**Journal of Networks**

ISSN 1796-2056

Volume 8, Number 1, January 2013

## Contents

### Special Issue: Selected Extended Best Papers from ICETE-2011

### Guest Editors: Mohammad S. Obaidat, Daniel Cascado-Caballero, José Luis (Sevi) Sevillano, and Joaquim Filipe

Special Issue on Selected Extended Best Papers from ICETE-2011

# Guest Editorial

This special issue of the Journal of Networks includes extended versions of extended versions of selected best papers accepted and presented at the 8th International Joint Conference on e-Business and Telecommunications (ICETE 2011). Eight papers were selected based on their excellent review scores. Their authors were invited to submit extended versions, which have undergone a second review process to guarantee that all the papers included in this Special Issue have been rigorously peer-reviewed. The accepted papers cover areas like empirical evaluations, simulation modeling and theoretical studies addressing a variety of topics related to security in telecommunications, wireless networking, and localization.

The first paper "Simplified Scheduling for Underwater Acoustic Networks" is authored by W. Van Kleunen, N. Meratnia and P. J. M. Havinga. It proposes an interesting solution to the scheduling and routing issues in underwater acoustic networks, giving a simplified set of constrains. These constrains are applied to three new scheduling algorithms: centralized for maximum throughput, distributed for low computational costs and centralized for low-latency end-to-end communications. The evaluation by simulation explores extensively the benefits obtained when the new set of constrains is applied to the proposed algorithms.

The second paper, entitled "Power Management Mechanism Exploiting Network and Video Information over Wireless Links" authored by C. Bouras, S. Charalambides, K. Stamos, S. Stroumpis and G. Zaoudis, describes some cross-layer mechanisms to manage power consumption in wireless communications. The authors focus their analysis on video transmissions. The scenarios proposed are simulated using ns-2 simulator, obtaining relevant results referred to the video quality and the associated power consumption.

In the third paper, "QoS-Aware Multipath Communications over MANETs", by M. Obaidat, M. Ali, I. Shahwan, M. S. Obaidat, and S. Obeidat, the authors propose a QoS aware routing protocol based on multipath routing. They try to minimize the packet delay providing more than one path between source and destination, and choosing the best path based on the quality of the link. The new protocol is compared against another well-known single path routing protocol (AODV) by means of simulation over OPNET. The results show that the new routing protocol outperforms AODV in terms of end-to-end delay, but at the cost of a greater protocol overhead.

The fourth paper, "Fingerprint Indoor Position System Based On Bitcloud and Openmac", authored by J. A. Gómez, A. V. Medina, E. Dorronzoro, O. Rivera and M. Merino, uses fingerprints as an alternative approach to estimate location of mobile nodes in WSNs. The fingerprints are used in two algorithms, being the second one (centroid) the algorithm that offers more accuracy. The algorithms are tested in two real implementations (bitcloud and openmac) over a specific scenario. The results show that openmac can overtake the problems detected in bitcloud and improve the precision on localization.

The fifth paper is entitled "Localization Method For Low Power Wireless Sensor Networks", authored by D. F. Larios, J. Barbancho, F. J. Molina and C. León. This paper proposes a range-free localization algorithm that uses fuzzy logic to process RSSI and estimate the position of mobile devices. This method causes less localization errors than other well-known localization methods and, at the same time, it improves power consumption of the mobile nodes.

The sixth paper, "Development Tools For Context Aware And Secure Pervasive Computing in Embedded Systems (PECES) Middleware", authored by R. Zhao, K. Selvarajah and N. Speirs presents a set of tools to enable the communication among heterogeneous devices across multiple smart devices and platforms in a secure manner. PECES and its tools are especially useful for experienced developers of services and applications that desire a clean deployment of their applications regardless of the platform for which they are developing. PECES has been evaluated by means of heuristic techniques that conclude that the majority of interviewed people assert that PECES is useful, and easy to use.

The seventh paper "Performance of OpenDPI in Identifying Sampled Network Traffic", authored by J. Khalife, A. Hajjar and J. Diaz, studies the impact of sampling traffic on the performance of OpenDPI, by means of two different ways of gathering information: per-packet sampling and per-flow sampling, in order to determine the reduction of the input data and the accuracy of the classification. Results show that the per-flow packet is more convenient for sampling; showing very high accuracy rates.

The last paper entitled, "Anomaly Detection Using Metaheuristic Firefly Harmonic Clustering", authored by M. H. A. C. Adaniya, T. Abrão and M. Lemes Proença presents a clustering algorithm based on K-Harmonic means (KHM) and Firefly Algorithm (FA) for the detection of network volume anomalies due to device failing or misconfiguration. As a consequence of the application to the real data collected from a proxy server, and a further statistic analysis, the algorithm achieves about 80% of true positive recognitions and 20% of false positive recognitions, demonstrating the satisfactory behaviour of this algorithm.

The guest editors would like to thank all the authors and reviewers for their valuable contributions to this special issue. We also thank the editorial staff of Journal of Networking for their support. We hope that the papers selected in this special issue will become useful resources for researchers and practitioners in these areas.


doi: 10.4304/jnw.8.1.1-3

## Guest Editors:

**Mohammad S. Obaidat**, Computer Science and Software Engineering, Monmouth University, W. Long Branch, NJ 07764, USA
E-mail: obaidat@monmouth.edu

**Daniel Cascado-Caballero**, Department of Computer Architecture, University of Seville, Seville, Spain
E-mail: danic@atc.us.es

**José Luis (Sevi) Sevillano**, Department of Computer Architecture, University of Seville, Seville, Spain

**Joaquim Filipe**, Polytechnic Institute of Setúbal, Escola Superior de Tecnologia de Setúbal, Rua do Vale de Chaves – Estefanilha, 2910-761 Setúbal, Portugal
E-mail: jfilipe@insticc.org

**Professor Mohammad S. Obaidat** (Fellow of IEEE and Fellow of SCS) is an internationally well-known academic/researcher/ scientist. He received his Ph.D. and M. S. degrees in Computer Engineering with a minor in Computer Science from The Ohio State University, Columbus, Ohio, USA. Dr. Obaidat is currently a full Professor of Computer Science at Monmouth University, NJ, USA. Among his previous positions are Chair of the Department of Computer Science and Director of the Graduate Program at Monmouth University and a faculty member at the City University of New York. He has received extensive research funding and has published over Ten (10) books and over Five Hundred and fifty (550) refereed technical articles in scholarly international journals and proceedings of international conferences, and currently working on three more books. Professor Obaidat has served as a consultant for several corporations and organizations worldwide. Mohammad is the Editor-in-Chief of the Wiley International Journal of Communication Systems, the FTRA Journal of Convergence and the KSIP Journal of Information Processing. He served as an Editor of IEEE Wireless Communications from 2007-2010. Between 1991-2006, he served as a Technical Editor and an Area Editor of Simulation: Transactions of the Society for Modeling and Simulations (SCS) International, TSCS. He also served on the Editorial Advisory Board of Simulation. He is now an editor of the Wiley Security and Communication Networks Journal, Journal of Networks, International Journal of Information Technology, Communications and Convergence, IJITCC, Inderscience. He served on the International Advisory Board of the International Journal of Wireless Networks and Broadband Technologies, IGI-global. Prod. Obaidat is an associate editor/ editorial board member of seven other refereed scholarly journals including two IEEE Transactions, Elsevier Computer Communications Journal, Kluwer Journal of Supercomputing, SCS Journal of Defense Modeling and Simulation, Elsevier Journal of Computers and EE, International Journal of Communication Networks and Distributed Systems, The Academy Journal of Communications, International Journal of BioSciences and Technology and International Journal of Information Technology. He has guest edited numerous special issues of scholarly journals such as IEEE Transactions on Systems, Man and Cybernetics, SMC, IEEE Wireless Communications, IEEE Systems Journal, SIMULATION: Transactions of SCS, Elsevier Computer Communications Journal, Journal of C & EE, Wiley Security and Communication Networks, Journal of Networks, and International Journal of Communication Systems, among others. Obaidat has served as the steering committee chair, advisory Committee Chair and program chair of numerous international conferences.

He is the founder of two well-known international conferences: The International Conference on Computer, Information and Telecommunication Systems (CITS) and the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS). He is also the co-founder of the International Conference on Data Communication Networking, DCNET. Between 1994-1997, Obaidat has served as distinguished speaker/visitor of IEEE Computer Society. Since 1995 he has been serving as an ACM distinguished Lecturer. He is also an SCS distinguished Lecturer. Between 1996-1999, Dr. Obaidat served as an IEEE/ACM program evaluator of the Computing Sciences Accreditation Board/Commission, CSAB/CSAC. Obaidat is the founder and first Chairman of SCS Technical Chapter (Committee) on PECTS (Performance Evaluation of Computer and Telecommunication Systems). He has served as the Scientific Advisor for the World Bank/UN Digital Inclusion Workshop- The Role of Information and Communication Technology in Development. Between 1995-2002, he has served as a member of the board of directors of the Society for Computer Simulation International. Between 2002-2004, he has served as Vice President of Conferences of the Society for Modeling and Simulation International SCS. Between 2004-2006, Prof. Obaidat has served as Vice President of Membership of the Society for Modeling and Simulation International SCS. Between 2006-2009, he has served as the Senior Vice President of SCS. Between 2009-2011, he served as the President of SCS. One of his recent co-authored papers has received the best paper award in the IEEE AICCSA 2009 international conference. He also received the best paper award for one of his papers accepted in IEEE GLOBCOM 2009 conference. Prof. Obaidat has been awarded a Nokia Research Fellowship and the distinguished Fulbright Scholar Award. He received the SCS Outstanding Service Award for his excellent leadership, services and technical contributions. Dr. Obaidat received very recently the Society for Modeling and Simulation Intentional (SCS) prestigious McLeod Founder's Award in recognition of his outstanding technical and professional contributions to modeling and simulation. He received in Dec 2010, the IEEE ComSoc- GLOBECOM 2010 Outstanding Leadership Award for his outstanding leadership of Communication Software Services and Multimedia Applications Symposium, CSSMA 2010. He received very recently the Society for Modeling and Simulation International's (SCS) prestigious Presidential Service Award for his outstanding unique, long-term technical contributions and services to the profession and society.

He has been invited to lecture and give keynote speeches worldwide. His research interests are: wireless communications and networks, telecommunications and Networking systems, security of network, information and computer systems, security of e-based systems, performance evaluation of computer systems, algorithms and networks, green ICT, high performance and parallel computing/computers, applied neural networks and pattern recognition, adaptive learning and speech processing. During the

2004/2005, he was on sabbatical leave as Fulbright Distinguished Professor and Advisor to the President of Philadelphia University in Jordan, Dr. Adnan Badran. The latter became the Prime Minister of Jordan in April 2005 and served earlier as Deputy Director General of UNESCO. Prof. Obaidat is a Fellow of the Society for Modeling and Simulation International SCS, and a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). For more info; se: http://bluehawk.monmouth.edu/mobaidat/.

**Dr. Daniel Cascado-Caballero** is assistant professor in the University of Seville. He received his degree in Computer Science in 1996 and his PhD. in 2003, both from the University of Seville. From 2009 until now, he has server as publication Chair in the following conferences: ACS/IEEE AICCSA2009, SPECTS 2009, SPECTS 2010. SPECTS 2011, SPECTS 2012, CITS 2012. His research field is focused in wireless communications, simulation, network traffic analysis, and e-health, where he has numerous papers published in referred international journals and conferences.

**Dr. José Luis (Sevi) Sevillano** received his degree in Physics (electronics) and his Ph.D. from the University of Seville (Spain) in 1989 and 1993 respectively. From 1989 to 1991 he was a researcher supported by the Spanish Science and Technology Commission (CICYT). After being Assistant Professor of Computer Architecture at the University of Seville, since 1996 he is Associate Professor at the same University. He has served as Vice Dean of the Computer Engineering School (2004-7) and as Director of Innovations for Teaching (2007-8) at the University of Seville. Currently, he is Coordinator of the Telefónica Chair on Intelligence in Networks, University of Seville, Spain.

Since 2007 Prof. Sevillano is Associate Editor of the International Journal of Communication Systems, published by John Wiley. He is also Associate Editor of Simulation (Sage). He also served as Vice-President for Membership of The Society for Modeling & Simulation International (SCS) (2009-2011). He has served on several international conferences: as General Chair (SPECTS-11, ICETE 2011), as Program Co-Chair (ACS/IEEE AICCSA 2009, DCNET 2010, SPECTS-2009, SPECTS-2010), as well as member of the TPC. He is also a member of the Steering Committee of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS). One of his co-authored papers received the Best Paper award of the 13th Communications & Networking Simulation Symposium (CNS 2010). He is author/co-author of more than 60 research reports and papers in refereed international journals and conferences, and has participated in more than 20 research projects and contracts.

**Joaquim B L Filipe** is currently a Coordinator Professor and Head of the Department of Systems and Informatics of the School of Technology of the Polytechnic Institute of Setúbal (EST-Setúbal). He got his M.Sc. in 1989 by the Technical University of Lisbon, an M.B.A., awarded in 1995 by the New University of Lisbon and a Ph.D. awarded by the School of Computing of the Staffordshire University, U.K, in 2000. His main areas of research involve Artificial Intelligence and Multi-Agent System theory and applications to different domains, with an emphasis on social issues in activity coordination, especially in organizational modeling and simulation, where he has been actively involved in several R&D projects, including national and international programs. He represented EST-Setúbal in several European projects. He has over 80 publications including papers in conferences and journals, edited books and conference proceedings. He started the ICEIS conference series, in 1999 and is conference chair since then. As President of INSTICC, the Institute for Systems and Technologies of Information, Control and Communication, he also launched several international conferences, including ICETE (the Int'l Conference on e-Business and Telecommunications). He participated in more than 40 conference and workshop program committees and he is in the editorial board of several international scientific journals.

# Simplified Scheduling for Underwater Acoustic Networks

Wouter van Kleunen, Nirvana Meratnia, Paul J.M. Havinga
Pervasive Systems, University of Twente
7522 NB Enschede, The Netherlands
Email: {w.a.p.vankleunen, n.meratnia, p.j.m.havinga}@utwente.nl

***Abstract*—The acoustic propagation speed under water poses significant challenges to the design of underwater sensor networks and their medium access control protocols. Similar to the air, scheduling transmissions under water has significant impact on throughput, energy consumption, and reliability.**

**In this paper we present an extended set of simplified scheduling constraints which allows easy scheduling of underwater acoustic communication. We also present two algorithms for scheduling communications, i.e. a centralized scheduling approach and a distributed scheduling approach. The centralized approach achieves the highest throughput while the distributed approach aims to minimize the computation and communication overhead. We further show how the centralized scheduling approach can be extended with transmission dependencies to reduce the end-to-end delay of packets.**

**We evaluate the performance of the centralized and distributed scheduling approaches using simulation. The centralized approach outperforms the distributed approach in terms of throughput, however we also show the distributed approach has significant benefits in terms of communication and computational overhead required to setup the schedule.**

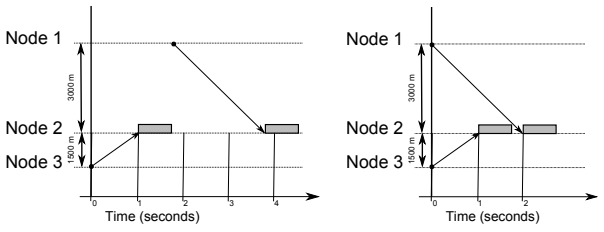
**We propose a novel way of estimating the performance of scheduling approaches using the ratio of modulation time and propagation delay. We show the performance is largely dictated by this ratio, although the number of links to be scheduled also has a minor impact on the performance.**

## I. Introduction

Acoustic communication is the most widely used type of communication for underwater networks. This is because acoustic communication is the only form of communication which allows long-range communication in underwater environments. Acoustic communication, however, poses its own set of challenges for the design of networking and communication protocols. The slow acoustic propagation speed of about 1500 m/s, limited available bandwidth, high transmission energy costs and variations in channel propagation are some of the challenges to overcome.

Examples of existing underwater MAC protocols include T-Lohi [1], Slotted-FAMA [2], and ST-MAC [3]. Scheduled communication approaches, such as ST-MAC [3] and STUMP [4] these have significant benefits over unscheduled approaches such as ST-MAC [3] or ALOHA [5] [6]. These benefits include improved success rate due to the avoidance of packet collision, reduced energy-consumption and improved throughtput. All scheduled based approaches use estimation of the propagation delay to schedule the reception of the packet.

Because of the slow propagation speed and the resulting large propagation times of the signal an uncertainty of the global state of the channel exists, this is called the space-time uncertainty [1]. Because of the spatial-temporal uncertainty, exclusive access to the medium is not required for collision free communication. Rather transmission times should be scheduled such that no collision occurs at reception. Figure 1 shows how two packets can be transmitted at the same time but are received without collision at the receiver. By exploiting the fact that we can have an estimation of the propagation delay, several transmissions can be scheduled at the same time as long as the reception of the packet is scheduled without interference.




(a) Exclusive access (b) Scheduled

Fig. 1. Exploiting spatial-temporal uncertainty in underwater communication with scheduling

Scheduling approaches such as ST-MAC [3] and STUMP [4] are able to exploit this spatial-temporal uncertainty but do so at the cost of complex scheduling algorithms. In [7] we have shown how to derive a simplified set of constraints which greatly simplifies the scheduling of underwater communication. In [8] we have shown how this set of constraints can be used to schedule large-scale networks and in [9] we have evaluated the performance of scheduled communication for an underwater localization system.

In this paper we will review the set of simplified constraints and present a centralized and distributed scheduling approach. We will also extend the centralized scheduling approach with scheduling dependencies which allows enforcing a certain order of transmission. We show that these scheduling constraints can be used to reduce the

This work is supported by the SeaSTAR project funded by the Dutch Technology Foundation (STW).


doi: 10.4304/jnw.8.1.4-14

end-to-end delay of packets in a multihop data collection network by delay the transmission of a parent node until all the packets of children nodes are received.

For performance evaluation, we compare the centralized and distributed scheduling approaches using simulation. We will show that the distributed scheduling approachehas significant benefits in terms of communication and computational overhead, while the centralized approach is able to achieve the highest throughput. We also show that scheduling constraints can be used to reduce the end-to-end delay of packets. We present a novel way to estimate the performance of scheduling approaches using the ratio of modulation time and scheduling time. Finally we evaluate broadcast scheduling and compare its performance with ALOHA.

## II. Related work

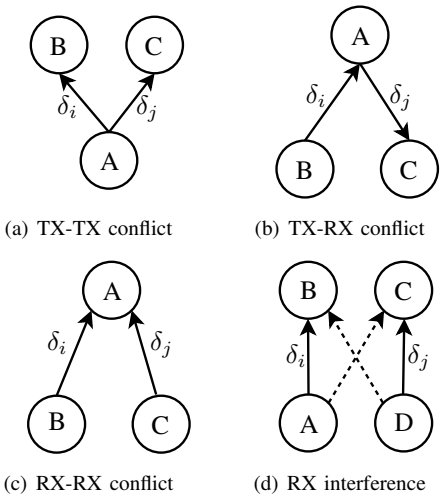
Scheduling communication in underwater communication is done through scheduling the reception of a packet in such a way that it is received without interference from other transmissions active within the network. To do so, the scheduling algorithm needs to know all transmissions and all nodes within the network beforehand and should be able to make an estimation of the propagation delay of the acoustic signal between two nodes. The propagation delay can be estimated by calculating the distance between two nodes using the position information. This distance divided by the propagation speed of the signal ( $1500m/s$) results in the propagation delay. Another approach would be to measure the propagation delay at runtime.

Because the propagation delay needs to be estimated and all transmissions should be known before scheduling the transmissions, scheduled communication is most suited for static networks. This is also because the benefits of using a schedule should outweigh the overhead of setting up a schedule. This can usually been done only when the schedule stays valid for a long period of time.

Scheduling algorithms schedule the transmission in time and therefore some form of time-synchronisation is required. This can be done by using a very accurate clock on the nodes or using some form of dynamic time-synchronisation. Time-synchronisation and position estimation have been researched extensively. An example of a time-synchronisation protocol is TSHL [10] and an overview of localization approach can be found in [11].

Scheduling communication can be done using different approaches. For example the scheduling can be done using a slotted approach, such as used by ST-MAC [3] and STUMP [4] or using unslotted approach as employed in our Simplified Scheduling approach [7]. Another difference between scheduling approaches is that they are done using a centralized approach (e.g. ST-MAC [3] and Simplified Scheduling [7]), using a distributed approach (e.g. STUMP [4]) and our Simplified Scheduling approach for large-scale networks [8].

All scheduling algorithms use similar scheduling constraints to model the possible conflicts that may arise. Figure 2 shows these possible conflicts. Our approach to scheduling communication is different from existing approaches because we have simplified these scheduling constraints into a simplified set of scheduling constraints. This allows development of considerably simpler scheduling algorithms.




(a) TX-TX conflict (b) TX-RX conflict

(c) RX-RX conflict (d) RX interference

Fig. 2. Illustration of all possible scheduling conflicts

## III. The set of simplified scheduling constraints

In this section we will review the set of simplified scheduling constraints. We will simplify this set of constraints further than what was described in [7] and we will also add support for scheduling broadcasts and scheduling in large-scale networks.

