Therefore, the NLOS effect could cause five circles and a line to intersect at various points, which will be offset from the true MS location. With NLOS propagation, the measured TOA values are always greater than the true TOA values due to the excess path length. The true MS location should be inside the region enclosed by the overlap of the five circles. The intersections that are within this are defined as feasible intersections. The feasible intersections must satisfy the following inequalities simultaneously:

$$(x-X_i)^2+(y-Y_i)^2\le r_i^2, \quad i=1,2,...5. \tag{10}$$

Calculating the feasible intersections of five TOA circles and the AOA line will give the proximate location of the MS. In order to enhance the performance of MS location estimation with less complexity, the hybrid geometrical positioning methods which we have proposed in [13] are applied in five BSs. In comparison, for the cases presented in [13], we used two AOA measurements to eliminate the least likely intersection. Note that the region of overlap of five circles is usually smaller than that of two circles, it is not necessary to eliminate the least likely intersection.

(1) Averaging Method

The simplest and most direct method of estimating the MS location is to calculate the average value of these feasible intersections.

Step 1. Find all the feasible intersections of the five circles and the line.

Step 2. The MS location $(\bar{x}_N,\bar{y}_N)$ is estimated by averaging these feasible intersections, where

$$\bar{x}_N=\frac{1}{N}\sum_{i=1}^{N}x_i \quad \text{and} \quad \bar{y}_N=\frac{1}{N}\sum_{i=1}^{N}y_i. \tag{11}$$

(2) Distance-Weighted Method

The weights can be dynamically adjusted with reference to the distance square between the estimated MS location and the average MS location. The detailed

steps are as follows:
Steps 1-2 are the same as those of the averaging method.
Step 3. Calculate the distance d_i between each feasible intersection (x_i, y_i) and the average location $(\bar{x}_N, \bar{y}_N)$.

$$d_i = \sqrt{(x_i - \bar{x}_N)^2 + (y_i - \bar{y}_N)^2},\ 1 \le i \le N \tag{12}$$

Step 4. Set the weight for the ith feasible intersection to $(d_i^{\,2})^{-1}$. Then the MS location (x_d, y_d) is determined by

$$x_d = \frac{\sum_{i=1}^{N} \left(d_i^{\,2}\right)^{-1} \cdot x_i}{\sum_{i=1}^{N} \left(d_i^{\,2}\right)^{-1}} \quad \text{and} \quad y_d = \frac{\sum_{i=1}^{N} \left(d_i^{\,2}\right)^{-1} \cdot y_i}{\sum_{i=1}^{N} \left(d_i^{\,2}\right)^{-1}}. \tag{13}$$

(3) Threshold Method

In this method, the decision of each weight is based on how close the feasible intersections are. The closer the feasible intersections, the more weight will be assigned. The detailed steps are as fellows
Step 1. Find all the feasible intersections of the five circles and the line.
Step 2. Calculate the distance d_{mn}, $1 \le m, n \le N$, between any pair of feasible intersections.
Step 3. Select a threshold value D_{thr} as the average of all the *d*istances d_{mn}.
Step 4. Set the initial weight I_k, $1 \le k \le N$, to be zero for all feasible intersections.
If $d_{mn} \le D_{thr}$, then $I_m = I_m + 1$ and $I_n = I_n + 1$ for $1 \le m, n \le N$.
Step 5. The MS location (x_t, y_t) is estimated by

$$x_t = \frac{\sum_{i=1}^{N} I_i \cdot x_i}{\sum_{i=1}^{N} I_i} \quad \text{and} \quad y_t = \frac{\sum_{i=1}^{N} I_i \cdot y_i}{\sum_{i=1}^{N} I_i}. \tag{14}$$