The scheduling constraints place restrictions on the transmissions start time of the "to be scheduled" transmissions. We denote the transmissions tasks as $\delta$ where a single transmission task $i$ from the complete set of transmission tasks is denoted as $\delta_i$. For each transmission we need to calculate the transmission start time $\delta_i.start$. Every transmission has a certain duration $\delta_i.duration$, source $\delta_i.src$ and destination $\delta_i.dst$. We assume the function $T$ will give the transmission delay between two nodes. This function can be implemented by calculating the distance between two nodes and using the estimated propagation speed to calculate the propagation delay.

The set of simplified scheduling constraints we have derived in [7] has been shown in Figure 3. We can simplify this set of constraints further by making the following observation: When the source of transmission $i$ is the same as the source of transmission $j$ then the equation (1) can be rewritten as equation (2).

$$T(\delta_i.src, \delta_i.dst) - T(\delta_j.src, \delta_i.dst), \tag{1}$$

$$T(\delta_i.src, \delta_i.dst) - T(\delta_i.src, \delta_i.dst), \tag{2}$$

Equation 2 will always evaluate to 0. The same can be done for the second equation of the maximum from the

$$\text{given } j \text{ for all } i < j, \begin{cases} \delta_j.start \geq \delta_i.start + \delta_i.duration & \text{if } \delta_i.src = \delta_j.src \\ \delta_j.start \geq \delta_i.start + \delta_i.duration + max( & \\ \quad T(\delta_i.src, \delta_i.dst) - T(\delta_j.src, \delta_i.dst), & \text{if } \delta_i.src \neq \delta_j.src \\ \quad T(\delta_i.src, \delta_j.dst) - T(\delta_j.src, \delta_j.dst)) & \end{cases}$$

Fig. 3. Set of simplified scheduling constraints

second rule from the simplified set of constraints. This makes the maximum term of the second rule to be the $max(0, 0)$ when $\delta_i.src = \delta_j.src$ and makes rule one not required.

In [8] we have extended the set of simplified scheduling constraints with an interference condition. This allows scheduling of large-scale networks where nodes may be outside of each others interference range. Two nodes are outside of interference range of each other if the signal of one node results in a received signal strength on the other node which is below a certain threshold ($TH_{cp}$). The value of this threshold ($TH_{cp}$) should be chosen in such a way that interfering signals are always below the receiver sensitivity of the node or the interfering signal can be guaranteed to be captured by the transmission.

The received signal strength is dependant on the output power of the sender and the attenuation between the sender and the receiver. The attenuation between nodes depends on the absorption rate of the water and the spreading of the signal. This path loss equation [12] can be written as follows:

$$10 \log(d, f) = k \cdot 10 \log d + d \cdot 10 \log a(f) \qquad (3)$$

The path loss depends on the carrier frequency ($f$) of the signal as well as the distance ($d$) between sender, and receiver. The spreading factor is constant, which can either be spherical ($k = 2$), cylindrical ($k = 1$), or something in between. The frequency dependant attenuation is given by the function $a(f)$.

Using this formula we can calculate whether two nodes interfer with each other. Consider two transmissions $\delta_i$ and $\delta_j$, which both have source ($\delta_i.src$ and $\delta_j.src$) and destination ($\delta_i.dst$ and $\delta_j.dst$). We will use the path loss function ($PL$) to calculate the difference of the received signal strengths at the destination of transmissions ($\delta_j$):

$$Interfer(\delta_i, \delta_j) = TRUE \text{ if } PL(\delta_j.src, \delta_j.dst) - PL(\delta_i.src, \delta_j.dst) \leq TH_{cp} \qquad (4)$$

Function (4) will return $false$ if transmission $\delta_i$ does not cause interference for transmissions $\delta_j$. We will now show how this equation can be applied to the set of simplified scheduling rules. The interference rule only applies when two nodes are able to interfer with each others transmissions. If $\delta_i.src$ is out of range of $\delta_j.dst$ and if $\delta_j.src$ is out of range of $\delta_i.dst$, there is no interference and therefore no constraint between the two transmissions.

Finally we will review how scheduling of broadcast messages can be added to the set of simplified scheduling constraints. A broadcast message should be scheduled in such a way that on all positions within the network the message can be correctly received. In other words, no collision should occur at any position in the network.

This can be done as follows: when node A broadcasts its message, node B will have to wait until the message of node A passes. Node B can then starts its transmission. When node B transmits immediately after the message from node A has passed node B, the propagation circle of the message from B will always stay within the propagation circle of node A. This means that on any position within the network both messages can be received without any interference. An example of this is shown in Figure 4.

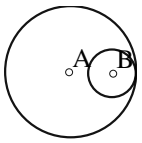



Fig. 4. An example of how two broadcasts can be transmitted without collisions

To put this in a scheduling constraint, node B will have to delay its transmission until the message from A has propagated to the position of B. Assuming the position of A (denoted as $\delta_i.src$) is the source of transmission $\delta_i$, and position of B (denoted as $\delta_j.src$) is the source of transmission $\delta_j$ and assuming that we can calculate the propagation time between two positions using the unspecified function $T$ (For example $T$ calculates the Euclidean distance between the two positions divided by an estimation of the sound speed under water), the minimum delay between transmission $\delta_i$ and $\delta_j$ can now be calculated as:

$$\delta_i.duration + T(\delta_i.src, \delta_j.src)$$

The complete set of simplified scheduling constraints, given in Figure 5, consists now of the following three scheduling rules:

1) If any of the two transmissions is a broadcast, the broadcast scheduling constraint should be used.
2) If both transmissions are unicasts, the interference rule from [7] should be used. This rule ensures that the second unicast arrives at the receiver when the first unicast has been received completely.

3) If both transmissions are unrelated, both can be scheduled at the same time.

## IV. SCHEDULING ALGORITHMS

The extended set of simplified constraints, described in Section III, can be applied to design a scheduling algorithm with low complexity. In this section we present the following three algorithms:

- A centralized scheduling approach with high throughput. This approach assumes all information is collected at a central node *prior* to scheduling. Because all information is available at a single point this approach will achieve the *highest throughput*.
- A distributed scheduling approach with *low computational and communication complexity* for large-scale underwater networks. This approach provides a trade-off between the efficiency of the resulting schedule and the amount of communication and computation required to setup the schedule.
- A centralized scheduling approach with transmission dependencies for *low-latency end-to-end communication*. While the first centralized scheduling approach optimizes for througput, this approach shows how transmission dependencies can be used for optimizing the end-to-end delay.

The centralized scheduling algorithm is a reduced complexity version of the algorithm shown in [7]. The algorithm tries to determine a schedule with minimal time length and therefore aims to achieve the highest throughput possible. The distributed algorithm uses the first algorithm to schedule clusters and allows scheduling of large-scale networks with reduced complexity. Finally the last algorithm is a centralized scheduling approach that uses dependencies between transmissions to optimize the end-to-end delay over multihop communication. Rather than trying to achieve the highest throughput possible, this approach minimizes the average time it takes for all packets to travel over multiple hops.

We have chosen these three approaches because we believe they can be used in existing and future underwater sensor network applications. The first approach provides a simple approach for scheduling small-scale networks. The second approach shows how scheduling can be done in large-scale networks and provides a balance between network setup overhead and throughput. The last approach can be used to reduce the end-to-end delay of packets or for applications requiring certain transmissions order such as data-aggregation and other distributed processing approaches.

### A. A centralized scheduling approach with high throughput

The extended set of simplified constraints can be applied to design a scheduling algorithm with low complexity for underwater networks. The algorithm from [7], which has $O(n^3)$ complexity, considers every transmission as the first transmission. To reduce the complexity, we can take the first transmission as the transmission to be scheduled at time 0. This will reduce the complexity of the algorithm from $O(n^3)$ to $O(n^2)$.

$V \leftarrow transmissions$ {Set of all transmissions}
$schedule \leftarrow [N] = 0$ {Resulting schedule}
$schedule[0] = 0$ {Schedule the first transmission}
$time = 0$
$last = 0$
$V \leftarrow V \setminus \delta_0$ {Remove transmission from set}
{Scheduling loop schedules transmissions greedy}
**while** $!empty(V)$ **do**
  $time_{min} \leftarrow infinity$
  {Calculate minimum starting time for remaining transmissions}
  **for** $\delta \in V$ **do**
    $schedule[\delta_{index}] = max(schedule[\delta_{index}], time + constraint(\delta_{last}, \delta_{index})]$
    {See if this transmission has the smallest starting time}
    **if** $schedule[\delta_{index}] < time_{min}$ **then**
      $time_{min} \leftarrow schedule[\delta_{index}]$
      $index \leftarrow \delta_{index}$
    **end if**
  **end for**
  {Schedule transmission with smallest starting time first}
  $time = time_{min}$
  $last = index$
  $V \leftarrow V \setminus \delta_{index}$
**end while**

Fig. 6. Reduced complexity algorithm for scheduling transmissions.

The algorithm initially schedules the first transmission. Inside the scheduling loop first all the minimum starting times for the remaining transmissions are calculated. The loop also finds the transmission with the minimum schedule time and removes this transmission from the set of "to be scheduled" transmissions. This is repeated until all transmissions are scheduled.

The algorithm continously updates the start time for the unscheduled transmissions. This is done by calculating the maximum of the previously calculated start time and the new start time calculated using the scheduling constraints. It ensures collision free reception by taking the maximum start transmission time.

When we calculate the schedule only once, there is also no need anymore to precalculate a table of delays for all transmission pairs. Any transmission pair will be considered at most once, but some will never be calculated. At the first iteration the algorithm will calculate the delays for $n-1$ pairs, in the second iteration for $n-2$, and so forth. This will further reduce the complexity from $O(n^2)$ to $O(\frac{1}{2}n^2)$. Because we do not calculate the delay table, the memory space complexity can also be reduced to $O(n)$.

The full algorithm can be seen in Figure 6.

### B. A distributed scheduling approach with low computational and communication complexity

The algorithm presented in Section IV-A requires multi-hop communication to gather information about all required transmissions within the network. This has a significant overhead and because it is done before

$$\begin{cases} \delta_j.start \geq \delta_i.start + \delta_i.duration + T(\delta_i.src, \delta_j.src) & \text{if } (\delta_i.dst = broadcast \text{ or } \delta_j.dst = broadcast) \text{ and } Interfer(\delta_i.src, \delta_j.dst) \\ \delta_j.start \geq \delta_i.start + \delta_i.duration + max( & \\ \quad T(\delta_i.src, \delta_i.dst) - T(\delta_j.src, \delta_i.dst), & \text{if } Interfer(\delta_i.src, \delta_j.dst) \\ \quad T(\delta_i.src, \delta_j.dst) - T(\delta_j.src, \delta_j.dst)) & \\ \delta_j.start \geq \delta_i.start & \text{otherwise} \end{cases} \tag{5}$$

Fig. 5. Extended set of simplified scheduling constraints allowing broadcast scheduling.

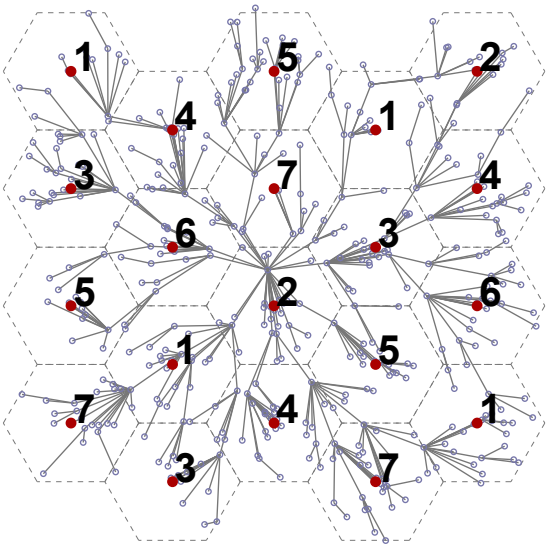


Fig. 7. Example of a deployment.

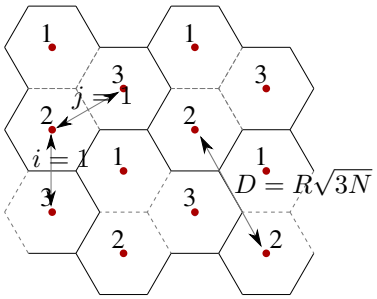



Fig. 8. Cellular network example.

scheduling, this communication will be done in an unscheduled way.

To reduce this communication overhead, we propose a distributed scheduling approach based on a clustering concept. We propose a technique in which cluster-heads are time-schedule arbiters for a cluster and nodes will send a request to the cluster-head to do a communication. The clusters are assigned a timeslot, which can span up to several seconds and will schedule all the requested transmissions in their timeslot. The timeslots can be reused in other clusters and this will ensure that minimal interference occurs between clusters.

Figure 7 illustrates an example deployment setup. The cluster-heads are in the center of their cluster and the numbers shown in the cluster indicate the used timeslot of the cluster. The small dots are sensor nodes scattered across the complete deployment area and the lines between nodes indicate communication links. Communication does not necessarily have to be done from or towards the cluster-heads and can be done to any node within the communication range. The links are set up in such a way that information is collected at a central sink.

The size of the clusters is dependant on the communication range of the nodes. We assume that all nodes in the network use the same output power and carrier frequency for transmissions and will therefore have the same communication range. All nodes within the cluster should be able to communicate with the cluster-head, therefore the cluster size should not be bigger than the communication range. We assume the radius of the cluster is exactly the size of the maximum communication range. The actual size can be calculated using the path loss expressed in Equation (3).

The clusters in our approach are similar to cells in a cellular network. If we assume that the shape of a cluster in our approach is hexagonal, we can then use the equations from cellular networks to calculate the number of timeslots required. The number of timeslots determines the reuse distance. One may recall that the reuse distance is the minimum distance between two clusters that share the same timeslot, see Figure 8 for an example where the number of timeslots can not arbitrarly be chosen and is determined from the following formula:

$$N = i^2 + ij + j^2 \tag{6}$$

The $i$ and $j$ parameters determine the reuse distance of a timeslot along two axises. The reuse distance ($D$) can be calculated from the number of cells per cluster ($N$) and the cell radius ($R$):

$$D = R\sqrt{3N} \tag{7}$$

The reuse distance is the minimum distance between two interfering senders in the network. The larger the distance between two interferers, the lesser interference will be experienced during communication. If a total of 3 timeslots are used, the closest distance between two interfering nodes is exactly the radius of the cluster. If more timeslots are used, the distance between two

| | Cluster | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | Max |
| Slot 1 | 1.33 | | | 1.57 | | | 1.57 |
| Slot 2 | | 1.61 | | | 1.43 | | 1.61 |
| Slot 3 | | | 1.37 | | | 1.45 | 1.45 |
| Slot length | 1.57 | 1.61 | 1.45 | 1.57 | 1.61 | 1.45 | |

Fig. 9. Results of calculating slot length based on cluster schedule lengths.

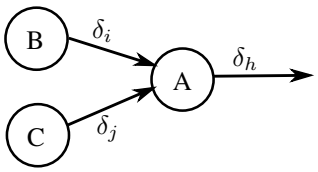



Fig. 10. Multihop scheduled communication.

interfering nodes will be larger, resulting in less noise from neighbouring clusters.

The nodes within a cluster all register their transmissions to the closest cluster-head. The cluster-head is therefore able to schedule all the transmissions within its cluster. After doing so, it will send the minimum length of its local schedule to the central cluster-head. The central cluster-head will assign timeslots to the clusters and determine the length of each timeslot. The timeslots do not necessarily have to be of equal time. The central cluster-head will assign the maximum schedule length of all clusters that share the same timeslot.

The cluster-heads will determine the order of transmissions within their cluster. This can be done using different optimization criteria as presented in [7]. We will be using the greedy approach in which transmissions are scheduled based on the minimum delay.

For scheduling the transmissions within a cluster we can use the algorithm from [7] or the reduced complexity algorithm from Section IV-A. The algorithm presented in Section IV-B will yield a smaller computational and memory space complexity, but because the number of transmissions per cluster is in practice limited, the algorithm presented in Section IV-A may as well be a good option.

Figure 9 shows an example of how the algorithm works. The table shows for all clusters the calculated cluster schedule lengths. The cluster-head schedules all transmissions within its cluster and determines the clusters schedule length. The central cluster-head determines the maximum of all schedule lengths per slot and assigns the maximum schedule length to the slot. The schedule length and slot lengths are then the only information the central cluster-head needs to communicate to the other cluster-heads.

### *C. A centralized scheduling approach with transmission dependencies*

The following algorithm is an extension of the approach shown in Section IV-A. This approach uses dependencies to force an ordering in the scheduling of transmissions. This can be used to reduce the average end-to-end delay for sending packets over multihop connections. An example in which dependencies can be used is shown in Figure 10. In this scenario two nodes (A and B) are sending their packets towards node C and node C is forwarding the (aggregated) packet. In this example it would be benificial for the end-to-end delay to first schedule the transmissions from A and B in such a way that they are received by C before C starts transmitting. If C receives the packets from A and B after it has transmitted its packet it will have to wait for a new iteration of the schedule to forward the packets from A and B.

To add scheduling to the algorithm of Section IV-A, we first have to define a list of dependencies. We store the dependencies in a list $D$ and every dependency is a tuple defining a dependency between two transmissions. In the example shown above, transmission $\delta_h$ will have to be transmitted after the reception of transmission $\delta_i$ and $\delta_j$. The dependency list $D$ therefore constains tuples $(\delta_i, \delta_h)$ and $(\delta_j, \delta_h)$.

Our algorithm for scheduling with dependencies is shown in Figure 11. When scheduling with dependencies, the options of possible transmissions are limited by the dependencies. When a transmission still has dependencies it is considered blocked and can not be scheduled. This transmission still has a tuple in the dependency list $D$. Only the transmissions without dependencies are considered as next transmission and from all these transmissions we greedily select the next transmission. Once we schedule a new transmission we can remove the tuples of dependent transmissions from the dependency list $D$, thereby freeing or unblocking possible new transmissions.

The first transmission we schedule is a transmission which has no dependencies. To find this transmission we go through the list of transmissions and find an entry which has no dependency entry in the list $D$, we do this by finding a transmission $\delta$, which does not have any tuple in D: $(*, \delta) \notin D$. After this, we schedule the first transmission and remove all dependencies for this transmission. Once we have selected the first transmission, we start scheduling the minimum next transmission with no dependency. While determining the minimum transmission to be scheduled we only consider transmissions which have no dependencies. Once the transmission is scheduled, we remove all dependencies related to this transmission. This frees up new transmissions which can be considered during a next round of the scheduling algorithm. We continue until all transmissions are scheduled. Because at every round we consider only transmissions which have no dependencies in the dependency list $D$, we ensure an ordering of the transmissions.

The dependencies can be derived from the routing algorithm. For example in a data-collection network all data is routed towards a lower hop node until it reaches a single central node in the network. When the parent for a node is selected by the routing algorithm, a transmission is generated from the node to the parent. Next to the transmission also a dependency for this transmission with

```
V ← transmissions {Set of all transmissions}
D ← dependencies {Set of all dependencies}
schedule ← [N] = 0 {Resulting schedule}
{Find the first transmission that can be scheduled}
for δ ∈ V do
  if (∗, δ) ∉ D then
    index ← δ_index
  end if
end for
schedule[index] = 0 {Schedule the first transmission}
time = 0
last = index
D ← D \ (δ, ∗) {Remove all dependencies}
V ← V \ δindex {Remove transmission from set}
{Scheduling loop schedules transmissions greedy}
while !empty(V) do
  time_min ← infinity
  {Calculate minimum starting time for remaining transmissions}
  for δ ∈ V do
    schedule[δ_index] = max(schedule[δ_index], time + constraint(δ_last, δ_index))]
    {See if this transmission has the smallest starting time}
    if (∗, δ) ∉ D and schedule[δ_index] < time_min then
      time_min ← schedule[δ_index]
      index ← δ_index
    end if
  end for
  {Schedule transmission with smallest starting time first}
  time = time_min
  last = index
  V ← V \ δ_index
  D ← D \ (δ, ∗)
end while
```

Fig. 11. Reduced complexity algorithm for scheduling transmissions with dependencies.

the parents transmission should be generated. This causes the transmission to be scheduled from the highest hop nodes first to the lower hop transmissions. Care should be taken that no dependency-cycles are generated, because this renders the dependencies to be unschedulable.

## V. EVALUATION OF COMMUNICATION AND COMPUTATION COMPLEXITY

To evaluate the different centralized and distributed scheduling approaches, we will first discuss briefly their complexity in terms of number of communications required as well as computational complexity of different approaches. The complexity overview of all scheduling approaches can be seen in Figure 12.

- Centralized Scheduling: In this case we assume all transmissions as well as position information are collected in a central location. The communication complexity is $n \cdot hops_{avg}$ (The average number of hops), because all transmission information needs to be sent over a multi-hop link to the central scheduler. For scheduling the links we will use the algorithm described in [7], whose complexity is $O(n^3)$.
- Reduced Complexity Centralized Scheduling: This is the algorithm described in Section IV-A. The computational complexity of this algorithm is $O(\frac{1}{2}n^2)$. The communication complexity is the same as the other centralized scheduling approach, namely $O(n^3)$.
- Distributed Scheduling: In the distributed situation, the transmissions are sent only to the cluster-head ($O(n)$ communications). The cluster-head will calculate a schedule for its own cluster and will forward the length of its schedule over a multi-hop link to the central scheduler. This results in $O(hops_{avg}k)$ number of communications. On average, the number of transmissions per cluster is $n/k$, which results in a computational complexity of $O((n/k)^3)$ per cluster, but also for the whole network.
- Distributed Reduced Complexity Scheduling: It is similar to the distributed approach, but the scheduling per cluster uses the reduced complexity centralized scheduling algorithm. This reduces the scheduling algorithm complexity to $O(\frac{1}{2}(n/k)^2)$ per cluster. The communication complexity remains $O(hops_{avg}k)$.

The packet size of all approaches is constant and does not grow with respect to the number of nodes in the network. From the evaluation of the complexity of the different approach, we can see that the distributed approaches have a much lower computational and communication overhead compared to the centralized approaches. The scalability of the distributed approaches is therefore much better than the centralized approaches.

## VI. EVALUATION OF SCHEDULING EFFICIENCY

To evaluate the scheduling efficiency of the different approaches, we implement them in c++. We evaluate the algorithms for different sizes of deployments. The parameters can be found in Figure 13(a). The network size ranges from 500 up to 8000 nodes scattered randomly over an area. The communications are set up in such a way that all data is collected at a central sink, similarly to the deployment illustrated in Figure 7.

For the different distributed scheduling approaches a reuse distance should be selected. We evaluated the distributed algorithms with both 3 as well as 7 timeslots. 3 timeslots is the minimum number of timeslots required and the reuse distance in this case will be exactly the interference range. Using 7 timeslots increases the reuse distance beyond the interference range, this provides a guard band for when the interference range in reality can not be that accurately estimated.

The evaluation results are shown in Figure 14(a). We see that the centralized approach performs the best, which is expected. This is due to the fact that the centralized approach has all link and deployment information of the network during the scheduling, while the distributed approach splits up the scheduling in sub-problems and uses local information only. The centralized approach places a lower bound on the achievable schedule length.

The reduced complexity centralized algorithm performs only slightly worse, the difference in schedule lengths is

| Scheduling approach | Computational | Communication | Packet size |
|---|---|---|---|
| Centralized | $O(n^3)$ | $2(n \cdot hops_{avg})$ | $O(1)$ |
| Reduced Complexity Centr. | $O(\frac{1}{2}n^2)$ | $2(n \cdot hops_{avg})$ | $O(1)$ |
| Distributed | $O((n/k)^3)$ | $2(n + k \cdot hops_{avg})$ | $O(1)$ |
| Distributed Reduced Complexity | $O(\frac{1}{2}(n/k)^2)$ | $2(n + k \cdot hops_{avg})$ | $O(1)$ |

$n$ = Number of transmissions
$k$ = Number of clusters

Fig. 12. Complexity of different scheduling approaches compared.

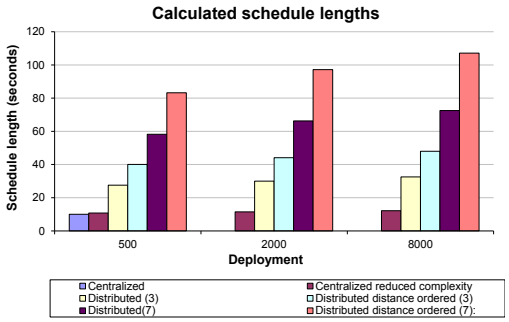
| Parameter | Value |
|---|---|
| Communication range: | 500m |
| Data rate: | 1000bps |
| Propagation speed: | 1500 m/s |
| Node placement: | random / uniform |

(a) General parameters

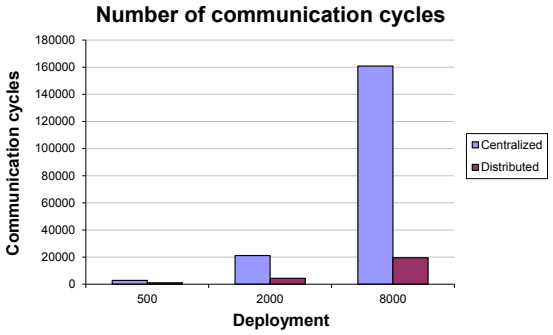
| Parameter | Small | Medium | Large |
|---|---|---|---|
| Clusters: | 4 x 3 | 7 x 7 | 14 x 14 |
| Area size: | 3.2 x 3.1km | 5.5 x 6.6km | 11 x 13km |
| Nodes: | 500 | 2000 | 8000 |

(b) Different deployment sizes

Fig. 13. Simulation parameters.




(a) Schedule length of different scheduling approaches




| | Scheduling approach | |
|---|---|---|
| Network size | Centralized | Distributed |
| 500 | 2813 | 1040 |
| 2000 | 21147 | 4418 |
| 8000 | 160878 | 19540 |

(b) Number of communication cycles required to setup schedule

Fig. 14. Results of simulation for different deployments and scheduling approaches.

only marginal. Therefore the reduced complexity centralized algorithm is a good alternative to the full complexity centralized algorithm. In Section IV-A and Section V we have already shown that the reduced complexity algorithm has large benefits in terms of computation and memory complexity. From the results of the simulation, we can conclude these benefits come at almost no cost in terms of schedule efficiency.

Among the distributed approaches, the distributed approach which minimizes schedule length and uses 3 timeslots, performs about twice as worse as the centralized approach. The approach that orders the transmissions based on distance of the transmission performs worse. The fact that the distributed approach performs worse when the network size increases is because for every timeslot the maximum schedule length from all clusters using that timeslot is used. If more clusters use the same timeslot, the maximum schedule length over all these clusters will go up.

The schedule lengths of the distributed approach are on average 270% of the centralized approach when 3 timeslots are used, and 580% when 7 timeslots are used. This shows that when the scalability, computational and communication benefits are irrelevant a centralized approach is still much preferred.

In Figure 14(b) the amount of communications cycles required to set up the network is shown. The difference between the centralized and distributed approach can be seen quite clearly. The centralized approach does not scale very well to large network sizes and requires large number of communication cycles. The distributed approach grows almost linearly with the size of the network. The number of communication cycles required is a little over 2 times the number of nodes in the network. The packet size of the messages is independant of the number of nodes in the network as has been noted before and contains only position and transmission information, or total schedule length for the cluster heads.

### A. End-to-end delay

A criteria for optimization, next to the criteria of optimizing for throughput, may be the time it takes for a packet to travel from the source node to the central node. In this section we will look at this end-to-end delay and we will look at how scheduling dependencies can be used to reduce this end-to-end delay. We have simulated the network in the same setup as before, we have used different number of nodes: 150 nodes, 500 nodes and 2000 nodes. However in this scenario we have setup the transmissions to aggregate the result of the child nodes. Using the shortest hop distance routing algorithm we determine for every node a parent. Every parent will have to send a packet of size 32 bytes plus the number of childs times 32 bytes. So every node is able to send its own data and forward the data from all its childs. The total size of data a parent will have to send depends on

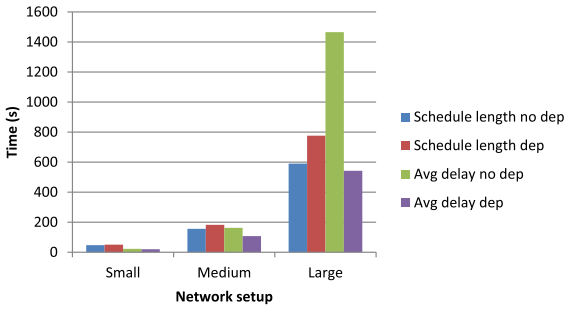



Fig. 15. Scheduling with and without dependencies

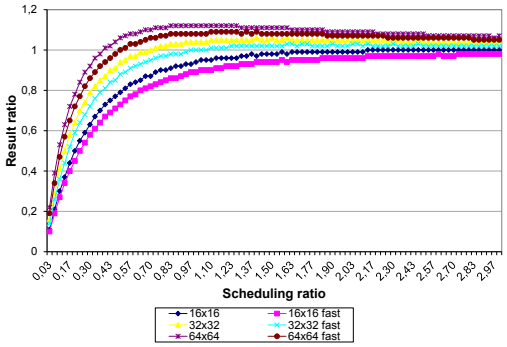



Fig. 16. Scheduling at different ratios

the number of children ($n$) as follows:

$$size_{total} = 32 + 32 * n \tag{8}$$

In this scenario we look at the length of a single run of the schedule, but we will also look at how long it takes for every packet to travel from the source node to the central node. If no dependencies are used, a packet may be in the network for several subsequent runs of the schedule. If dependencies are used, the parent will wait for all packets to arrive from the children and then starts transmitting his packet and forwards all of his children packets. The results are shown in Figure 15.

What can be seen from these results is that when scheduling with dependencies, the schedule length is increased. This results in a lower throughput. However the average end-to-end delay is decreased because the packets can all be delivered to the central node in a single run. One can see that for the small network the difference in end-to-end delay is not that substantial. This is because in these small networks the number of hops a packet has to travel is small and the length of a schedule run is still short. For medium and large networks the difference is substantial.

Looking at the results, one should consider the application and whether it makes sense to optimize for end-to-end delay. Considering that quite large networks with large number of hops need to be constructed before the difference becomes noticable. One scenario where using dependencies does make sense is when a very low duty cycle is used. The network may sense, send data and then go to sleep for a considerable time. For example the network may sense at a rate of every 10 minutes or every hour. In this scenario it would make sense to optimize for end-to-end delay, because every subsequent run of the schedule may add a delay of 10 minutes or an hour. Packets that require multiple runs of the schedule before being delivered add a delay of many minutes between every iteration of the schedule.

Another scenario for optimizing the end-to-end delay would be when distributed processing such as aggregation is used. In such a scenario a parent can not send before it has received all data from its children. In such a situation the ordering of transmissions is required and the scheduling algorithm with dependencies can be used to achieve a good throughput for the communication in such a network.

### *B. Scheduling efficiency at different propagation time / modulation time ratio*

In this section we evaluate the performance of scheduling independent of the size of the deployment or the modulation rate used at the low-level radio. To do so, we realise that the two most important factors of the performance of the schedule are (I) the propagation time between nodes and (II) the modulation time of the packets. We can simulate the scheduling approach under different sizes of deployments, different packet sizes and different modulation rates, however we can also simulate independent of these parameters by taking the modulation time / propagation time ratio. We define this ratio as follows:

$$ratio_{scheduling} = \frac{time_{modulation}}{time_{propagation}}$$

To give an example, say we want to send 125 byte packets at a modulation rate of 1000bps. The modulation time for a packet in this example is 1 second. Say the average distance between nodes in our network is 1500 meters, this is an equivalent of a propagation time of 1 second. In this scenario our scheduling ratio is 1.

In another scenario in which we want to send 125 byte packets at a modulation of 10kbps, the modulation time for a packet is 100ms. Having an average distance of 150m between nodes results in the same scheduling ratio.

To define a result value independent of the modulation time, we use the following ratio:

$$ratio_{result} = \frac{throughput_{schedule}}{throughput_{radio}}$$

We look at the resulting throughput of the schedule in relation to the throughput available of the radio. This makes the result independent of the chosen radio throughput. We calculate the schedule with different scheduling ratios, for a deployment of nodes uniformly deployed. We simulate this for different number of nodes: 16 nodes, 32 nodes and 64 nodes.

Interesting to see is that there is a certain optimum for the performance around a scheduling ratio of 1. When

going to lower scheduling ratios, the efficiency starts to decrease rapidly. In these scenarios the propagation time takes the overhand in the schedule and the schedule effectively becomes sparse. When going to the higher scheduling ratios than 1, the efficiency of the schedule decreases a little bit but seems to converge to a result ratio of 1.

What can be seen from Figure 16 is that when the propagation time takes the overhand, the efficiency of communication in terms of bandwidth starts to decrease. We have shown this for scheduled MAC protocols but the result may also be valid for unscheduled MAC protocols in the underwater environment. Therefore if the propagation delays are large it makes sense to aggregate more packets into a single transmission or switch to a lower modulation rate. This can be done without losing too much efficiency.

Another observation from Figure 16 is that for different number of nodes the performance slightly differs. We tried to define an indicator for the performance of the scheduling using the modulation time and the propagation time. However from the results it becomes clear that this does not fully describe the performance. The ratio gives a good indication on what the expected performance will be, but when more nodes and therefore more links are in the network the performance does slightly increase. We believe this is because when more links are available the greedily approach of the scheduling algorithm works better because at every step it has more options to choose from.

Using ratios we determined the efficiency of communication scheduling independent of data-throughput of the radio, data packet sizes and node distances. The performance of the schedule is dependent on the number of links. We have shown that communication scheduling achieves the highest throughput when the ratio modulation time and propagation time is 1. This shows that when a certain average distance is dictated by a deployment, the modulation time and data size should be selected accordingly.

### C. *Efficiency of broadcast scheduling*

Scheduling broadcasts is possible but is not optimal as scheduling unicasts. To evaluate the efficiency of broadcast scheduling we placed four beacons on a surface and transmit broadcast messages. On the surface there are ten nodes which should receive the broadcast message. We measure the time it takes until all ten nodes have succesfully received all four broadcast messages.

We compare the performance of scheduled communication with ALOHA [6]. We set the sending rate of the beacons to $G = \frac{1}{2}$ which is the optimal sending rate for pure ALOHA [6]. We run the simulation at different distances between beacons and at two different modulation rates. The results are shown in Figure 17.

What can be seen from the results is that scheduling broadcasts is not optimal. This is because when scheduling broadcasts the messages are scheduled very

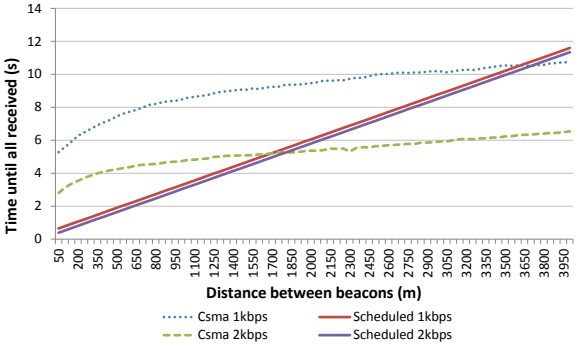



Fig. 17. Result of broadcast scheduling at different beacon distances and different modulation rates

pessimisticly, while in reality collisions occur much less often. The graph shows therefore a cut off point where ALOHA communication performs better than scheduled communication.

From the results it also becomes clear that the performance of scheduled communication is very dependent on the distance between beacons. This is because the delays for transmitting a packet are calculated based on the distance between beacons. The modulation rate, however, does not have a significant impact on the performance. The graph shows two lines for scheduled communication where the higher modulation rate only shows a slight improvement in performance (the 2kbps simulation is the lower line while the 1kbps simulation is the top line).

The ALOHA performance is not very dependent on the distance between the beacons, i.e. the distance is a factor in the performance but the modulation rate is the biggest factor. When the modulation rate doubles the performance of ALOHA almost doubles. The biggest factor in the performance of ALOHA is the modulation rate.

## VII. CONCLUSION

Scheduling algorithms for underwater communication allows mitigating the effects of the long propagation delay of the acoustic signal. Scheduling has significant benefits in terms of throughput, energy consumption, and reliability.

In this paper we discussed the extended set of simplified scheduling constraints and introduced a centralized and a distributed scheduling technique for underwater acoustic communication systems. The centralized approach achieves the highest throughput of all scheduling approaches but does this at the cost of high computational and communication overhead.

The distributed approach groups all transmissions together in clusters from which they originate. Nodes within a cluster communicate with the cluster-head only for scheduling their link. Our approach does not place any restrictions on the communication patterns. It does not restrict communication between sink and node and nodes can communicate directly with other nodes within communication range. Each cluster-head will calculate a schedule for its cluster and will forward the total schedule

length of its cluster to a central scheduler. The central scheduler will schedule the timeslots and assign a timeslot to each cluster. Compared to the centralized approach, the distributed approach has a much lower communication and computational overhead.

Comparing communication and computational complexity of the proposed algorithms shows that the distributed approach is much more scalable in larger networks. We also evaluated the schedule lengths of different scheduling approaches. The reduced complexity centralized approach calculates only marginally less efficient schedules, and is therefore a good replacement for the full complexity approach.

We have also introduced an approach to allow scheduling with dependencies between transmissions. This allows to restrict scheduling transmissions in a certain order. We have used this to schedule transmissions from higher hop nodes before transmissions from lower hop nodes. This allows aggregation of data from the outside of the network to a central data collection node and can be used to reduce the end-to-end delay of packets in a data-collection network.

We presented a novel way of estimating the performance of scheduling based on the ratio of modulation time and propagation time. We have shown that this ratio gives a good indication of the expected performance even though the number of links also has a small impact on the performance.

We have also evaluated the performance of broadcast scheduling and compared its performance with ALOHA. We show the performance of broadcast scheduling is dependent on the distance between nodes, while the performance of ALOHA is dependent on the modulation rate. We also show that ALOHA is able to outperform broadcast scheduling. This shows that, as opposed to unicast scheduling, broadcast scheduling is not always a better choice than ALOHA.

Finally we evaluate the end-to-end delay performance of scheduling with dependencies and discuss in what cases dependencies are benificial.

Our future work includes considering the effects of acoustic signal such as refraction, multipath and propagation speed variability on performance. Other effects that will be considered are node dynamics, position estimation errors and time-synchronisation errors. We also want to verify the results of the simulation in real experiments.

## ACKNOWLEDGMENT

This work is supported by the SeaSTAR project funded by the Dutch Technology Foundation (STW).

## REFERENCES

[1] A. A. Syed, W. Ye, and J. Heidemann, "T-Lohi: A New Class of MAC Protocols for Underwater Acoustic Sensor Networks," in *IEEE INFOCOM*, 2008.

[2] M. Molins, "Slotted FAMA: a MAC protocol for underwater acoustic networks," in *In IEEE OCEANS06, Singapore*, 2006, pp. 16–19.

[3] C.-C. Hsu, K.-F. Lai, C.-F. Chou, and K. C.-J. Lin, "ST-MAC: Spatial-temporal mac scheduling for underwater sensor networks." in *INFOCOM*. IEEE, 2009, pp. 1827–1835.

[4] P. M. Kurtis Kredo II, "Distributed scheduling and routing in underwater wireless networks," *Globecom 2010*, 2010.

[5] L. F. Vieira, J. Kong, U. Lee, and M. Gerla, "Analysis of aloha protocols for underwater acoustic sensor networks," in *Work in Progess poster at the First ACM International Workshop on UnderWater Networks (WUWNet)*. Los Angeles, California, USA: ACM, September 2006.

[6] A. Tanenbaum, *Computer Networks*, 4th ed. Prentice Hall Professional Technical Reference, 2002.

[7] W. van Kleunen, N. Meratnia, and P. J. Havinga, "A set of simplified scheduling constraints for underwater acoustic mac scheduling," *AINA*, 2011.

[8] W. van Kleunen, N. Meratnia, and P. J. Havinga, "Mac scheduling in large-scale underwater acoustic networks," in *8th International Joint Conference on e-Business and Telecommunication, ICETE 2011, Sevilla, Spain*. USA: IEEE Computer Society, July 2011, pp. 27–34.

[9] W. van Kleunen, N. Meratnia, and P. J. Havinga, "Scheduled mac in beacon overlay networks for underwater localization and time-synchronization," in *The Sixth ACM International Workshop on Underwater Networks (WUWNet) , Seattle, USA*. New York: ACM, December 2011, p. 6.

[10] A. A. Syed and J. Heidemann, "Time synchronization for high latency acoustic networks," in *In Proc. IEEE InfoCom*, 2006.

[11] V. Chandrasekhar, W. K. Seah, Y. S. Choo, and H. V. Ee, "Localization in underwater sensor networks: survey and challenges," in *WUWNet '06: Proceedings of the 1st ACM international workshop on Underwater networks*. New York, NY, USA: ACM, 2006, pp. 33–40.

[12] D. E. Lucani, M. Stojanovic, and M. Médard, "On the relationship between transmission power and capacity of an underwater acoustic communication channel," *CoRR*, vol. abs/0801.0426, 2008.

# Power Management Mechanism Exploiting Network and Video Information over Wireless Links

Christos Bouras[1,2], Savvas Charalambides[1,2], Kostas Stamos[1,2,3], Stamatis Stroumpis[1,2], Giannis Zaoudis[1,2]
[1]Computer Technology Institute and Press "Diophantus", Greece
[2]Computer Engineering and Informatics Dept., University of Patras, Greece
[3]Technological Educational Institute of Patra, Greece
E-mail: {bouras, charalampides, stamos, stroumpis, zaoudis}@cti.gr

***Abstract*—This article examines the ways in which cross-layer information from higher network layers may be utilized for more efficient power management in wireless networks and energy constrained mobile devices. In particular, we present and evaluate mechanisms that fine-tune transmission power according to information received from the transport (feedback reports from TFRC) and application (type of the video frame encoded) layers. Further improvements may be applied if the video encoding is done using capabilities of the SVC standard. We also describe power management adaptation techniques for wireless video transmission using the TFRC protocol that take into account feedback about the received video quality and try to adapt transmitting power accordingly. The purpose of the mechanisms is to utilize TFRC feedback and thus achieve a beneficial balance between the power consumption and the received video quality. The mechanisms proposed, offer significant improvements when used in terms of both power consumption and received video quality. All proposals are compared and evaluated using simulation. [1]**

***Index Terms*— cross-layer, SVC, TFRC, power management, wireless, video transmission**

## I. INTRODUCTION

Dealing with networking architecture by dividing functionality in layers is a tested and successful concept, especially for wired networks. It reduces complexity and makes issues more manageable and architectures more flexible and upgradeable. However, it may lead to suboptimal designs, since operations of one layer are not always aware of information available to different layers. A careful cross-layer approach, where selected communication and interaction between layers is allowed, can have performance advantages without negating the successful layer separation that has guided network design so far. A theoretical discussion of the cross-layer problem framework can be found at [1].