(4) Sort Averaging Method

Since some of the feasible intersections are too far away from the averaged MS location, these feasible intersections may not provide improved MS location accuracy. Therefore, we proposed sort averaging method and sort-weighted method, which does not consider the influence of those far from feasible intersections.
Steps 1-3 are the same as those of the distance-weighted method.
Step 4. Rank the distances d_i in increasing order and re-label the feasible intersections in this order.
Step 5. The MS location $(\bar{x}_M, \bar{y}_M)$ is estimated by the mean of the first M feasible intersections.

$$\bar{x}_M = \frac{1}{M}\sum_{i=1}^{M} x_i,\ \bar{y}_M = \frac{1}{M}\sum_{i=1}^{M} y_i \quad (M = 0.75 * N \le N) \tag{15}$$

(5) Sort-Weighted Method

Steps 1-4 are the same as those of the sort averaging method.
Step 5. The MS location is estimated by a weighted average of the first M feasible intersections with weight $= (d_i^{\,2})^{-1}$.

$$x = \frac{\sum_{i=1}^{M} \left(d_i^{\,2}\right)^{-1} \cdot x_i}{\sum_{i=1}^{M} \left(d_i^{\,2}\right)^{-1}},\ y = \frac{\sum_{i=1}^{M} \left(d_i^{\,2}\right)^{-1} \cdot y_i}{\sum_{i=1}^{M} \left(d_i^{\,2}\right)^{-1}} \quad (M = 0.75 * N \le N) \tag{16}$$

V. Simulation Results

Computer simulations are performed to show the proposed methods is appropriate for location estimation. The distance between these BSs is $d = 3464$ m and the MS locations are uniformly distributed in the center cell, as shown in Fig. 1. 10,000 independent trials are performed for each simulation. Three different NLOS propagation models were used to model the measured ranges and angle, the circular disk of scatterers model (CDSM) [19] [20], the biased uniform random model [16] and the uniformly distributed noise model [19].

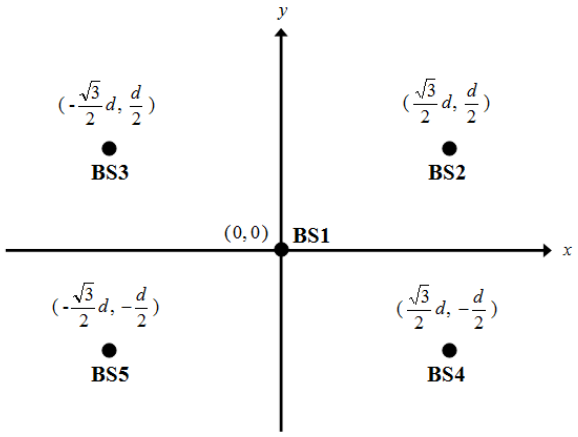


Figure 1. Five-cell system layout.

The CDSM assumes that there is a disk of scatterers around the MS and that signals traveling between the MS and the BSs undergo a single reflection at a scatterer. The BS1 serving a particular MS is called the serving BS which can provides more accurate measurements. The radius of the scatterers for BS1 is 100 m and the other BSs were taken from 100 m to 500 m. Figure 2 shows how the average location error is affected by radius of the CDSM. As the radius of the scatterers increases, the NLOS error will increase and lead to less accurate MS location estimation. By comparing the root mean square (RMS) error of location estimation, the proposed methods can predict the MS location accurately. When the NLOS errors increase, both TSA and HLOP provide relatively poor location estimation. The proposed methods can give a more accurate MS location and thus reduce the RMS errors.

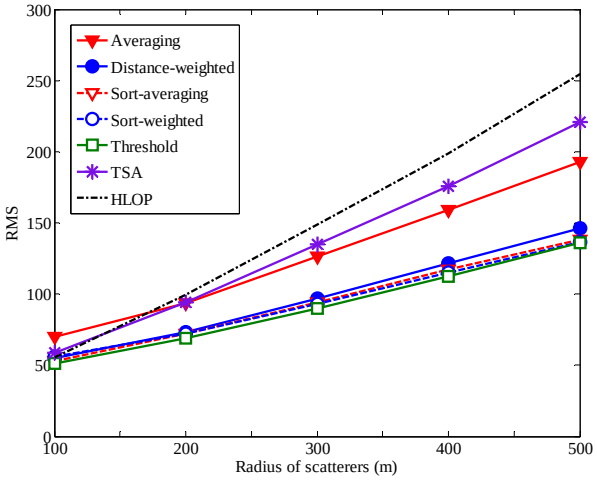


Figure 2. Effect of radius of the CDSM on the average error performance.