Wireless transmission differs in important ways from wired communication. While increased power generally correlates with a stronger signal and therefore improved transmission characteristics, in many wireless scenarios reduced power consumption is desired. This tradeoff has been explored by various researchers studying TCP (Transmission Control Protocol) modifications ([2], [3], [4]) trying to combine reduced power consumption with increased data throughput. Wireless standards such as IEEE 802.11 specify power saving mechanisms [5], although studies have shown that PSM (Power Saving Mechanisms) and other similar mechanisms carry a significant performance penalty in terms of throughput ([6], [7] [8], [9]).

In this context, an important issue for the efficiency of wireless networks is to accurately determine the cause of packet losses. Packet losses in wired networks occur mainly due to congestion in the path between the sender and the receiver, while in wireless networks packet losses occur mainly due to corrupted packets as a result of the low SNR (Signal to Noise Ratio), the multi-path signal fading and the interference from neighboring transmissions. This is information that may potentially be utilized for adjustment of transmission power. The sender of a traffic flow can be informed for packet losses through a transport protocol that provides such feedback, such as for example TFRC (TCP Friendly Rate Control) [10]. TFRC is more suitable for applications such as telephony or streaming media where a relatively smooth sending rate is important.

The same consideration applies for application-layer traffic of video. Typical video encoding standards define various types of frames with varying importance in terms of information and compressibility. I-frames are independent of other frames, P-frames are dependent on previous frames, and B-frames are dependent both on previous and future frames. Therefore, a video stream is expected to suffer more quality degradation when an I-frame is lost or delayed instead of a B-frame. The latest standard H.264/MPEG-4 defines slices instead of frames,

[1] This article is an extended version of the paper titled "Utilizing Video Encoding for Power Management over Wireless Networks", which has been presented in the 8th International Joint Conference on e-Business and Telecommunications (ICETE 2011).


doi: 10.4304/jnw.8.1.15-25

which are more fine-grained elements that make part of a video picture. The MPEG-4 protocol with the enhancements of the FGS (Fine Granularity Scalability), AVC (Advanced Video Coding) and SVC (Scalable Video Coding) provides adaptive video coding by taking into account the available bandwidth. SVC [25] enables the transmission and decoding of partial bit streams to provide video services with lower temporal or spatial resolutions or reduced fidelity while retaining a reconstruction quality that is high relative to the rate of the partial bit streams. Hence, SVC provides functionalities such as graceful degradation in lossy transmission environments as well as the possibility for bit rate, format, and power adaptation.

In this paper we investigate the above possibilities and propose cross layer mechanisms for wireless scenarios and in particular WiFi transmission scenarios. The main objective of the mechanisms is to limit WiFi power consumption while maintaining satisfactory user experience and low computational complexity. The rest of this paper is organized as follows: Section 2 gives an overview of related work in the area of cross layer optimization, and section 3 provides an introduction to the SVC standard. Section 4 describes the main proposals of the paper related to utilization of information from transport and application layers. Section 5 presents the test bed setup that was used for the experiments, which are presented along with their results in in section 6, while section 7 concludes the paper and discusses possible future work. Source code for our implementation and installation instructions can be found at [14].

## II. Related Work

Many cross-layer design proposals can be found in the literature. It is worthwhile to present how the layers are coupled, in other words, what kind of architecture change has taken place in a particular cross-layer design. The layered architecture can be bypassed in several ways, including the creation of new interfaces, the merging of adjacent layers, the design coupling without new interfaces and the vertical calibration across layers [1].

The cross-layer design approach in this paper is categorized in “Creation of new interfaces” category. The cross-layer approach is useful for wireless networks, because of the unique problems created by wireless links, the possibility of opportunistic communication on wireless links, and the new modalities of communication offered by the wireless medium.

Several researchers have focused on various issues of cross layer optimization for wireless ad hoc networks, when there is no infrastructure assumed. The author in [17] proposes a jointly optimal design of the three layers (physical, MAC, routing) for wireless ad-hoc networks and studies several existing rate-maximization performance metrics for wireless ad-hoc networks in order to select appropriate performance metrics for the optimization. In [18] the authors propose an application adaptive scheme based on priority based ARQ (Automatic Repeat Request) together with a scheduling algorithm and FEC (Forward Error Correction) coding combined with RLP (Radio Link Protocol) layer granularity. In [1] the need of a cross-layer optimization is examined and an adaptation framework is proposed amongst the application (APP), the Medium Access Control (MAC) and the Physical (PHY) layers. In the same publication a number of different methodologies for cross-layer adaptation are proposed, named “top-down” approach, “bottom-up”, “application centric” and “MAC centric”.

The work in [19] summarizes the recent developments in optimization based approaches for resource allocation problems in wireless networks using a cross-layer approach. Paper [20] deal’s with 802.16 WiMax (Worldwide Interoperability for Microwave Access) networks. This paper presents an adaptive cross-layer scheduling algorithm for the IEEE 802.16 BWA (Broadband Wireless Access) system. The algorithm uses adaptive modulation and coding (AMC) scheme at the physical layer according to the SNR on wireless fading channels. In [21], the gap between existing theoretical cross-layer optimization designs and practical approaches is examined.

Power management in wireless networks is surveyed in [13] and techniques classified according to the layer where they are applied (application, transport, network, data link, MAC or physical). The authors in [15] propose a power management scheme for intra-frame refreshed image sequences of the wireless video service in code-division multiple-access (CDMA) systems, while [16] introduces coordinated power management policies for video sensor networks. In [23] , transmission power is one of the parameters that were jointly optimized in order to minimize power consumption. A thorough survey of power-awareness in mobile multimedia transmissions can be found in [24]. To the best of our knowledge the cross-layer design presented in this paper, is the first one taking into consideration parameters such as receiver’s perceived video quality, while using TFRC in wireless video transmission.•

## III. Scalable Video Coding

Scalable video coding (SVC) is a highly attractive solution to the problems posed by the characteristics of modern video transmission systems. It was standardized as an extension of H.264/AVC. Deriving from H.264/AVC, it maintains the concepts of using a Video Coding Layer (VCL) and a Network Abstraction Layer (NAL). There are three main kinds of scalability that SVC can support:

- Temporal scalability: A bit-stream provides temporal scalability when the set of access units (a set of NAL units that always contains exactly one primary coded picture) can be partitioned into a temporal base layer and one or more temporal enhancement layer(s). A strictly requirement for a bit-stream to be called temporal scalable is that, when we remove all access units of all temporal enhancement layers with a temporal layer identifier higher than k ($1 < k <$ max-layer), then the remaining layers still form a valid bit-stream for a SVC decoder (when k=1, then we have a baselayer bit-stream

which must be compatible with conventional H.264/AVC decoders). Due to its non-reference property, B slices are often used to form temporal enhancement layers.

Spatial scalability: A bit-stream contains of multiple layers, in which each layer corresponds to a supported spatial resolution and can be referred to by a spatial layer with a dependency identifier. In each spatial layer, motion-compensated prediction and intra-prediction are employed as in single-layer video coding.

- Quality (SNR) scalability: This scalability can be considered as a special case of spatial scalability with identical picture sizes of base and enhancement layers. Quality scalability comprises of coarse-grain quality scalable (CGS) coding, medium-grain quality scalable (MGS) coding and fine-grain quality scalable (FGS) coding.
- Combined scalability: In some cases, quality, spatial, and temporal scalability can be combined.

## IV. Power Management Mechanisms

The target of the mechanisms presented in this section is to minimize or eliminate packet losses, with an emphasis on packets containing crucial information, since even a small packet loss rate can result to important reduction of multimedia quality in the end user and result to a bad end user experience.

This section is divided in two subsections dealing with the utilization of transport layer feedback (we call the relevant mechanism "binary") and two subsections dealing with the utilization of application layer information.

### *A. The Binary Mechanism*

Combined with TFRC's limited variation in transmission rates, we aim for improved media parameters such as PSNR (Peak Signal-to-Noise Ratio) and MOS (Mean Opinion Score), which better represent the end user experience. At the same time, we have to make sure that power consumption will be bounded and will only increase when this results to noticeably improved video quality. A new interface has been provided to TFRC in order to set the power transmission accordingly.

The Binary mechanism, originating from our previous work [27], uses the TFRC receiver's reports to the sender in order to calculate the packet loss rate percentage. The algorithm considers only a constant number of previous packet losses, so that it is more adaptive to the most recent conditions of the network. This cross-layer mechanism uses information provided by the TFRC protocol which is a transport layer protocol and needs to act upon the physical layer to adjust the transmission power. The parameters involved by each layer include the transmission power at the physical layer, and the packet loss information at the transport layer.

The essence of the mechanism is to provide a better and quicker convergence to efficient power levels in scenarios where the mobile nodes follow more random movement patterns. We have to note that the objective of the binary mechanism is therefore to accommodate rapidly changing movement patterns, but not necessarily fast-moving nodes. The "binary" name comes from the dichotomic ("divide and conquer") nature of the mechanism, since it tries to divide the possible power level ranges through their middle, as will be shown below.

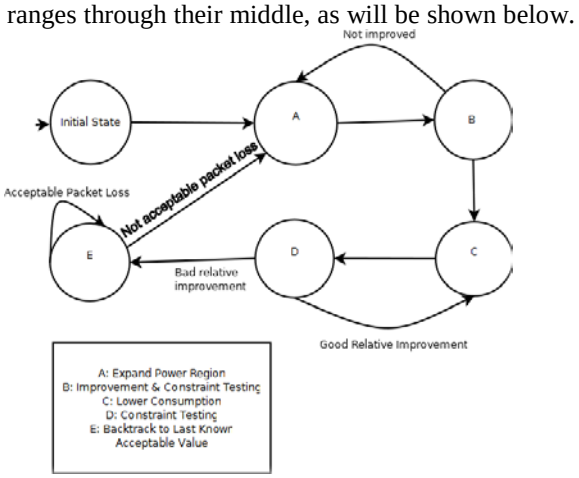



Figure 1. Finite state machine for the proposed mechanism for the sender.

The finite automaton presented in Figure 1. is the mechanism used by the sender of the video via TFRC. Every time the sender receives a TFRC report from the receiver changes its state according to the state it is in and the new data. The mechanism after receiving the first report, if packet loss is not satisfactory, defines a region in which it will try to approximate the optimum power. The optimum power is the one that produces a desired value of packet loss. After defining the region, the sender will increase its power to the maximum possible in that region and send the next TFRC packet with that power (state A). When the sender receives the next report, it tests whether there has been as significant improvement. If there has been an improvement and packet loss is below a predetermined threshold goes to state C or else repeats the actions of state A. In state C, the mechanism sets the power to the middle of the defined region and the sender goes to state D. In state D the algorithm tests whether the packet loss constraints are still satisfied and if this is the case it repeats state C. If this is not the case the algorithm goes to state E where it goes back to the previous known acceptable power value. The mechanism stays at state E while the packet loss value is acceptable, and if not it goes back to state A.

### *B. Power Management Mechanism Extended for Multiple Receivers*

In this section we consider the implications of having multiple wireless receivers and present a mechanism that extends the power management approach for efficient operation in such a scenario. In this case, the transmitting station has to calculate the most efficient transmission rate, so that a maximum number of receivers experience a satisfactory quality.

We assume that the transmitting node has a variety of nodes within its transmission range, which all wish to receive the same broadcast transmission. The problem in this case is for the transmitting node to decide on an

optimal strategy for all their varying reception capabilities.

Several approaches can be examined for generalizing the original binary mechanism for the multiple receivers' problem. Every case follows the binary mechanism described above, where step B changes accordingly.

**Follow the worst-case receiver**

Calculating an average does not guarantee that nodes with high mobility and bad channel characteristics will receive fair quality video. On the other hand taking into account extreme values could lead to high energy consumption. This mechanism variation is used in order to be efficient for every wireless node, which is included in the hop. Such an approach is suitable for a set of receivers that do not have wide differences in reception quality and capabilities, or do not quickly distance from each other or approach the transmitting node. In any case, this approach is expected to maintain a minimum quality level for every one participating node. However, the existence of outlier nodes that for some reason are not able to receive the stream properly may have a large influence on the performance of the whole system. Such an approach may be more suitable when minimum quality thresholds should be guaranteed.

B: Improvement and constraint testing according to the TFRC reports with the most packet losses. If qualified, go to state C, else go to state A

**Calculate an average**

In this scenario the mechanism variation calculates the transmission power based on every the TFRC report from all the wireless and mobile nodes, thus making our mechanism less power-consuming, although some nodes may experience transmission problems due to wireless transmission characteristics.

B: Improvement and constraint testing, by calculating the average amount of packet losses from the last five TFRC reports. If qualified, go to state C, else go to state A

**Follow the median**

Sometimes the median can be a more robust estimator than the mean in the presence of outliers, so we investigate its applicability as a criterion for feeding the power management mechanism.

B: Improvement and constraint testing, taking into account the median value from the TFRC reports of all receivers. If qualified, go to state C, else go to state A

### *C. Exploiting Video Frame Information*

Since I-frames contain the most important information compared to the rest of the frames, and their loss may affect multiple frames before and after in the frame sequence, it is reasonable to make sure that they reach their destination. If the receiving mobile node has moved further away from the transmitting node, a transmission power increase may mitigate weak signal reception problems. However, packet losses may also be due to other factors, such as channel congestion, and then power increases offer no benefit. This is where the binary mechanism is needed: its operation is to quickly identify the optimal level of power for a given network condition depending on available information about packet losses. However, since the identification of an efficient power level unavoidably has to examine several iterations of packet loss reports, it is complemented by direct changes depending on the frame type as discussed below.

We therefore introduce a modification to the adaptive algorithm presented above that tries to heuristically increase power levels only when it is expected to produce some tangible beneficial effect.

```
onBackground(BinaryMechanism())
while (true) {
   frameType=checkMPEG4FrameType()
   currPower=getCurrentPower()
   if (frameType == I)
         setPower(P_I*currPower)
   else if (frameType == P)
         setPower(P_P*currPower)
   else
         setPower(P_B*currPower)
}
```

The $P_I$, $P_P$, $P_B$ values are fixed for a transmitting node and quantify the amount of importance that each type of frame has relative to the rest. It is therefore imperative that $P_I>=P_P>=P_B$. In the test-bed and experiment sections we present the selected values for the type of encoding that was simulated and tested.

The proposed approach for selecting the values of the values for the $P_I$, $P_P$, $P_B$ parameters is to make them dependent on the statistical distribution of the I, P and B frames of the video respectively, so that average consumption does not exceed the theoretical consumption if all frames were treated identically. In other words, the $P_I$, $P_P$, $P_B$ values are also constrained by the fact that we want the sum $N_I*P_I+N_P*P_P+N_B*P_B$, where $N_I$, $N_P$, $N_B$ are the percentage of I, P and B frames respectively, to equal one.

Since power adaptation in this case is dependant upon information available at the sender (frame type), no special considerations for multiple receivers are needed in this case.

The combinations of the above mechanism with the binary mechanism lead to a new cross-layer design between Application-Transport-Physical layers.

### *D. SVC Mechanism*

In case where the transmitted video is encoded using H.264 SVC, we propose to exploit Network Abstraction Layer (NAL frames), which are segmented into a number of smaller UDP packets before feeding them to a real or simulated network. The video server component is responsible for the above procedure. In the case of a simulated transmission, this component also logs video frame number, frame type, frame size, number of segmented UDP packet, and timestamps down to a video trace file, which can then be used to simulate video transmission.

The objective of the SVC standardization has been to enable the encoding of a high-quality video bit stream that contains one or more subset bit streams that can themselves be decoded with a complexity and reconstruction quality similar to that achieved, using the

existing H.264/AVC design with the same quantity of data as in the subset bit stream.

The most important part of the NALU header for our purposes is the PRID field, which designates the priority of the specific frame, as considered by the video encoding algorithm. A lower value of PRID indicates a higher priority [26]. The proposed cross-layer design creates an interface from the application layer to the physical layer, by taking into consideration the priority information from the application layer of the transmission and passing this info to the physical layer which then adjusts its transmission power in order to achieve minimum packet loss for important SVC frames that heavily influence the perceived end-user experience.

The main idea is to exploit the video bit stream at the physical layer according to the priority of the packet that will be transmitted as specified by the SVC architecture. This information may then be used to adjust the transmission power of the sender node, making sure that frames of higher importance are transmitted with higher average power, while balancing overall power consumption with low importance frames. According to the SVC standard packets with higher priority are considered quit important for the decoding process, so our approach focuses on these packets that will lead to better end-user experience. The mechanism is actually improving the overall quality of a video especially in cases where the distance between the nodes is above a certain threshold and is increasing.

We consider beneficial a power transmission increase only in packets that carry payload information for NAL units with higher priority. Since NAL units with higher priority are important for the decoding procedure, additional transmission power will typically result in a decrease in packet loss ratio of this kind of packets which will lead to improved end user experience.

The proposed mechanism's goal is twofold. On the one hand PSNR values will increase and on the other hand transmission power will be used efficiently.

```
while (true) {
   nalu = processNALU();
   prid = getPRID(nalu);
   currPower=getCurrentPower();
   if (prid < HIGH)
         setPower(P_H)
   else if (prid < MEDIUM)
         setPower(P_M)
   else
         setPower(P_L)
}
```

Since packets with high PRID contain the most important information compared to the rest of the packets, and their loss may affect multiple frames before and after in the frame sequence, it is reasonable to make sure that they reach their destination. If the receiving mobile node has moved further away from the transmitting node, a transmission power increase may mitigate weak signal reception problems.

We expect this approach to be beneficial in cases where the distance between the nodes is large (and signal strength is correspondingly small), and especially when the receiving nodes tend to further distance themselves from the transmitting node. In such cases, signal weakness is harmful for the overall quality of the perceived video. On the other hand, we want our approach to use transmission power efficiently, even when signal strength is adequate, so that no excessive power consumption takes place.

The $P_H$, $P_M$, $P_L$ values are fixed for a transmitting node and quantify the amount of importance that each type of frame has relative to the rest. It is therefore imperative that $P_H>=P_M>=P_L$. The interaction of these parameters is explained in the pseudo-code above. Their absolute values are related to the absolute power levels available at a specific environment, with $P_M$ typical being chosen close to the average power used in a default setting, and $P_H$ and $P_L$ symmetrically above and below the $P_M$ power level.

## V. TESTBED SETUP

For our experiments we have used the Network Simulator 2 (ns-2.34) as a basic tool for simulating multimedia data transmission over wireless networks. In order to simulate MPEG-4 video transmission using ns-2, another software package is needed, namely Evalvid-RA [11]. Evalvid-RA supports rate-adaptive multimedia transfer based on trace file generation of an MPEG video file. The multimedia transfer is simulated by using the generated trace file and not the actual binary multimedia content. The simulator keeps its own trace files holding information on timing and throughput of packets at each node during simulation. Combining this information and the original video file Evalvid-RA can rebuild the video file as it would have been received on a real network. Additionally, by using the Evalvid-RA toolset the total noise introduced can be measured (in dB PSNR) as well as MOS can be calculated. An example implementation is illustrated in [12]. For experiments using SVC video transmission, we used the extension EvalvidSVC ([2], [3]). Evalvid SVC supports scalable video coding extension of the H.264 mechanism based on trace file generation of an MPEG video file, similarly to Evalvid-RA.


Figure 2. Topology in experiments

In our experiments we used the network topology illustrated in Figure 2. . The akiyo sample video found in media.xiph.org was used for video streaming for the purposes of our experiments.

The simulation environment consists of three parts and is depicted in the Figure 2. . During the pre-processing phase, a raw video file, which is usually stored in YUV format, is encoded with the desired video encoder into 30 different encoded MPEG-4 video clips with quantizer scale values in the range 2–31. Quantizer scale 2 provides an encoded video with the highest quality. We use the

ffmpeg free video encoder for the creation of the video clips. For our simulations, all video clips have temporal resolution of 25 frames per second and GoP (Group of Pictures) pattern IBPBPBPBPBPB, with a size of 12 frames. The frame size of all clips is 352x288 pixels, which is known as the Common Intermediate Format (CIF). After all the video files are encoded they are then traced to produce 30 frame-size traces files. At the end of the pre-processing phase, we thus have 30 m4v files with their associated frame size files.

Briefly, the video file was preprocessed and many video files were produced of different quality and resolution using the ffmpeg tool and shell scripts included in the Evalvid-RA toolset. Then, trace files were generated for all these files and by using these trace files the simulation took place. Ns-2 scripts were created to simulate video transmission over a wireless network over TFRC. After simulating the transfer of the video in several different resolutions, ns-2 trace files were obtained which then were used to reconstruct the videos as it would have been sent over a real network.

The third part of the simulation environment consists of the reconstruction of the transmitted video and the measurement of the performance evaluation metrics. The reconstruction of the received video traces is implemented off-line by comparing the transmitted and the received traces with those of the original video sequence of all the transmitted simulcast streams. In this phase, several measurements and calculations can be done involving network and video metrics such as PSNR, MOS, jitter, throughput and delay. With the above described procedure we are able to make extensive comparisons between algorithms and reach conclusions about the efficiency of each one.

For SVC experiments, we used the DownConvertStatic resampler. This tool is used for spatial/temporal resampling of video sequences. In our procedure we used it to spatially resample our video to a resolution of 176x144 at 30 Hz, from 352x288 in order to have the same video sequence but with two different spatial characteristics. The next step was to encode the two separate video sequences into one spatial scalable bit-stream. To accomplish this we used the H264AVCEncoderLibTestStatic AVC/SVC encoder. The encoder is used for generating AVC or SVC bit-streams depending on the encoding mode you select in the main configuration file of the encoder. The parameter that defines the encoding is AVC mode. After defining the parameters of the encoder's configuration files and encoding our video sequences we get a spatial scalable bit-stream. Following the encoding we used the MP4Box tool that came with the EvalSVC tool to create an ISO MP4 file which will contain the video samples and a hint track to describe how to packetize the frames for transport. Furthermore we used we used the mp4trace tool from EvalSVC to create the mp4 file.

The output of the mp4trace tool was used as an application in ns-2 to produce traffic in our simulated scenario and by enabling tracing we produced the needed trace file.

We used the EvalSVC toolset to generate the appropriate trace files for transmission over the network simulator ns-2. Through EvalSVC toolset we exported the PRID of the NALU header, by using of a modified version of mp4trace tool. The trace files that were used had spatial scalability, where two resolutions of the same video were used.

Several modifications of the network simulators were needed in order to build a working instance of the proposed mechanisms. Firstly, a module that implements the logic of the proposed mechanisms was added in the simulator. Then, the module that implements the TFRC protocol was changed so that it provides information about packet losses to our mechanism. The mechanism calculates the power needed to improve PSNR and then this information is passed to the modified wireless physical layer module that is able to increase or decrease power according to the mechanism.

Furthermore, the module that implements the UDP protocol was modified in order to retrieve the video frame and priority information. The mechanisms run constantly throughout the whole simulation process at the agent of the transmitting node, which is an integrated agent of the toolset in ns2, where PRID info is available with the modifications we made.

Additionally, by using the EvalvidSVC toolset the total noise introduced can be measured (in dB PSNR) as well as Mean Opinion Scores (MOS) can be calculated. Objective PSNR measurements can be approximately matched to subjective MOS according to the standardized TABLE 1.

TABLE I. ITU-R QUALITY AND IMPAIRED SCALE AND POSSIBLE PSNR TO MOS MAPPING [22]

| PSNR [dB] | MOS | Impairment |
|---|---|---|
| >37 | Excellent (5) | Imperceptible |
| 31-37 | Good (4) | Perceptible, but not annoying |
| 25-31 | Fair (3) | Slightly annoying |
| 20-25 | Poor (2) | Annoying |
| <20 | Bad (1) | Very annoying |

## VI. PERFORMANCE EVALUATION EXPERIMENTS

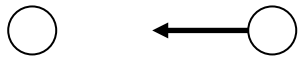
For our experiments, we transfer H.264 video over TFRC over a wireless link and in particular over a single hop in a wireless ad hoc network. Selection of $P_I$, $P_P$, $P_B$ values for this specific video encoding was 1.3, 1.1 and 0.9 respectively. In order to model various instances of network degradation, we have performed a series of experiments with various scenarios, with both stationary and mobile nodes:

- Scenario 1: Two nodes, both stationary

- Scenario 2: Two nodes, one stationary, one moving away

• Scenario 3: Two nodes, one stationary, one moving closer and then moving away

• Scenario 4: Two nodes, one stationary, one moving closer



In all scenarios, the nodes communicate wirelessly using 802.11 MAC protocol and the distributed coordination function (DCF) from the Carnegie Mellon University. Propagation model used was two-ray ground reflection model.

TABLE II. STATIONARY NODES

| Power management | Triple cross-layer | Binary | None |
|---|---|---|---|
| PSNR average | 37.8 | 37.6 | 37.1 |
| Energy Consumption | 0.051 W | 0.046 W | 0.046 W |
| MOS | Excellent (5) | Excellent (5) | Excellent (5) |

In this scenario both nodes are stationary, so power requirements do not vary. Nevertheless, power management mechanisms offer a better PSNR with slightly increase in transmission power.

TABLE III. ONE NODE MOVING AWAY

| Power management | Triple cross-layer | Binary | None |
|---|---|---|---|
| PSNR average | 35.3 | 34.8 | 30.2 |
| Energy Consumption | 0.049 W | 0.047 W | 0.047 W |
| MOS | Good(4) | Good(4) | Fair (3) |

This time, the proposed mechanism displays a noticeable performance advantage over the approach without any mechanism. We observe that it actually achieves Good Mean Opinion Score while the value for the same scenario without any power management mechanism in fair.

TABLE IV. ONE NODE MOVING CLOSER AND THEN AWAY

| Power management | Triple cross-layer | Binary | None |
|---|---|---|---|
| PSNR average | 36.2 | 36.1 | 33.3 |
| Energy Consumption | 0.050 W | 0.048 W | 0.048 W |
| MOS | Good(4) | Good(4) | Good (4) |

The same applies to this scenario, where the power management mechanisms significantly improve received video quality as shown by the PSNR values. Power increase is non-existent or very small in both cases. The reason is that both mechanisms are capable to adapt to the changing distances between the nodes and tweak the power levels accordingly.

When a node is moving closer it is natural to achieve a better PSNR value in all methods. By also using rapid adjustment of power even better results occur, whereas power consumption again stays relatively low.

TABLE V. ONE NODE MOVING CLOSER

| Power management | Triple cross-layer | Binary | None |
|---|---|---|---|
| PSNR average | 38.8 | 37.9 | 34.6 |
| Energy Consumption | 0.049 W | 0.046 W | 0.046 W |
| MOS | Excellent (5) | Excellent (5) | Good (4) |

The results from all scenarios demonstrate that in all cases the proposed mechanism significantly outperforms the default behaviour (without any power management mechanism) as it achieves higher video quality reception, with only slight increases of average power levels. Figure 3. summarizes the results of the experiments in terms of the ratio PSNR/power which gives us an estimation of how well the trade-off between power consumption and video quality is balanced.

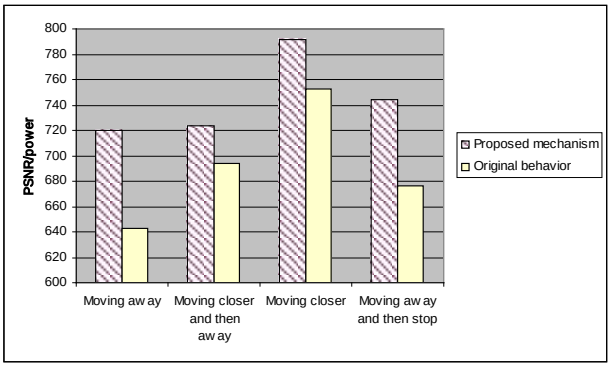



Figure 3. PSNR/power ratio

We can see that the proposed mechanism achieves a significantly improved trade-off, which means that the mobile nodes may gain in either quality or power consumption or both.

### *A. Experiments with SVC Encoding*

In this set of ns-2 experiments, we transfer H.264, in particular SVC extension, video over UDP using the same network setup described above. In order to model various instances of network degradation, we have performed a series of experiments with various scenarios, with both stationary and mobile nodes.

We then compare the achieved throughput in terms of PSNR and power consumption. Objective PSNR measurements can be approximately matched to subjective MOS according to the standardized TABLE I.

During the preprocessing phase a raw video file, which is usually stored in YUV format, is encoded with the desired video encoder. For our simulations, all video clips have a spatial scalability where the frame size of clips is 352x288 and then is down sampled and merged with 177x144 frame size using the EvalSVC toolset.

TABLE VI. EXPERIMENTS WITH STATIONARY NODES

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 32.76 | 31.81 |

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| Energy Consumption | 0.272W | 0.28W |
| MOS | Good (4) | Good(4) |

In the first scenario, both nodes are stationary, so power requirements do not vary. Nevertheless, power management mechanisms offer a better ratio of PSNR to transmission power. The proposed mechanism proves especially capable in taking advantage of the available transmission power.

TABLE VII. EXPERIMENTS WITH ONE NODE MOVING AWAY

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 27.53 | 23.49 |
| Energy Consumption | 0.272W | 0.28W |
| MOS | Fair (3) | Poor(2) |

This is a scenario where the cross-layer mechanism significantly affects perceived end-user experience. Its handling of higher priority frames leads to noticeably better PSNR values for the same average power consumption. We observe that the optimization also leads to an upgrade of the PSNR-equivalent MOS score. The improvement in the result can be understood if we consider the fact that while the moving node is distancing itself from the transmitting node, it crosses at some point the threshold where signal strength is no longer adequate for proper packet reception. Due to the increased power allocated to high importance packets, the proposed mechanism is able to keep video transmission at an acceptable level for a significantly longer time period.

TABLE VIII. EXPERIMENTS WITH ONE NODE MOVING CLOSER

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 34.67 | 32.65 |
| Energy Consumption | 0.272W | 0.28W |
| MOS | Good (4) | Good(4) |

Since a node is moving closer it is natural to achieve a better PSNR value compared to the other scenarios. Usage of the proposed mechanism again achieves better results occur, without adversely affecting power consumption.

TABLE IX. EXPERIMENTS WITH ONE NODE MOVING CLOSER AND THEN MOVING AWAY

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 30.25 | 28.76 |
| Energy Consumption | 0.272W | 0.28W |
| MOS | Good (4) | Good(4) |

In this case the node changes its movements rapidly but our mechanism seems to react better in terms of PSNR values though MOS level is the same. In cases where the receiving node is moving away our mechanism leads to better overall video quality.

TABLE X. EXPERIMENTS WITH ONE NODE MOVING CLOSER THEN MOVING AWAY AND THEN MOVING CLOSER AGAIN

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 32.23 | 29.65 |
| Energy Consumption | 0.272W | 0.28W |
| MOS | Good(4) | Fair(3) |

The proposed approach demonstrates a significant performance lead for the cross-layer approach, including an upgrade of the PSNR-equivalent MOS score compared to the default approach.

TABLE XI. EXPERIMENTS WITH ONE NODE MOVING CLOSER AND THEN STOPS MOVING

| Measurement | Nalu mechanism | Without mechanism |
|---|---|---|
| PSNR average | 33.14 | 32.02 |
| Energy Consumption | 0.272W | 0.28W |
| MOS | Good(4) | Good(4) |

In this case both mechanisms achieve comparable results, with no benefit of the mechanism but also no negative effects.

The results from all scenarios demonstrate that in almost all cases the proposed mechanism outperforms the default behavior (without any power management mechanism) as it achieves higher video quality reception, with negligent increase of average power levels. The results from all scenarios are summarized in Figure 4. , which displays the ratio of PSNR/Power for all mechanisms and scenarios. A higher value means that the mechanism achieved better video quality with lower power consumption, which is our main objective.

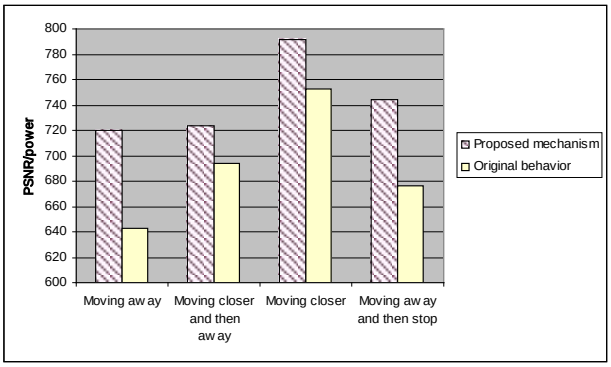



Figure 4. Test results

We can see that the proposed mechanism achieves a significantly improved trade-off, which means that the mobile nodes may gain in either quality or power consumption or both, compared to the original approach that does not utilize the cross-layer information.

### *B. Multiple Receivers*

In the last set of experiments we transfer H.264 video over TFRC over wireless links. In this set of experiments we assume that the video is received by multiple mobile nodes, so that we can verify that the proposed mechanism scales satisfactorily. We simulate various patterns of movement for the sending and receiving nodes as detailed

below. We compare the results depending on the type of policy for configuring the cross-layer mechanism as explained in the relevant part of section 4.

**Scenario 1:** Results from two nodes moving randomly and the other two approaching the transmitting node are summarized in the following table.

| Mechanism | PSNR average | Energy Consumption | MOS |
|---|---|---|---|
| None | 30.1 | 0.034 | Fair |
| Worst case | 30.8 | 0.041 | Fair |
| Median | 30.7 | 0.035 | Fair |
| Average | 32.3 | 0.034 | Good |

In this scenario, we observe that the average approach obtains clearly superior results (the only one that gets a "Good"-equivalent in the MOS scale), while it also ties for best energy consumption.

**Scenario 2:** Two nodes move randomly, one node is stationary and the other is leaving the hop.

| Mechanism | PSNR average | Energy Consumption | MOS |
|---|---|---|---|
| None | 31.0 | 0.038 | Good |
| Worst case | 31.2 | 0.040 | Good |
| Median | 30.8 | 0.035 | Fair |
| Average | 28.4 | 0.035 | Fair |

As we can see in the above table the average approach did not excel in the quality of the transmitted video, although it did achieve the best energy result among compared approaches. We conclude that the average approach is not aided by a scenario where the behavior of the nodes varies widely. On the other hand, the median approach in this case was able to achieve the best results as it weighs down extreme values that heavily influence the calculation of the average.

**Scenario 3:** Two nodes move randomly, one node is stationary and the other is approaching the base station.

| Mechanism | PSNR average | Energy Consumption | MOS |
|---|---|---|---|
| None | 28.2 | 0.031 | Fair |
| Worst case | 33.4 | 0.041 | Good |
| Median | 29.5 | 0.035 | Fair |
| Average | 29.6 | 0.033 | Fair |

The best behavior in this scenario in terms of video quality was displayed by the worst-case approach, although its energy consumption was the highest among all tested mechanisms, as we can see in the above table. This behavior was common for all scenarios, and is due to the worst-case approach's tendency to favor video quality guarantees for all nodes at the cost of increased energy consumption, sometimes just for the benefit of a single node.

**Scenario 4:** Two nodes move randomly and the other two are moving away.

| Mechanism | PSNR average | Energy Consumption | MOS |
|---|---|---|---|
| None | 27.1 | 0.031 | Fair |
| Worst case | 28.7 | 0.042 | Fair |
| Median | 30.4 | 0.036 | Fair |
| Average | 29.8 | 0.032 | Fair |

Since half of the nodes are moving away from the transmitting node in this case, this has been the most adverse scenario for almost all mechanisms. Especially the worst-case approach displayed heavily increased energy consumption, as it tried to accommodate nodes that were moving out of transmission range. The average approach was though able to obtain fair quality results with very low energy consumption. Overall results are presented in the above table.

**Scenario 5:** Three nodes move randomly and the one left is stationary.

| Mechanism | PSNR average | Energy Consumption | MOS |
|---|---|---|---|
| None | 27.3 | 0.031 | Fair |
| Worst case | 29.6 | 0.038 | Fair |
| Median | 31.2 | 0.037 | Good |
| Average | 29.7 | 0.033 | Fair |

In our final experiment more nodes than ever performed random movements. The results, which are summarized in the above table, were similar with most of the previous scenarios, in that the median and average approaches yielded best results. This time however differences were somewhat diminished, as the random movements did not allow a single approach's advantage on specific type of movements to sum up.

The results from all scenarios with multiple receivers are summarized in Figure 5. , which displays the ratio of PSNR/Power for all mechanisms and scenarios. A higher value means that the mechanism achieved better video quality with lower power consumption, which is our main objective. As we can see, the worst case approach obtained a relatively low ratio in all cases. This is an expected result, as this is the trade-off that we have to pay in order for all receivers to achieve high video quality. On the other hand, selecting the average approach yields the best results in most cases, while only scenario 2 outcomes favor the median approach.

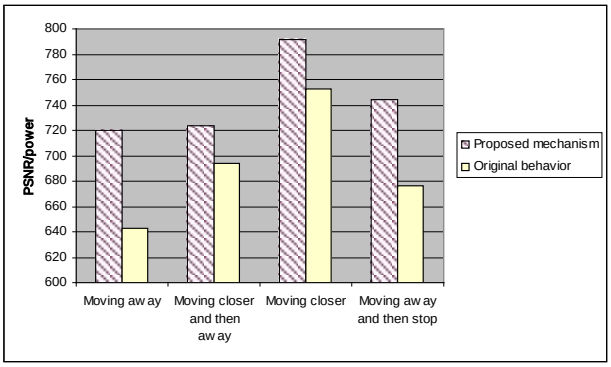



Figure 5. Summary of results

## VII. Conclusions and Future Work

In this paper we have proposed some advanced power management cross-layer mechanisms for power management in wireless TFRC and UDP transmission, which significantly improve both the objective quality of the transmitted video, and make more optimal usage of available power utilizing information from three different layers of the TCP/IP stack.

Utilizing the video encoding properties of H.264 and the SVC extension we can manage power in order to favor the most important packets. Furthermore, exploiting feedback information from the transport layer, allows the algorithm to benefit from the knowledge of the network

status. When feedback is provided by multiple receivers, we have seen that minor tweaks to the algorithm can achieve better results and can be fine-tuned depending on the specific requirements of each particular situation. Most of the presented approaches have their strong and weak points, depending on the specific type of movement performed by the nodes. The complexity cost of the mechanisms is quite small, and slightly larger power consumption in measurements seems to be the only remaining trade-off.

The proposed cross-layer mechanisms could be further improved in a wide range of ways. We plan to estimate power consumption by taking into account both power consumption for the computational complexity of encoding and the power consumption for the transmission, create an SVC rate adaptive mechanism that could be extended to support temporal, snr and combined scalability and extend the mechanism to take into account the PSNR metric along with packet loss and adjust the transmission rate, the power and the video transmission quality in order to optimize the perceived video quality. We are also working on real implementations of the algorithms in order to more accurately estimate any computational performance trade-offs.

## REFERENCES


[1] M. van der Schaar, D. Sai Shankar N, "Cross-layer wireless multimedia transmission: Challenges, principles and new paradigms", IEEE wireless Communications, Aug. 2005

[2] V. Tsaoussidis, H. Badr, "TCP-Probing: Towards an Error Control Schema with Energy and Throughput Performance Gains" 8th IEEE Conference on Network Protocols, Japan, November 2000

[3] C. Zhang and V. Tsaoussidis, "TCP Real: Improving Real-time Capabilities of TCP over Heterogeneous Networks" 11th IEEE/ACM NOSSDAV, NY, 2001

[4] Christine E. Jones, Krishna M. Sivalingam, Prathima Agrawal, Jyh Cheng Chen. A Survey of Energy Efficient Network Protocols for Wireless Networks. Wireless Networks. Volume 7, Issue 4 (Aug. 2001). pp. 343-358

[5] IEEE 802.11 PSM Standard. Power Management for Wireless Networks. Section 11.11.2

[6] Dave Molta. Wi-Fi and the need for more power. Network Computing. December 8, 2005.

[7] Huan Chen and Cheng-Wei Huang. Power management modeling and optimal policy for IEEE 802.11 WLAN systems. IEEE Vehicular Technology Conference 2004.

[8] G. Anastasi, M. Conti, E. Gregori, A. Passarella. A performance study of power-saving polices for Wi-Fi hotspots. Computer Networks: The International Journal of Computer and Telecommunications Networking. Volume 45 , Issue 3 (June 2004). pp. 295-318. 2004

[9] T. Simunic, "Power Saving Techniques for Wireless LANs", Conference on Design, Automation and Test in Europe - Volume 3. pp. 96-97. 2005

[10] M. Handley, S. Floyd, J. Pahdye and J. Widmer, "TCP Friendly Rate Control (TFRC): Protocol Specification", RFC 3448, January 2003

[11] Arne Lie, Jirka Klaue, "Evalvid-RA: Trace Driven Simulation of Rate Adaptive MPEG-4 VBR Video", Multimedia Systems, Springer Berlin / Heidelberg, Volume 14, Number 1 / June, 2008, pp. 33-50

[12] Torgeir Haukass, "Rate Adaptive Video Streaming over Wireless Networks Explained & Explored", MSc thesis, Norwegian University of Science and Technology, Department of Telematics

[13] K. Klues, "Power Management in Wireless Networks", Advanced Topics in Networking: Wireless and Mobile Networking by R. Jain, Wash. Univ. St. Louis, 2006

[14] http://ru6.cti.gr/ru6/research_ns.php

[15] Il-Min Kim and Hyung-Myung Kim "An Optimum Power Management Scheme for Wireless Video Service in CDMA Systems", IEEE Transactions on Wireless Communications, Vol. 2, No. 1, January 2003

[16] N. H. Zamora, J.-C. Kao, R. Marculescu, "Distributed Power-Management Techniques for Wireless Network Video Systems", Design, Automation & Test in Europe Conference & Exhibition, pp. 1-7, 16-20 April 2007.

[17] Bozidar Radunovic, "A Cross-Layer Design of Wireless Ad-Hoc Networks", Ph.D. Thesis, July 2005

[18] Yufeng Shan, Zakhor, A., "Cross layer techniques for adaptive video streaming over wireless networks", IEEE International Conference on Multimedia and Expo, 2002 (ICME '02), Volume: 1, pp 277- 280

[19] Xiaojun Lin, Shroff, N.B., Srikant, R. ,"A tutorial on cross-layer optimization in wireless networks", Selected Areas in Communications, IEEE Journal, Aug 2006, Volume: 24, Issue: 8,On page(s): 1452-1463

[20] Li, X., Wu, X., Li, W., Wang, X., "An adaptive cross-layer scheduling algorithm for multimedia networks", 4th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIH-MSP 2008, art. no. 4604006, pp. 52-55

[21] Warrier, A., Le, L., Rhee, I., "Cross-layer optimization made practical", 4th International Conference on Broadband Communications, Networks, Systems, BroadNets, art. no. 4550507, pp. 733-742

[22] ITU-R Recommendations BT.500-11. Methodology for the subjective assessment of the quality of television pictures (2002)

[23] J. Zhang, D Wu., S. Ci, H. Wang and A. Katsaggelos, "Power-Aware Mobile Multimedia: a Survey "Journal of Communications, Vol. 4, No. 9, October 2009.