The improvement in MS location accuracy using the proposed method can be obtained in the cumulative distribution function (CDF) curves, as illustrated in Fig. 3. The radius of the scatterers for BS1 and the other BSs were taken to be 100m and 300m, respectively. From the simulation results, it is clear that TSA and HLOP predict the MS location with poor accuracy and the proposed methods always achieve the best performance.

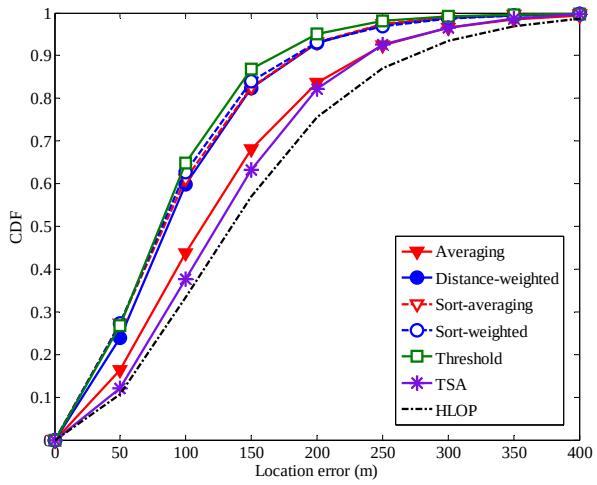


Figure 3. CDFs of the location error when CDSM is used to model the NLOS error.

The second NLOS propagation model is based on a biased uniform random variable [16], in which the measured error of TOA between the MS and BSi is assumed to be $\eta_i = p_i + u_i \cdot q_i$, where p_i and q_i are constants and u_i is a uniform random variable over [0, 1]. Similarly, the measured error of AOA, is modeled as $|f_1| = \alpha_1 + u_1 \cdot \beta_1$, where α_1 and β_1 are constants. The error variables are chosen as follows: $p_1 =$ 50m, $p_2 = p_3$ =100 m, $q_1 =$ 150 m, $q_2 = q_3$ =300 m, $\alpha_1 = 2.5°$, and $\beta_1 = 2°$. Figure 4 shows CDFs of the location error for different algorithms. It can be observed that the proposed methods can promote the location precision effectively.

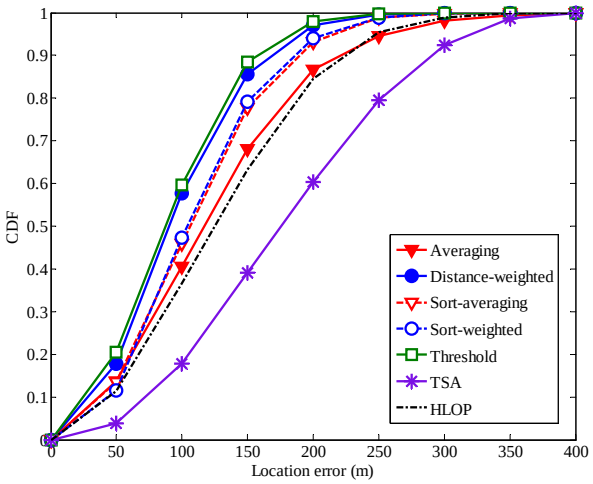


Figure 4. Comparison of error CDFs when NLOS errors are modeled as biased uniform random variables.