[24] Z. Li, F. Zhai, and A. Katsaggelos, "Joint video summarization and transmission adaptation for energy-efficient wireless video streaming," EURASIP Journal on Advances in Signal Processing, 2008.

[25] H. Schwarz, D. Marpe, Member, IEEE, and Thomas Wiegand, Member, IEEE, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard"

[26] S. Wenger, Y. K. Wang, T. Schierl, and A. Eleftheriadis, "RTP payload format for SVC video," draft, Internet Engineering Task Force (IETF), February 2011.

[27] C. Bouras, A. Gkamas, V. Kapoulas, V. Papapanagiotou, K. Stamos, G. Zaoudis, "Power Management Adaptation Techniques for Video transmission over TFRC" International Journal of Network Management, Wiley InterScience, 2011


**Christos Bouras** is Professor in the University of Patras, Department of Computer Engineering and Informatics. Also he is a scientific advisor of Research Unit 6 in Computer Technology Institute and Press - Diophantus, Patras, Greece. His research interests include Analysis of Performance of Networking and Computer Systems, Computer Networks and Protocols, Mobile and Wireless Communications, Telematics and New Services, QoS and Pricing for Networks and Services, e-learning, Networked Virtual Environments and WWW Issues.

**Savvas Charalambides** was born in Pafos, Cyprus in 1987. He entered Computer Engineering and Informatics Department in 2007 and joined RU6 in 2011. He is interested in Network Design, Network Protocols and Wireless Networks.

**Kostas Stamos** was born in Patras, Greece in 1978. He has received his Diploma, Master Degree and PhD from the Computer Engineering and Informatics Department at the University of Patras. Since July 2001 he works as a R&D Computer Engineer with Research Unit 6 of CTI. His research interests include multicast video transmission, QoS, bandwidth on demand and network applications and protocol design.

**Stamatis Stroumpis** was born in Chios, Greece in 1989. He entered Computer Engineering and Informatics Department in 2007 and joined RU6 in 2010. His research interests include wireless video transmission (mpeg4-h264) and network technologies.

**Giannis Zaoudis** was born in Chios, Greece in 1987. He entered Computer Engineering and Informatics Department in 2004. His interests include Web Technologies, Network programming and Network Protocols. Currently he is a post-graduate student in CEID and he is doing research on cross-layer techniques over wireless networks.

# QoS-Aware Multipath Communications over MANETs

Muath Obaidat[1], M. Ali [2], Ihsan Shahwan[3]
[1,2, 3]City College at the Graduate Center of the City University of New York, Department of Electrical Engineering NY, USA
{Muobaidat, ali}@ccny.cuny.edu, ishahwan@hunter.cuny.edu

M.S. Obaidat[4]
[4] Monmouth University, Department of Computer Science and Software Engineering, NJ, USA, Fellow of the IEEE
obaidat@monmouth.edu

Suhaib Obeidat[5]
[5]Bennett College, Department of Mathematics and Computer Science, Greensboro, NC, USA
obeidat@bennet.edu

***Abstract*—To enhance the Quality of service (QoS) communications over mobile ad hoc networks (MANETs), this paper proposes QoS-Aware Multipath Routing Protocol (QMRP). Delay is the most crucial factor for multimedia applications which can be minimized by providing more than one path between source-destination pair as well as choosing the path based on the quality in terms of reliability and stability of the link. To the best of our knowledge no one before included projected load; load introduced by the node requesting a path to a destination into the delay computation for a path between source-destination pair as well as maintaining loop freedom through the neighbor hop list of the source. The originality of the proposed protocol comes from the fact that it introduces this new parameter into route quality computation which makes QMRP unlike its precursors providing more accurate measure of the realistic delay as well as maintaining loop freedom of multiple node disjoint paths using neighbor hop list.**
**Cross layer communications between physical (PHY), MAC and routing layers interact to achieve QoS against the network and channel dynamics by minimizing delay and choosing more reliable and stable paths without requiring any additional resources. Performance evaluation of the proposed protocol against a single path AODV routing protocol using OPNET has been conducted. Results show that QMRP outperforms AODV in terms of E2E delay, packet delivery fraction (PDF) and route discovery frequency. However, routing overhead for QMRP is more than that of AODV due to the discovery of more one path in each route discovery process.**

***Index Terms*— Multipath routing; Quality of Service (QoS); Cross layer; MANET**

## I. INTRODUCTION

MANET is an autonomous infrastructure-less multihop network that can be built on demand without the need for any backbone. In MANETs each mobile node operates not only as a host but also as a router by forwarding the traffic of other mobile nodes in the network. The ease of deployment of MANETs make it attractive for many applications such as rescue and recovery in disaster areas, education and research expeditions in remote places and emergency mobile medical units [1]. Nonetheless, MANETs have distinctive challenges including security, power management, efficient dynamic routing and QoS guarantees.

This paper investigates one of the above mentioned challenges; it focuses on routing to support QoS. To design an effective routing protocol for MANETs it's essential to understand the fundamental characteristics of these networks, MANETs are characterized by (i) dynamic nature; nodes are mobile and changing their location randomly and so topology is unpredictable which means the network status changes in a very short time (ii) radio properties; topology changes can occur even in low, or complete absence of mobility due to variation in the wireless medium as a result of attenuation, interference, multipath effects, shadowing and fading. The previous mentioned constraints of MANETs make it very difficult to achieve hard QoS (e.g., guaranteed delay and constant bit rate (CBR)) without wasting much of the network resources. Therefore, the aim is to develop a QoS-aware routing protocol that provides soft QoS; for more details on soft QoS see [2].

Two factors are required in order to provide quality assurance for delay-sensitive; real time applications in MANETs. First requirement, route selection criterion needs to be QoS-aware. Second, instantaneous response to the dynamics of the network is required so that switching routes are seamless to the user experience over the lifetime of a session [3]. To fulfill the first requirement is to be able to find a path with sufficient resources and including links in the paths that are stable to meet QoS constraints. To address the second issue, multipath routing has gained attention in the research community in the past several years [4]. Multipath

Manuscript received December 31 2011; revised June 1, 2012; accepted July 1, 2012.
Corresponding author. Email: Muobaidat@ccny.cuny.edu.


doi: 10.4304/jnw.8.1.26-36

protocols establish multiple paths between source-destination pairs; this approach has many advantages such as fault tolerance, load balancing and QoS assurance [5].

The rest of the paper is organized as follows. Section II describes related work. Section III presents background information. Section IV defines the problem and explains QoS-aware multipath routing protocol. Simulation environment and parameters are detailed in V. Section VI performance evaluation results of the proposed protocol are presented. Section VII draws conclusions.

## II. Related Work

The fundamental proposed routing protocols for MANETs are single-path routing protocols use single path to set up communications between source-destination pairs. Multipath routing protocols based on single-path *Ad Hoc on Demand Distance Vector* (AODV) proposed by Perkins and Royer [6] have been proposed in the literature. As a modified multipath version of AODV M. Marina et al. proposed *Ad hoc On demand Multipath Distance Vector* (AOMDV) [7]. AOMDV establishes link-disjoint and loop-free paths based on the minimum hop count similar to AODV criteria. Link-disjointness is achieved by a special flooding mechanism, while loop-freedom is ensured by using the notion of advertised hop count value at node N for destination D; this value represents the maximum hop count at N available for D. As a result, alternative paths at node N for D are accepted only if they have a lower hop count than the advertised hop count.

Z. Ye et al. proposed *AODV-Multipath* (AODVM) [8] which finds multiple node disjoint paths with no limit on the number of paths. Duplicate Route REQuests (RREQ) for the same source-destination pair are not discarded instead recorded in the RREQ table, the destination consequently replies to all RREQs. When intermediate node overhears broadcasting of a Route REPly (RREP) message from neighboring node, it deletes the corresponding entry of the transmitting node from its RREQ table. When an intermediate node receives a RREP which it cannot forward any further it generates route discovery error message to the node from which it received the RREP; this node will try to look up an alternate path from its table to the source.

The *Shortest Multipath Routing Using Labeled Distanced* (SMLDR) uses the shortest path regardless of the link quality [9]. SMLDR introduces a metric called limiting distance; that is the minimum distance to the destination known at each node. Lee and Gerla proposed *AODV Backup Route* (AODV-BR) [10], this is an extension of AODV with a back up route in case of the primary route failure without considering the link quality also it has been shown that the protocol does not perform well in heavy load conditions. Perkins and Royer proposed *QoS AODV* (QS-AODV), which considers delay joint with hop count as a criterion for choosing the route [11]. Nonetheless, the protocol does not consider the dynamics of MANET; such as topology changes due to mobility and/or link /node failure that will lead to changes in the estimated delay.

A. Valera et al. proposed *Caching and Multipath* (CHAMP) routing protocol [12]. CHAMP uses the joint packet caching and shortest multipath routing to minimize packet loss ratio due to route failure. *Split Multi-path Routing* (SMR) protocol proposed by Lee, S. et al. [13]. SMR is an extension of *Dynamic Source Routing* (DSR) [14]. This protocol attempts to establish maximally disjoint paths. The source broadcasts a RREQ message; however, unlike DSR the intermediate nodes do not send RREP if they have a path to the destination, from the received RREQs, the destination identifies multiple disjoint paths and sends a RREP packet back to the source for each individual route. SMR performs poorly in highly dense networks due to immense routing over head due to source routing nature of the protocol.

Resent years have shown increased interest in routing protocols that utilize Cross Layer (CL) in MANETs [15][16][17]. H. Sun et al. proposed an adaptive QoS routing protocol by cross layer cooperation [18], based on the current network conditions QoS requirements assured by adaptively using multipath routing and Forward Error Correction (FEC). In [19] M. Li et al presented cross layer multipath routing protocol (EMRP), which exchange information between PHY, MAC and routing layer to utilize the network resources.

Most of previous discussed protocols use minimum hop count as a criterion in finding and establishing paths between source-destination pairs. However, R. Draves et al. [19] have shown that minimum hop count routs without considering link quality could degrade the network performance since they might include wireless links that are bad or congested along the path causing the overall throughput of the network to degrade and cause even more delays than longer paths that consist of good links.

This work represents a follow-up to a previous work of ours [20] where the performance of QMRP over wireless link has been studied and compared with single path routing protocol AODV. On the contrary of the previous work that uses Max_Tx_out this work uses a more realistic measure Actual_Tx_out; actual transmission rate out from the MAC layer is used in order to better capture the channel dynamics to enhance the chances of QoS support over MANETs. In addition the average queuing delay is used in the delay computation instead of using current queue size solely along with queue occupancy factor. Moreover, a salvaging mechanism adopted similar to that in [21]. Also, a preemptive handoff mechanism is used to switch to paths with lower *Expected Path Delay* (*EPD*) value as in (1) through the use of an update packet to check the status of the paths, see section V for more details.

## III. Background and Terminology

### A. AODV

*AODV* is one of the most studied on demand reactive routing protocols [22]. Sequence number plays an important role in AODV and serves as a time stamp.

Each node maintains a monotonically increasing sequence number, every time a node generates a routing message it increments its sequence number. The node also keeps the highest known sequence number for each destination in its routing table. The highest sequence number means a fresher up to date route.

When a source node *S* needs to communicate with a destination node *D* in the network and *S* does not have a route to *D*, it initiates a route discovery process that starts by broadcasting a RREQ packet tagged with a sequence number to achieve limited flooding of the RREQ. Every node that receives the RREQ checks its routing table to see if it has a route to *D*. If it does, it sends a RREP back to *S*; otherwise it rebroadcasts the RREQ incrementing the hop count by one. This way, when a node receives several RREQ through multiple routes, it discards RREQs that result in a higher hop count. Intermediate nodes between *S* and *D* create an entry for the neighbor ID in its routing table from which the RREQ was received. The destination *D* responds to the first RREQ it receives by unicasting a RREP, intermediate nodes forward the RREP back to the source according to their routing table.

Every node maintains an entry in its routing table that updates the route expiry time. Every route in considered valid for a certain time after which the route entry is deleted from the routing table. Whenever a route is used to forward data packets the route expiry time is updated to the current time plus the Active Route Timeout. When a route expires the node deletes the entry for the route and invalidates it. When a link to the next hop is broken the node generates a Route ERRor (RERR) message to all nodes listed as active neighbors to the node in its routing table and invalidates all routes through the link [23][24] the node increments the sequence number and sets up the hop count to $\infty$ making AODV loop free at all times, for more details see[6].

### *B. Multipath and Disjointness*

Multi-path routing protocols are of special interest in mobile ad hoc network because of limited bandwidth of mobile nodes. Multi-path routing in Ad hoc networks permits the establishment of more than one path between source and destination to assure the QoS requirements. Because of dynamic nature of Ad hoc networks due to mobility and nodes joining and leaving the network, limited transmission range and limited source power of the nodes, multi-path routing is needed to increase network resilience and load balancing, which decreases congestion and bottlenecks, increases aggregate bandwidth, reduces end to end delay, delay variation and packet loss ratio (PLR) [25] [26] [27]. It was found in [27] that the performance of multipath outperforms multiple descriptions techniques.

Multi-path routing protocols could be classified into two types; node disjoint and link disjoint routes through the network. Node disjoint paths; which have no node and so no link in common provide higher degree of fault tolerance than link disjoint paths; which are paths that have no link in common. Since node failure in link disjoint can cause many links to fail while node failure in node disjoint will cause only one link to fail which means paths fail independently. However, node disjoint paths are harder to find, therefore, less abundant than link disjoint paths [25] [26]. Nonetheless, if the aim is to accomplish load balancing then node disjoint paths are more effective also node disjoint can increase the life time of the whole network by avoiding draining the resources of a node that is located in strategic location as node *I* in Fig. 1.

Suppose that node S needs a path to destination *D*, as can be seen *S* has two link disjoint paths to *D*; *S-A-I-E-D* and *S-B-I-F-D* but one node disjoint path to *D* since *I* is common node between the two paths. Since *I* is part of two link disjoint paths this can cause node *I* to reach exhaustion point and causes *I* to use its resources in a very short period of time which might cause node *I* to fail and hence all paths through *I* fail increasing the number of dropped packets and causing longer delays which may also lead to network partitioning this is similar when a node participating in more than one path move out of range due to mobility, hence the choice was to choose node disjoint multipath routing protocol.

### *C. Cross-Layer Design*

The exploitation of dependence between different layers of the protocol stack to maximize the performance gain is referred to as Cross-Layer Design. Taxonomy of cross-layer schemes is suggested based on the violations done to the layered approach [27].

Such a design has its pros and cons. On one hand, it violates the layered architecture compromising the independence of protocol design at one layer from other layers. In addition, many such violations will result in the collapse of the layered approach affecting system longevity [29]. On the other hand, nevertheless, cross-layer design addresses issues arising from the nature of the wireless medium which cannot be addressed otherwise; such issues include the broadcast nature of the wireless medium; the channel response time variations and the ability to receive multiple packets at the same time (e.g., see [30]).

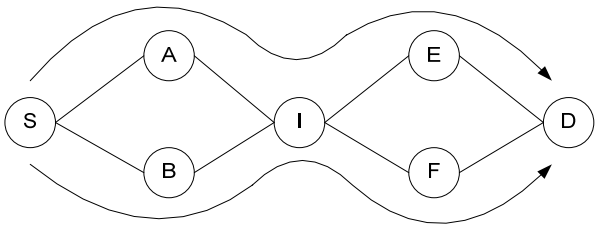



Figure 1. Node vs. Link disjointness

Supporting QoS over ad hoc networks while being aware of resources availability cannot be accomplished without a cross layer design as reported by different studies [31][32]. In addition, many schemes involving a cross-layer design have reported better performance when compared to the traditional approaches [33][34][35] [36][37][38][39].

## IV. Problem Statement and Multipath Approach

The objective is to develop a multipath routing protocol, the emphasis is on providing QoS for Qos-stringent applications rather than on accommodating many connections and not fulfilling the QoS requirements. However, the limited resources and the dynamics of MANETs in addition to the wireless medium unpredictability make the task even harder to achieve. The question has two parts; (i) what approach to take and (ii) how to apply it.

To answer the first part, a multipath approach will be taking. The multipath aim is to find more than one path between Source-destination pair to maintain QoS assurance for the life time of a connection. Multipath increases network resilience (i.e., reliability through fault tolerance), load balancing which decreases congestion and bottlenecks also, it increases aggregate bandwidth, reduces end to end (E2E) delay, delay variation and packet loss ratio (PLR). Additional benefits of multipath routing include reduction of computation time that CPUs require, high call admission ratio in voice applications [1] [2].

The main quality impairment for delay sensitive application voice is delay. ITU-T G.114 recommends [39] the one way transmission time between 0-150 ms delay is acceptable for voice. The time when a frame is generated at the source until it reaches the destination is the End-to-End (E2E) delay. E2E delay consists of packetization, queuing, propagation, transmission, and play out delay. Play out delay is the time a packet spends in the buffer at the destination for smooth play out as in Fig. 2. [40].

Bit Error Rate (BER) as a function of Signal to Noise Ratio (SNR) which is a function of distance that translates into modulation reflects the link quality and stability. Figure 3 [42] shows BER vs. SNR for five digital modulation schemes: binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), and quadrature amplitude modulation (QAM) with different number of bits per symbol.

How to employ multipath to accomplish QoS, will be through cross layer approach. The cooperation and the interaction between different layers; by extracting some crucial parameters from the physical and MAC layers and feed them into the routing layer.

## V. System Architecture and Protocol Dynamics

The proposed architecture is general and works with any application as well as any routing protocol that supports multipath as shown in Fig. 4. The proposed QMRP extracts the SNR from the physical layer and passes it to the MAC layer where the later compute the actual transmission rate out from the node and the queue size then pass these values to the routing layer where delay computation takes place to find the path with the lowest delay between source-destination pair to assure QoS based on link quality. Fig. 5 shows the dynamics of the protocol.

### *A. Protocol Overview*

This section describes the details of QMRP protocol which computes multiple node disjoint paths based on the feedback from the physical and MAC layers. QMRP improves AODV significantly by modifying the phases of route discovery, route selection and route maintenance. In this work route broadcasting packet refers to RREQ/RREP.

The QMRP protocol establishes multiple node-disjoint paths that will experience the lowest delay. Most delay-aware routing protocols in literature use the current delay or history to estimate end-to-end (E2E) delay as a metric. However, this is not an accurate measure of the delay that is going to be experienced by the route-requesting node since this node will increase the total network load. Once the network load increases, E2E delay that was obtained through route broadcasting is no longer accurate. Introducing the projected increase in load into the computation of delay a more accurate delay value will be obtained once the node starts injecting its traffic into the network. The protocol phases are explained below.

### *B. Route Discovery: Reverse Path*

The route discovery process of QMRP starts when node *S* needs to communicate with another node *D* and *S* does not already have a path to *D*, *S* broadcasts a RREQ tagged with a sequence number to achieve limited flooding of the RREQ. In AODV every node that receives the RREQ rebroadcasts the RREQ incrementing the hop count by one. When a node receives several copies of the same RREQ, it uses only the first copy to form the reverse path; all duplicates that arrive later are discarded. Nonetheless, since the aim is to find multiple paths these duplicates of route broadcasting could be utilized to establish multiple paths, however, only those route broadcasting that guarantee loop freedom and node disjointness will be used to establish reverse paths.

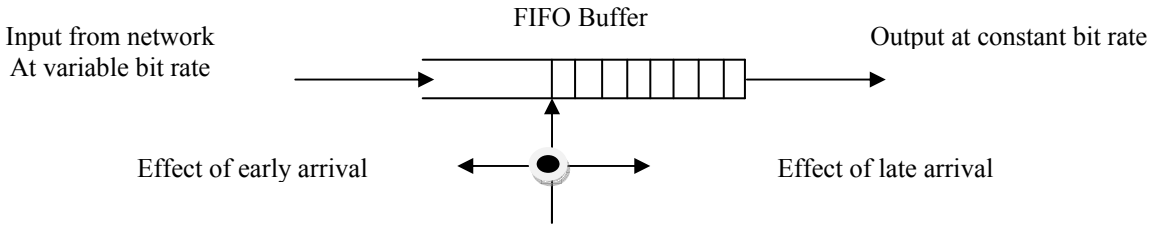



Figure 2. Play out buffer

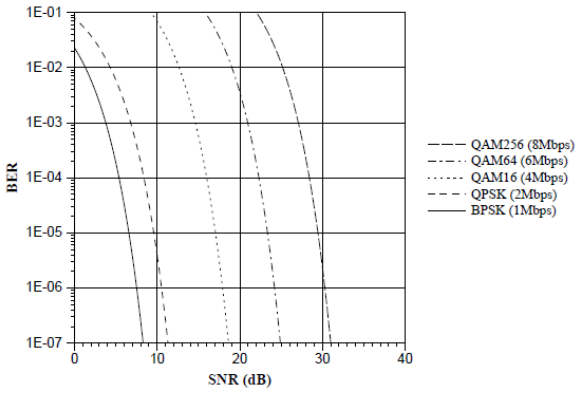



Figure 3. BER vs. SNR

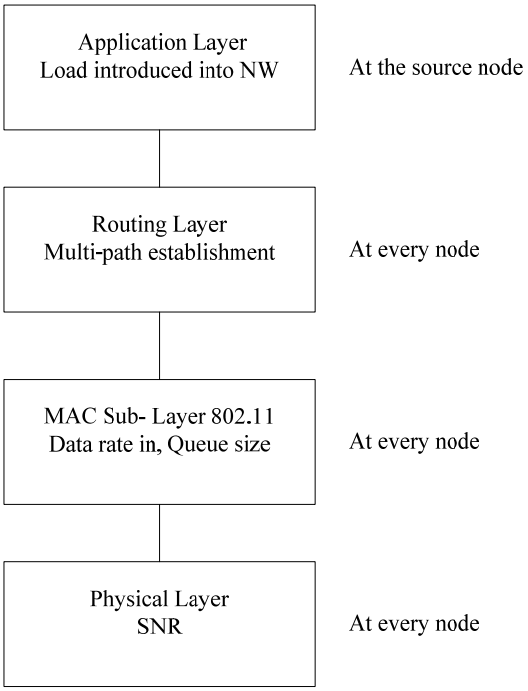



Figure 4. System Architecture

QMRP introduces two additional fields to the route broadcasting packet the *Expected Path Delay* (*EPD*) field; which is the cumulative delay up to and including the node itself and a *load* field; which is the new load that will be added to the network by a node requesting a path to a destination. *EPD* is initialized to zero while the *load* initialized to the new amount of traffic that will be added into the network by the source requesting a path to a destination.