The final NLOS propagation model is based on the uniformly distributed noise model [19], in which the TOA measurement error is assumed to be uniformly distributed over $(0,U_i)$, where U_i is the upper bound and the AOA measurement error is assumed to be $f_1 = w_1 \cdot \tau_1$, where w_1 is a uniformly distributed variable over [-1, 1] [21]. The variables are chosen as follows: $U_1 = 200$ m, $U_i = 500$ m, for $i = 2, 3, ...5$, and $\tau_1 = 2.5°$. Figure 5 shows CDFs of the average location error of different algorithms when the range errors were using the uniformly distributed noise model. It can be seen that the proposed hybrid TOA/AOA methods provide much better location estimation as compared with other existing algorithms.

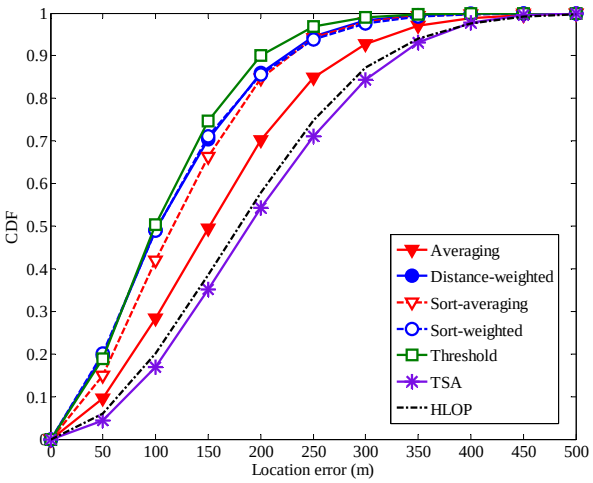


Figure 5. Comparison of error CDFs when NLOS errors are modeled as the upper bound.

Figure 6 provides the RMS error as the upper bound on uniform NLOS error increases. The upper bound for BS1 is 200 m and the other BSs are taken from 200 m to

700 m. As expected, it is observed that the location error increases with the upper bound of NLOS. The proposed methods always give better accuracy than TSA and HLOP for the error model considered. The performance degradation of the proposed methods is not pronounced under harsher NLOS error conditions.

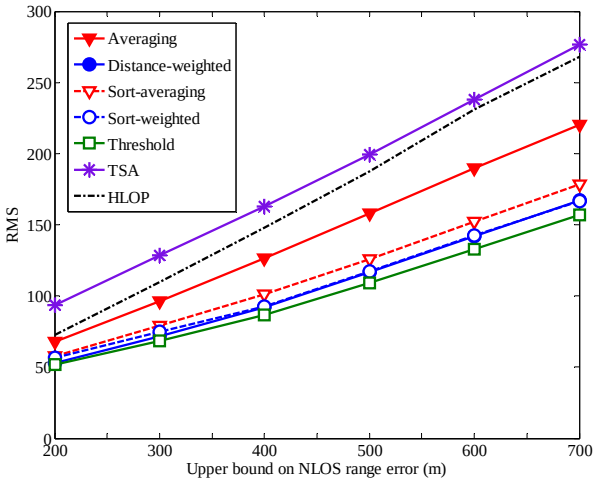


Figure 6. Performance comparison of the location estimation methods when the upper bound is used to model the NLOS.

VI. Conclusions

Based on the NLOS situation and the knowledge of NLOS error statistics is not obtained, we proposed the hybrid methods that utilize all the possible intersections of five TOA circles and the AOA line to provide the improved MS location estimation. The proposed methods mitigate the NLOS errors by the weighted sum of the feasible intersections of five circles and a line. Simulation results demonstrate that the proposed methods with different chosen weights generate more accurate MS location estimates than the conventional TSA and HLOP.

References

[1] W. G. Figel, N. H. Shepherd, and W. F. Trammell, "Vehicle location by a signal attenuation method," *IEEE Trans. Veh. Technol.*, pp. 105–109, 1969.

[2] K. J. Krizman, T. E. Biedka, and T. S. Rappaport, "Wireless position location: fundamentals, implementation strategies, and sources of error," in *Proc. IEEE Vehicular Technology Conf.*, vol. 2, pp. 919–923, May 1997.

[3] S. Al-Jazzar, J. Caffery, and H.-R. You, "A scattering model based approach to NLOS mitigation in TOA location systems," in *Proc. IEEE Vehicular Technology Conf.*, vol. 2, pp. 861–865, 2002.

[4] B. T. Fang, "Simple solution for hyperbolic and related position fixes," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 5, pp. 748–753, Sep. 1990.

[5] J. J.Caffery, and G. L. Stuber, "Overview of radiolocation in CDMA cellular systems," *IEEE Commun. Mag.*, vol. 36, no. 4, pp. 38–45, Apr.1998.

[6] T. S. Rappaport, J. H. Reed, and B. D. Woerner, "Position location using wireless communications on highways of the future," *IEEE Commun. Mag.*, vol. 34, no. 10, pp. 33–42, Oct. 1996.

[7] R. Muhamed, and T. S. Rappaport, "Comparison of conventional subspace-based DOA estimation algorithms with those employing property-restoral techniques: simulation and measurements," in *Proc. IEEE Int. Universal Personal Communications Conf.*, vol. 2, pp. 1004–1008, Oct. 1996.

[8] L. Cong and W. Zhuang, "Nonline-of-sight error mitigation in mobile location," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 560–573, Mar. 2005.

[9] M. Silventoinen and T. Rantalainen, "Mobile station emergency locating in GSM," in *IEEE Int. Conf. on Personal Wireless Communications*, pp. 232–238, Feb. 1996.

[10] L. Cong, and W. Zhuang, "Hybrid TDOA/AOA mobile user location for wideband CDMA cellular systems," *IEEE Trans. Wireless Commun.*, vol. 1, no. 3, pp. 439–447, Jul. 2002.

[11] M. A. Spirito, "Mobile station location with heterogeneous data," in *IEEE Vehicular Technology Conf.*, vol. 4, pp. 1583–1589, Sep. 2000.

[12] N. J. Thomas, D. G. M. Cruickshank, and D. I. Laurenson, "Performance of a TDOA-AOA hybrid mobile location system," in *Int. Conf. on 3G Mobile Communication Technologies*, pp. 216–220, March 2001.

[13] C.-S. Chen, S.-L. Su, and Y.-F. Huang, "Hybrid TOA/AOA geometrical positioning schemes for mobile location," *IEICE Trans. Commun.*, vol. E92-B, no. 2, pp. 396–402, Feb. 2009.

[14] W. H. Foy, "Position-location solutions by Taylor series estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-12, no. 2, pp. 187–193, Mar 1976.

[15] D. J. Torrieri, "Statistical theory of passive location systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-20, vol. 2, pp. 183-197, Mar. 1984.

[16] S. Venkatraman, and J. Caffery, "Hybrid TOA/AOA techniques for mobile location in non-line-of-sight environments," in *Proc. IEEE Wireless Communications and Networking Conf.*, vol. 1, pp. 274–278, Mar. 2004.

[17] M. McGuire, K. N. Plataniotis, and A. N. Venetsanopoulos, "Location of mobile terminals using time measurements and survey points," *IEEE Trans. Veh. Technol.*, vol. 52, no. 4, pp. 999–1011, Jul. 2003.

[18] J. J. Caffery, "A new approach to the geometry of TOA location," in *Proc. IEEE Vehicular Technology Conf.*, vol. 4, pp. 1943–1949, 2000.

[19] S. Venkatraman, J. Caffery, J., and H.-R. You, "A novel TOA location algorithm using LOS range estimation for NLOS environments," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1515–1524, Sep. 2004.

[20] P. van Rooyen, M. Lotter, and D. van Wyk, Space-time processing for CDMA mobile communications. New York: Kluwer, 2000.