To guarantee loop freedom, (i) QMRP preserve the following update rule; keep paths for the highest known destination sequence number. When a route broadcasting packet received by a node with higher sequence number all paths correspond to lower sequence number will be invalidated. (ii) QMRP guarantees loop freedom by utilizing a neighbor hop list; a list of one hop a way neighboring node of the source node that initiates a route discovery process by generating a RREQ requesting a path to a destination that is the first hop traversed by a RREQ. (iii) Maintaining the invariant that all nodes can only broadcast one copy of a RREQ per unique neighbor hop. A maximum of three route broadcasting with unique neighbor hop is allowed.

Obtaining the neighbor hop list is as follows: if an intermediate node, *I*, receives a RREQ and the hop count is zero, the node checks the source of the RREQ and the node from which it received the RREQ if they match it increments the hop count and adds itself into the neighbor hop field of the RREQ in addition, an entry for the node from which it received the RREQ is added. Every subsequent node on the path maintains a list of neighbors called neighbor hop list associated with every RREQ message received. When a RREQ is received the neighbor hop field of the RREQ is checked against the neighbor hop list before adding an entry for the path into the node's routing table. Additionally, before rebroadcasting the RREQ, intermediate node of *S*, *I*, increments the hop count of the RREQ and updates *EPD* field with its computed delay as in equation (1).

When a node receives duplicates of the RREQ with unique neighbor hop it records the information from these packets in a RREQ table which contains the following fields, source_id, dest_id, neighborhop_list which contains neighbor Id, *EPD*, last hop, Exp_timer. Where the source_id is the source that generates the RREQ, dest_id is the destination to which the RREQ is intended, neighbohop_list is the neighbor hop list, *EPD* for each neighbor in the list, last hop is the last hop on the path and Exp_timer is the expiration timer.

To explain how QMRP guarantees loop freedom using the neighbor hop list Fig. 6 shows the source node *S* wants to communicate with destination node *D*, *S* broadcasts a RREQ packet to its neighboring node within transmission range, in this case nodes *A* and *B* will receive the RREQ. Each of *A* and *B* will add itself in the neighbor hop field in the RREQ and rebroadcast the RREQ, node *I* receives both RREQ from *A* and *B* and accepts both of them since they arrive via different neighbor hop of the source node *S*, now node *C* receives the RREQ from node *B* with neighbor hop field *B* and broadcast it, node *I* receives it and checks the neighbor hop field which is *B* in this case against its neighbor hop list, however, node *I* has already received a RREQ with the same neighbor hop *B* so it discards the packet and don't broadcast it any further.

In order to update the node's routing table, an entry is only added or updated if route broadcasting satisfies any one of the following criteria: (i) No route entry exists for the originator of the route broadcasting. (ii) Sequence Number of the route broadcasting packet is greater than the sequence number of the existing route entry. (iii) Sequence numbers are equal and the *EPD* of the route broadcasting is less than the *EPD* of the existing route entry and number of valid RREQs is less than three.

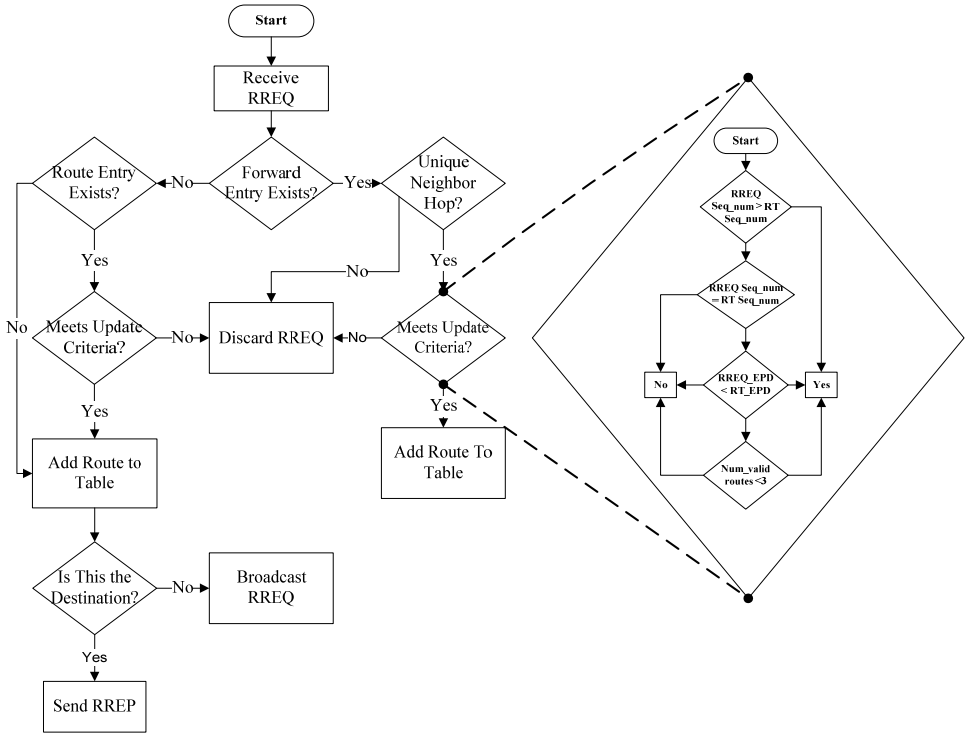



Figure 5. Protocol Dynamics

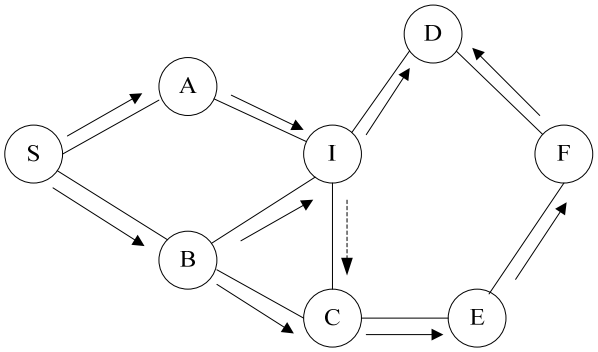



Figure 6. Neighbor hop list and loop freedom

### C. Route Reply: Forward Path

If a node receives a RREQ and it is the destination of the RREQ, the node adds/updates the entry in its routing table and generates a RREP. The RREP is unicasted back to the source node and the *EPD* field of the RREP is initialized to zero. Subsequent nodes that receive the RREP maintain their routing table according to the conditions specified, increment the *EPD* field of the RREP with their computed delay up to and including the node itself, and then forward the RREP to the next hop towards the source node that was found through the reverse path during route discovery process based on the minimum *EPD;* so the node checks its RREQ table and forward the RREP to the next hop with the lowest *EPD.*

When the destination receives subsequent RREQs from different last hops it generates a RREP to each distinct last hop and per unique neighbor hop. Additionally, all intermediate nodes can only forward one copy of the RREP per source and per destination sequence number; this is to guarantees node disjointness of the paths. If no reverse path is available, the RREP is discarded; this is also in case the node is already participating in an active path for the same source-destination pair. When the source receives all RREPs back from the destination; a maximum of three paths in this study, it uses the path with the lowest *EPD* value and saves the other two as a backup paths in case the primary path fails for any reason.

The *EPD* includes various parameters from the MAC; data rate received, current queue size, and SNR from the PHY layer that is reflected in the actual transmission out rate. The equation for computing the *EPD* is give by (1):

$$EPD = \sum_{i=0}^{n} \left[ \overline{D}_i + \frac{\Delta t * (DR_i + l - Actual\_Tx_{out(i)})}{Actual\_Tx_{out(i)}} \right] \quad (1)$$

Where $\overline{D}_i$ is the average queuing delay at a node and is given by (2)

$$\overline{D}_i = \alpha * \overline{D}_{j-1} + (1 - \alpha) * \overline{D}_j \quad (2)$$

Where α is the queue occupancy and is given by (3)

$$\alpha = \frac{queue_{size} - queue_{length}}{queue_{size}} \quad (3)$$

*i*: a node along the path

*queue_size*: is the size of the queue at node *i*

*queue_length:* is the length of the queue at node *i*

*j:* is the current period;

$DR_i$: data rate calculated based on all traffic received at node *i*, this parameter is passed from the MAC.

$\Delta t$ : time difference between the current time and an arbitrary time after the new load has been introduced into the network, this can vary according to how long routes are expected to remain active based on mobility and active route timeout value, for simplicity purposes $\Delta t$ is assumed to be 2 seconds.

$l$ : is the proposed new traffic load that is added by the source initiating a route discovery process into the network.

$Max\_Tx_{out(i)}$: represents the maximum data rate that a node can transmit at and given is by:

$$Max_{Tx_{out}} = Data\,Tx_{Rate} * \beta * (1 - BER) \quad (4)$$

In this study we are taking the Actual _Tx_out actual transmission out from a node that is extracted from the MAC layer, which is based on the above equation (4).

$Data\,Tx\_Rate$ : rate at which a node is able to transmit/receive;

$\beta$ : network efficiency factor, which is typically between 0.7-0.8.

BER: Bit Error Rate.

C-Route Maintenance

Route maintenance in QMRP is an extension to that of AODV and is achieved by the means of generating a Route ERRor (RERR) packet. When an intermediate node, *I*, discovers link/node failure; due to mobility, undetected hello packet, etc…, it generates a RERR packet. The RERR packet propagates towards all nodes that have a path through the failed link and invalidates all available paths in all nodes along the way that have a path through the failed link. When the RERR packet reaches the source and the source still in need for a path to the destination it switches to the second available path it the path list at the source node. If all paths are invalid and the source is still in need for a path to the same destination the source stars a new route discovery process as explained previously in A.

Each path has an ID that is a combination of the neighbor hop of the source as well as the last hop on the path which is the previous hop to the destination. At periodic interval every $\frac{\Delta t}{2}$ the source unicasts an update packet to check the status of the path with *EPD* value initialized to zero towards the destination, when the destination receives the update packet it initializes the *EPD* to zero and sends it back to the source when the source node receives the update packet it checks the *EPD* field against the other *EPD* for alternate paths in its routing table. The preemptive handoff basically takes place when the path *EPD* value increases and the routing table has a lower *EPD* for another path it switches to the path with lower *EPD*; this is in order to maintain QoS for the life time of a connection and always use the best available path in terms of *EPD* as the primary path for data transmission Also, QMRP uses a salvaging mechanism where packets transmitted over failed path are retransmitted over back up paths.

## VI. Simulation Environment and Evaluation

### A. Simulation Environment

OPNET [43] simulation package is used to evaluate the performance of the proposed QMRP protocol and compare it with the AODV. The random way point mobility [44] model is used as the mobility model. Constant Bit Rate (CBR) where generated by all nodes of size 512 bytes each plus headers of different layers (UDP/IP/MAC). All different traffic connections set up between source-destination pairs are at random. Each reading across all scenarios and experiments is the average of 10 runs. The rest of the simulation environment parameters are summarized in the Table I.

TABLE I

Simulation parameters

| Parameter | Value |
|---|---|
| Transmission Range | 250 m |
| Simulation Time | 500s |
| Simulation Area | 1200 m x 600 m |
| Number of nodes | 50 |
| Number of Transmitting nodes | Varies between 5-40 |
| Traffic Type | CBR |
| Packets Generation Rate | 3 packets/sec |
| Packet size | 512 bytes |
| Mobility Model | Random Waypoint |
| Traffic Model | Spread Randomly |
| Speed | Varies between 5-25 (m/s) |
| Pause time | 0 s |
| Channel rate | 2 Mbps |
| Packet Reception Power Threshold | -95 |
| Hello loss | 2 |
| Wireless channel model | Error free, bidirectional |
| Queue size | 64 packets |

### B. Performance metrics

The following four key metrics considered to evaluate the performance of the protocols.

1) Average E2E delay: is the difference in time between the reception of a packet by the destination and the moment it was generated by the source; it includes all possible delays encountered by a packet.

2) Packet Delivery Fraction: is the amount of traffic received successfully at the destination as a fraction of the traffic generated by the source node.

3) Route Discovery Frequency: number of all RREQ packets generated per sec by all sources

4) Normalized Routing Load: is the number of control packets transmitted per data packet delivered at the destination, each hop wise communication is counted as a new transmission.

### C. Results and Discussion

1- Varying mobility: (experiment I)

The first set of experiments Fig. (7-10) show the four key performance metrics as a function of mobility; speed in m/s. Traffic is 20 connections through the network and the speed varies between 0 static to 25 m/s, the rest of parameters are given in table I. To analyze the performance of all protocols more accurately, we turned off one node and turned on another node randomly every 50 seconds throughout the simulation that sums up to a total of 10 nodes in the network to account for node failure due to power exhaustion; this will cause paths to fail through that node.

Fig. 7 depicts average E2E delay. QMRP has about 32% improvement over AODV in average E2E delay at high speed; this is due to the fact that AODV path selection criteria is based on minimum number of hops between

source-destination pair regardless of the nodes status along the path. On the other hand paths in QMRP are established based on the *EPD*. QMRP avoids congested nodes by choosing paths based on minimum *EPD* instead. QMRP considers delay encountered at each node into the computation of the *EPD* when establishing paths between a source–destination pair. The *EPD* includes not only the current delay, but also the expected delay due to the new load introduced by the source node into the network as well as queuing delay the most contributive factor in E2E delay. Due to mobility frequent link breaks occur in an environment such as MANET which causes AODV's need to initiate additional route discovery process/es to continue transmission. During the search for a new route packets are being delayed and/or dropped.

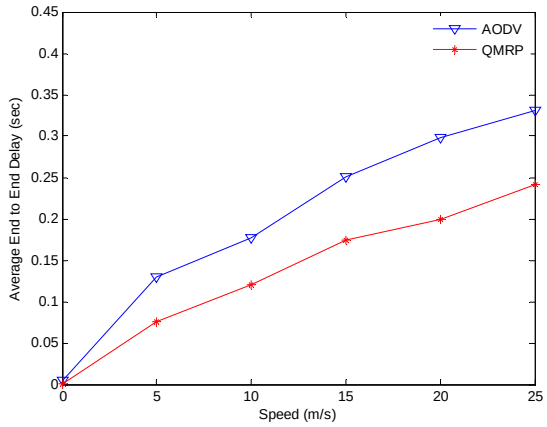



Figure 7. Average End to End Delay

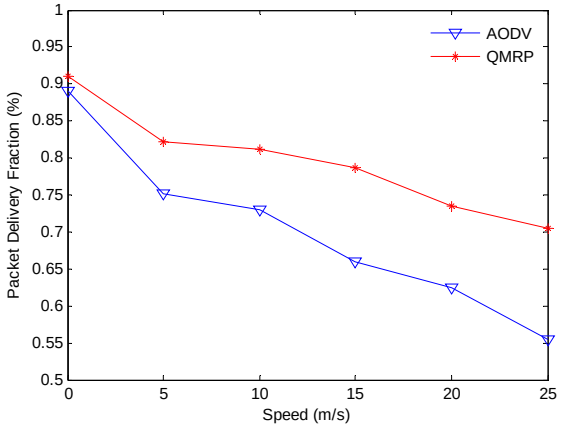



Figure 8. Packet Delivery Fraction

Fig. 8 shows packet delivery fraction; the difference between the single path AODV and QMRP is about 27% increase in successfully transmitted packets this is due to the fact that QMRP is a multipath protocol and takes into consideration the channel reliability through the SNR from the PHY layer and queuing delays at each node along the path as per to equation (1).

Most of the proposed routing protocols such as AODV don't consider channel conditions, load balancing causing heavily loaded nodes along the path between source-destination. As a result heavily loaded nodes with longer queues will cause longer delays.

Also this may cause congested nodes to exhaust their power energy quickly causing failure of a session, more dropped packets and may lead to network partitioning. QMRP decreases the number of dropped packets due to the fact that the protocol allows self load balancing by avoiding bottle necks; along the path when establishing multiple paths based on many factors included in (1). Therefore, packet losses are due to mobility as well as node/link failure as mentioned previously.

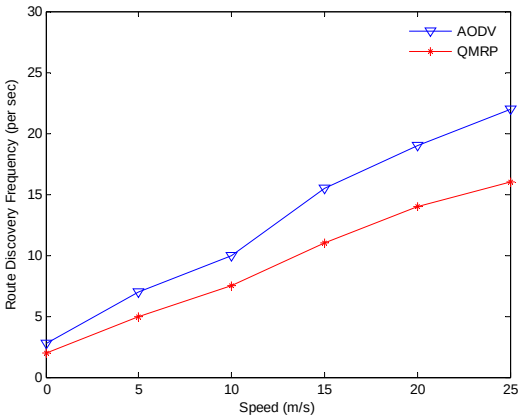



Figure 9. Route Discovery Frequency

As can be seen in Fig. 9 QMRP has up to 35% less than AODV in terms of route discovery frequency. This is due to the nature of the single path AODV vs. multipath in QMRP as well as node disjointness that guarantees nodes/links fail independently. In addition path selection criterion in AODV can increase heavily loaded nodes along minimum hop count path which will also increase the probability of node failure and link breaks accordingly. Also, as mobility increases that will cause links to fail which means AODV has to initiate a new route discovery process during this search for an alternative path, packets are being queued; delayed and/ or dropped.

Fig. 10 shows the normalized routing load, even though QMRP has less route discovery frequency than AODV, QMRP has more overhead per route discovery process due to more forward of control packets. However this overhead is relatively low.

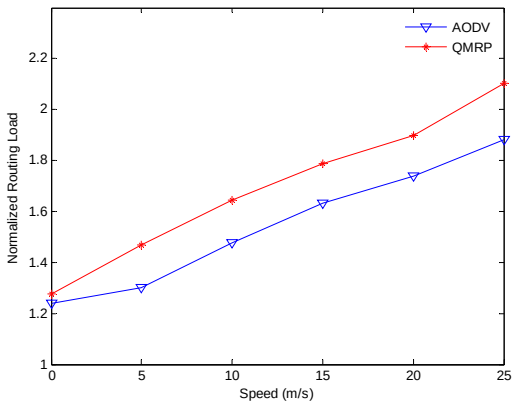



Figure 10. Normalized Routing Load

2- Varying number of connections: (experiment II)

In this second set of experiments, Fig. (11-14) the number of connections is varied between 0-40 connections while fixing the speed at 5 (m/s). Increasing number of connections will test the routing protocol under tense conditions. The rest of parameters remain the same as in experiment I.

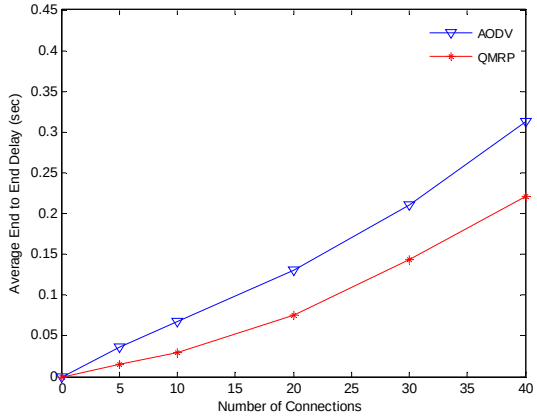



Figure 11. Average End to End Delay

As the number of connections increases so does the average E2E delay for both protocols as in Fig. 11, however, QMRP has the lower delay; this is due to the use of *EPD* in choosing the route. As (1) shows it has parameters from the PHY and MAC layers, so the path selection criterion chooses more stable and reliable links, this means packets along more reliable paths encounters less E2E delay. In addition QMRP does some indirect self load balancing strategy by avoiding congested nodes with longer queues.

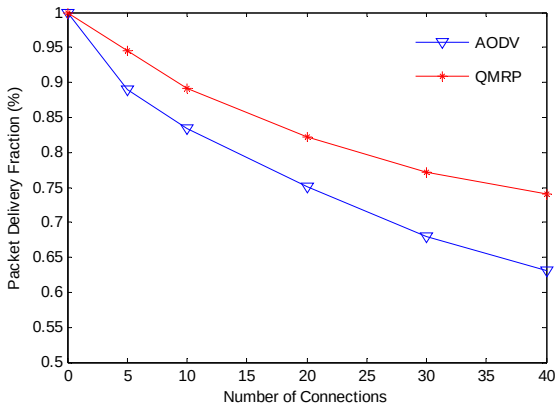



Figure 12. Packet Delivery Fraction

Packet delivery fraction shown in Fig. 12 has the same behavior; as number of connections increases packet loss increases again with QMRP more packet delivered successfully, this is because QMRP has more reliable and stable paths so the probability of dropping packets decreases as compare to AODV. In the case of AODV which is based on minimum number of hops per route without considering the node's queue status and channel conditions causing congested nodes along the path which translates into longer delays and more dropped packets.