[21] C.-L. Chen, and K.-T. Feng, "An efficient geometry-constrained location estimation algorithm for NLOS environments," in *Proc. Int. Conf. on Wireless Networks, Communications and Mobile Computing*, vol. 1, pp. 244–249, Jun. 2005.

Chien-Sheng Chen received the BS degree in Electrical Engineering from Feng Chia University, in 1994 and received the MS degree in Aeronautics and Astronautics from Chung Hau University, in 1996. He received the Ph.D. degree in the institute of computer and communication engineering, National Cheng Kung University in 2010. He is currently with the Tainan University of Technology. His current research interests include mobile communication networks and wireless location systems.

Yi-Jen Chiu received his BSc and MSc degrees in Electronic Engineering from Feng Chia University, Taiwan, in 1992 and from Chung Yuan Christian University, Taiwan, in 1998, respectively. Since 2009, he has been pursuing the PhD degree in Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University and has been doing research on ultra-wideband radio technologies, wireless communication. He is working in the Department of Digital Entertainment and Game Design, Taiwan Shoufu University, Tainan, Taiwan.

Ho-Nien Shou received the B.S. degree in electrical engineering from National Taiwan Institute of Technology, Taipai, Taiwan, R.O.C., in 1986 and the M.S. degree from the Department of Aeronautics and Astronautics and the Ph.D. degree in electrical engineering, both from National Cheng Kung University, Tainan, Taiwan, R.O.C., in 1990 and 2002, respectively.
From 1990 to 1991, he was with the Institute of Science and Technology (CSIST) as an assistant researcher, working with flight digital control system. From 1999 to 2001, he was with the National Space Organization (NSPO) as an assistant researcher, working with satellite attitude control system. From 2001, he was with the Department of Aviation & Communication Electronics Air Force Institute of Technology Assistant Professor. His main research interests include nonlinear system control, satellite attitude control system.

Ching-Lung Chi was born in Chayi, Taiwan, R.O.C., in 1965. He received the B.S. and M.S. degrees from Chung-Cheng Institute of Technology, Taiwan, in 1988 and 1996 and the Ph.D. degree from National Cheng Kung University, Tainan, Taiwan, in 2006, all in electrical engineering. Since 2008, he has been with She-Te University, Kaohsiung, Taiwan, where he is currently an associate Professor of computer and communication. His research interests are in the areas of wireless communications, and channel coding techniques.

Call for Papers and Special Issues

Aims and Scope.

Journal of Networks (JNW, ISSN 1796-2056) is a scholarly peer-reviewed international scientific journal published monthly, focusing on theories, methods, and applications in networks. It provide a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on networks.

The Journal of Networks reflects the multidisciplinary nature of communications networks. It is committed to the timely publication of high-quality papers that advance the state-of-the-art and practical applications of communication networks. Both theoretical research contributions (presenting new techniques, concepts, or analyses) and applied contributions (reporting on experiences and experiments with actual systems) and tutorial expositions of permanent reference value are published. The topics covered by this journal include, but not limited to, the following topics:

- Network Technologies, Services and Applications, Network Operations and Management, Network Architecture and Design
- Next Generation Networks, Next Generation Mobile Networks
- Communication Protocols and Theory, Signal Processing for Communications, Formal Methods in Communication Protocols
- Multimedia Communications, Communications QoS
- Information, Communications and Network Security, Reliability and Performance Modeling
- Network Access, Error Recovery, Routing, Congestion, and Flow Control
- BAN, PAN, LAN, MAN, WAN, Internet, Network Interconnections, Broadband and Very High Rate Networks,
- Wireless Communications & Networking, Bluetooth, IrDA, RFID, WLAN, WMAX, 3G, Wireless Ad Hoc and Sensor Networks
- Data Networks and Telephone Networks, Optical Systems and Networks, Satellite and Space Communications

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jnw/.

(Contents Continued from Back Cover)