Fig. 13 shows that after 20 connections the difference increases significantly between AODV and QMRP in terms of route discovery frequency as a result of node failure and/or link break which means for the single path AODV a new route discovery process.

Even though multipath routing protocols less route discovery processes, still it has more routing overhead over all as compare to single path routing protocol such as AODV which is illustrated in Fig. 14. QMRP has more routing overhead per each route discovery process, yet this difference is insignificant; less than 10% only when number of connections is more than 30 connections.

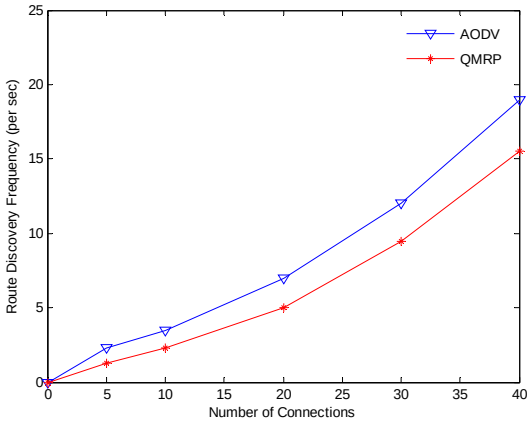



Figure 13. Route Discovery Frequency

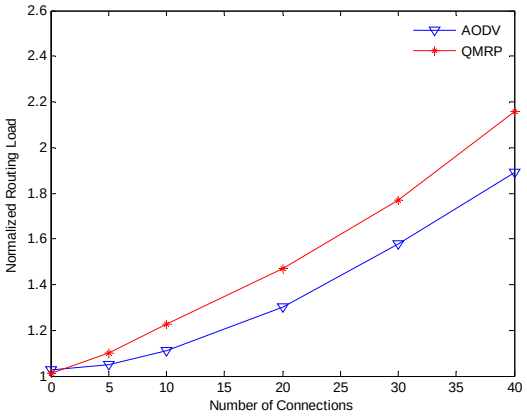



Figure 14. Normalized Routing Load

## VII. Conclution and Future Work

In a dynamic environment like MANET on demand multipath routing protocols can achieve many aims such as but not limited to lower delay, load balancing, network resilience through network fault tolerance and increase packet delivery fraction.

This paper proposes QoS-Aware multipath routing protocol (QMRP) a node-disjoint protocol that considers channel conditions when establishing multipath between source-destination pair in wireless ad hoc networks to overcome the limitations of other single path; AODV.

QMRP uses the *EPD* as a metric to choose the route which takes into consideration the SNR at the physical layer as well as the actual data rate from the MAC layer in addition to the node's average queuing delay to reflect the link quality and the medium utilization around the node, respectively. Without loss of generality QMRP introduces the new load -significant in case of delay sensitive applications-; that is introduced by the node requesting a path to a destination in the computation of the *EPD* to capture the real channel conditions. QMRP does self load balancing by avoiding congested nodes along the path by avoiding nodes with longer queues. As the results show, QMRP protocol outperforms the AODV protocol in terms of average E2E, packet delivery fraction, route discovery frequency. On the other hand, QMRP has insignificance more routing overhead than that of AODV.

Future work will focus on the analytical and statistical analysis for QMRP vs. other routing protocols. Also more investigation of the performance of QMRP with IEEE 802.11e since it was developed to offers QoS capabilities to WLAN. More thorough research of the protocol is needed under Rayleigh fading channel since this is usually the case for MANET's environment.

## REFERENCES


[1] C. Siva Ram Murthy & B.S. Manoj, "Ad Hoc Wireless Networks Architecture and Protocols," Prentice Hall, 2004.

[2] S. Chen and K. Nahrstedt, "Distributed quality-of-service in ad hoc networks,"IEEE J. Sel. Areas Communications., vol. 17, no. 8, Aug. 1999.

[3] Pradeep Macharla, Rakesh Kumar, Anil Kumar Sarje, "A QoS routing protocol for delay-sensitive applications in mobile ad hoc networks," COMSWARE pp. 720-727, 2008

[4] [4]Chen, W. Wu Z. Li, "Multipath Routing Modeling in Ad Hoc Networks," Proc. of IEEE ICC, pp. 2974-2978. May, 2005.

[5] Adibi, S.; Erfani, S.; , "A multipath routing survey for mobile ad-hoc networks," Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE , vol.2, no., pp. 984- 988, Jan. 2006

[6] C. E. Perkins and E. M. Royer. Ad Hoc On-Demand Distance Vector Routing. Proc. of the IEEE WMCSA, pp. 90–100, 1999.

[7] M.K. Marina, S.R. Das, "On-demand Multipath Distance Vector Routing for Ad hoc Networks," Proceedings of the Ninth International Conference on Network Protocols, Washington, DC, USA, pp.14-23, 2001.

[8] Z. Ye, S. V. Krishnamurthy, and S. K. Tripathi, "A Framework for reliable routing in mobile ad hoc networks," Proceedings of IEEE INFOCOM, San Francisco, CA, USA, pp. 270–280, March 2003.

[9] Chandramouli, Balasubramanian, J.J. Garcia-Luna-Aceves "Shortest Multi-path Routing Using Labeled Distances" IEEE International Conference on Mobile Ad-hoc and Sensor Systems, 2004.

[10] Sung-Ju Lee and Mario Gerla, "AODV-BR: Backup Routing in Ad hoc Networks", Wireless Communications and Networking Conference, 2000. WCNC 2000 IEEE, Volume 3, pp. 1311-1316, September 2000.

[11] C.E. Perkins, and E.M. Belding-Royer, "Quality of Service for Ad Hoc On Demand Distance Vector Routing," draft-perkins-manet-aodvqos-02.txt, Mobile Ad Hoc Networking Working Group Internet Draft, 14 October 2003.

[12] A. Valera, W. Seah, and S. Rao, "Cooperative packet caching and shortest multipath routing in mobile ad hoc networks," Proceedings of IEEE INFOCOM, San Francisco, CA, USA, pp. 260–269, March 2003.

[13] Lee, S.-J.; Gerla, M.; , "Split multipath routing with maximally disjoint paths in ad hoc networks," Communications, 2001. ICC 2001. IEEE International Conference on , vol.10, no., pp.3201-3205 vol.10, 2001

[14] D., J., D, M.: Dynamic source routing in ad hoc wireless networks. Mobile Computing (ed. T. Imielinski and H. Korth), Kluwer Academic Publishers. Dordrecht, Netherlands. (1996).

[15] V. Srivastava and M. Motani, "Cross-layer design: A survey and the road ahead," IEEE Communications Magazine, 43 (12), pp. 112–119, 2005.

[16] V. Kawadia and P. R. Kumar, "A cautionary perspective on cross-layer design," IEEE Communications Magazine, 12 (1), pp. 3–11, 2005.

[17] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross layer optimization in wireless networks," IEEE Journal on Selected Areas in Communications, 24(8), pp. 1452–1463, 2006.

[18] H. Sun and H. D. Hughes, "Adaptive QoS routing by cross-layer cooperation in ad hoc networks," EURASIP Journal on Wireless Communications and Networking 5, pp. 661–671, 2005.

[19] R. Draves, J. Padhye, and B. Zill, "Comparison of routing metrics for static multi-hop wireless networks," Proceedings of ACM SIGCOMM, Portland, Oregon, USA, pp. 133–144, August 2004.

[20] Mu'ath Obaidat, M. Ali, Ihsan Shahwan, M.S. Obaidat, Suhaib Obeidat " QoS-Aware Multipath Routing Protocol for Delay Sensitive Applications in MANETs: A Cross-Layer Approach", Wireless Information Networks and Systems, ICETE, Seville, Spain, July 2011

[21] Johnson DB, Maltz DA, Hu Y. The dynamic source routing protocol for mobile ad hoc networks (DSR). http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-10.txt, July 2004. IETF Internet Draft (work in progress).

[22] "Ad hoc on demand distance vector (AODV) routing," IETF Internet Draft, Work in Progress, [Online]. Available: ttp://www.ietf.org/internet- drafts/draft-ietf-manet-adov.03.txt, Jun. 1999

[23] C.Siva Rama Murthy and B.S. Manoj, "Adhoc Wireless Networks: Architectures and Protocols", Second Edition, Prentice Hall.

[24] T.G.Basavaraju and Subir Kumar Sarkar, "Adhoc Mobile Wireless Networks: Principles, Protocols and Applications", Auerbach Publications, 2008.

[25] S. Mueller, R. P. Tsang, and D. Ghosal. Multipath Routing in Mobile Ad Hoc Networks: Issues and Challenges. Performance Tools and Applications to Networked Systems, Lecture Notes in Computer Science Vol. 2965, Springer 2004, pp. 209--234.

[26] G. Parissidis, V. Lenders, M. May, and B. Plattner. Multi-path routing protocols in wireless mobile ad hoc networks: A quantitative comparison. In NEW2AN, 2006

[27] Jagadeesh Balam, and Jerry D. Gibson, Multiple descriptions and path diversity for voice communications over wireless mesh networks. In IEEE Transactions on multimedia, vol. 9, no. 5, August 2007

[28] V. Srivastava and M. Motani. Cross-layer design: a survey and the road ahead. Communications Magazine, IEEE, 43(12):112–119, 2005.

[29] V. Kawadia and P. R. Kumar. A cautionary perspective on cross layer design. IEEE Wireless Communications, 12(1):3–11, 2005.
[30] L. Tong, V. Naware, and P. Venkitasubramaniam. Signal processing in random access. In IEEE Signal Processing, volume 21, pages 29–39, 2004.
[31] J. Al-karaki and A. E. Kamal. Quality of Service Routing in Mobile Ad hoc Networks: Current and Future Trends. CRC Publishers, 2004.
[32] A. Goldsmith and S. B. Wicker. Design challenges for energy-constrained ad hoc wireless networks. In IEEE Wireless Communications, pages 8–27, 2002.
[33] K. Chen, S. H. Shah, and K. Nahrstedt. Cross-layer design for data accessibility in mobile ad hoc networks. In Wireless Personal Communications, volume 21, 2002.
[34] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, andW. Kellerer. Application-driven cross-layer optimization for video streaming over wireless networks. Communications Magazine, IEEE, 44(1):122–130, 2006.
[35] B. Raman, P. Bhagwat, and S. Seshan. Arguments for cross-layer optimizations in bluetooth scatternets. In IEEE 2001 Symposium on Applications and the Internet, pages 176–184, 2001.
[36] D. Wu, S. Ci, and H. Wang. Cross-layer optimization for video summary transmission over wireless networks. Selected Areas in Communications, IEEE Journal on, 25(4):841–850, 2007.
[37] T. Yoo, E. Setton, X. Zhu, A. Goldsmith, and B. Girod. Cross-layer design for video streaming over wireless ad hoc networks. In IEEE Wireless Communications Magazine, volume 12, pages 59–65, 2005.
[38] W. Yuan, K. Nahrstedt, S. V. Adve, D. L. Jones, and R. H. Kravets. Design and evaluation of a crosslayer adaptation framework for mobile multimedia systems. In SPIE/ACM Multimedia and Networking Conference (MMCN), 2003.
[39] W. H. Yuen, H. Lee, and T. D. Andersen. “A simple and effective cross layer networking system for mobile ad hoc networks.” IEEE International Symposium on personal, indoor, and mobile radio communications (PIMRC2002), pages 1952–1956, 2002.
[40] ITU-T, “One-way transmission time,” in ITU Recommendation G.114, 1996.
[41] F. Halsall, Multimedia Communications. Addison-Wesley,2001.
[42] G. Holland, N. Vaidya, and P. Bahl, A rate-adaptive MAC protocol for multi-hop wireless networks, in Proceedings of ACM MOBICOM’01, 2001.
[43] www.OPNET.com
[44] D.B. Johnson and D.A. Maltz. Dynamic source routing in ad hoc wireless networks, in: T. Imielinski and H. Korth, editors, Mobile Computing, volume 353 (1996) pp. 153–181. Kluwer Academic Publishers

# Fingerprint Indoor Position System Based on Bitcloud and Openmac

José A. Gómez, A. Verónica Medina, Enrique Dorronzoro, Octavio Rivera and Sergio Martín
Departamento Tecnología Electrónica,Seville University, ETSI-INF, Avda. Reina Mercedes s/n, Seville, Spain Email: (jgomezdte, vmedina)@us.es,( enriquedz, octavio)@dte.us.es,smartin@us.es

***Abstract*—This paper presents a research and a development of a fingerprint-indoor-positioning system using the Received Signal Strength Indication (RSSI) of a Wireless Sensor Network (WSN). The WSN implementation is based on two different protocol stacks: BitCloud and OpenMAC, a certified ZigBee Compliant Platform (ZCP) and an IEEE 802.15.4 embedded software implementation respectively, both from Atmel, and the system uses two different fingerprint algorithms, Simple and Centroid. A comparative analysis of both algorithms using both protocol stacks implementations have been performed to ascertain the best WSN protocol stack and the best algorithm for positioning purposes.**

***Index Terms*— IEEE 802.15.4, RSSI, Centroid, Indoor position, ZigBee, WSN, BitCloud, OpenMAC**

## I. INTRODUCTION

WSNs are present in many applications. Examples of WSN applications are found in Ambient Living [1], [2], [3], [4] or Smart building [5], [6], [7], [8], [9] researching fields for solving data acquisition process. Depending on its applications, ambient or user sensors and actuators can be used for making decisions.

The knowledge of a subject's position is very useful in these kinds of systems because depending on it the decisions to be made are different. As stated in [11] and [12], an amount of indoor location tracking systems have been proposed in the literature, based on Radio Frecuency (RF) signals, ultrasound, infrared, or some combination of modalities.

Using RF signal strength, it's possible to determine the location of a mobile node with an acceptable accuracy. Given a model of radio signal propagation in a building or other environment, received signal strength can be used to estimate the distance from a transmitter to a receiver, and thereby triangulate the position of a mobile node. However, this approach requires detailed models of RF propagation and does not account for variations in receiver sensitivity and orientation.

An alternative approach is to use empirical measurements of received radio signals, known as RSSI, Receiver Signal Strength Indicator, to estimate location. By recording a database of radio ''signatures'' along with their known locations, a mobile node position can be estimated by acquiring the actual signature and comparing it to the known signatures in the database, also known as fingerprints. A weighting scheme can be used to estimate location when multiple signatures are close to the acquired signature.

All of these systems require the signature database to be manually collected prior to system installation, and rely on a central server (or the user's mobile node) to perform the location calculation. Several systems have demonstrated the viability of this approach, one of those is MoteTrack [11], [12].

Motetrack's basic location estimation uses a signature based approach that is largely similar to RADAR [10] that obtains a 75th percentile location error of just under 5 m, but in MoteTrack decreased the location error by 1/3.

We have implemented a similar system to MoteTrack, a signature-based localization scheme, but using other motes, Meshnetics´ ones (http://www.meshnetics.com/), that use different RCB (microcontroller and transceiver) and, also, different software, the BitCloud Stack, [13], a ZigBee PRO certified platform. (Atmel acquires MeshNetics´ ZigBee Intellectual Properties). The BitCloud-stack system has been tested and the same precision as MoteTrack has been obtained, but an amount of drawbacks have been found while its implementation was performed.

These drawbacks will be exposed later but in order to solve them, we have been working in the same way by using other WSN stack called OpenMAC, an IEEE 802.15.4 MAC level implementation from Meshnetics too [15] instead of BitCloud. This lower level protocol has enabled to implement applications by taking control of the RSSI measurements.

Two different fingerprint positioning algorithms, Simple and Centroid, have been implemented too in a desktop application, and a comparative analysis has been made in order to study the improvements from the OpenMAC protocol stack relative to the BitCloud one as well as from the first algorithm to the second one.

In Section 2 an overview of the system is presented. In Section 3 the used hardware is shown BitCloud Implementation is explained in Section 4. OpenMAC solution is presented in section 5. Finally conclusions are established in section 6.

## II. SYSTEM OVERVIEW

An overview of the system is shown in Fig. 1. In our system, a building or an area is populated with a number


doi: 10.4304/jnw.8.1.37-44

of MeshNetics´s motes acting as fixed nodes, one of them acting as coordinator, C making up the WSN.

Fixed nodes send to C periodic beacon messages, beacon2. Each beacon2 sent by a fixed node, consists of an n-tuple of the format {MobileID, RSSI}, where n is the number of mobile nodes. MobileID is a unique identifier of a mobile node, and RSSI is the received signal strength from the last beacon message, beacon1, sent by aforementioned mobile node and received by the fixed node.

The location estimation problem consists of a two-phase process: an offline collection of reference signatures followed by an online location estimation. As in other signature-based systems, the reference signature database (the off-line phase) is acquired manually by a user with a mobile node and a PC connected to C. The reference signature database consists of a number of reference signatures. Each reference signature, shown as black dots in Fig. 1, is formed by a set of signature tuples of the format {sourceID, meanRSSI}, where sourceID is the fixed node ID and meanRSSI is the mean RSSI of a set of beacon messages received over some time interval. Each signature is mapped to a known location by the user acquiring the signature database (P1-P5 in Fig. 1).

In the online phase, given a mobile node's received signature, s, received from the fixed nodes, and the reference signature set R, the mobile node's location can be estimated applying one of the two algorithms described below.

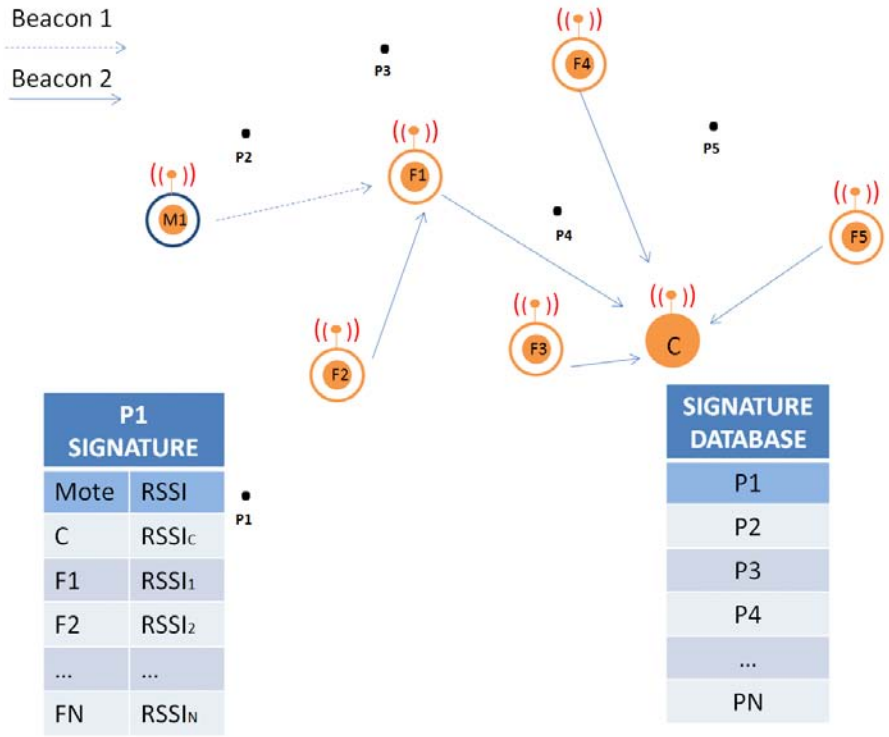



Figure 1. System Overview. M1 is a mobile node, F1-F5 are fixed nodes, and C is the coordinator, also a fixed node. M1 periodically sends a beacon message, beacon 1, to inform the others node that is present, all fixed node that receives it, save the RSSI of that message in a table. Fixed node periodically sends a message to C, beacon 2, to inform about the RSSI that they receive from mobiles node, M1 in this case.

The first step is to compute the signature distances, from s to each reference signature $r_i \in R$. We employ the Manhattan distance metric,

$$M(r,s) = \sum_{t \in T} \left| RSSI(t)r - RSSI(t)s \right| \quad (1)$$

where T is the set of signatures tuples presented in both signature, RSSI(i)r is the RSSI value in the signature appearing in signature $r_i$ and RSSI(i)s is the RSSI value in the signature appearing in signature s.

Given the set of signature distances, the location of a mobile node can be calculated in several ways applying one of the following fingerprint mechanisms.

### A. Simple Algorithm

In this algorithm the location point will be one of the stored in the fingerprints database, where associated signature Manhattan distance is the lower from the one obtained in the online phase.

$$\min(M(r,s)) \quad (2)$$

It must be noticed that through single fingerprint algorithm, it's only pretended to locate in rooms (not in hallways or courtyard) and only with room-level accuracy.

### B. Centroid Algorithm

Centroid algorithm is similar to the previous one but considers the centroid of the set of signatures within some ratio of the nearest reference signature. Given a signature

s, a set of reference signatures R, and the nearest signature $r^* = \arg\min r \in RM(r,s)$ we select all reference signatures r ∈ R that satisfy

$$\frac{M(r,s)}{M(r^*,s)} \qquad (3)$$

for some constant c, empirically-determined. The geographic centroid of the locations of this subset of reference signatures is then taken as the mobile node's position. Small values of c work well, generally 1.1 or 1.2.

When a more completely localization is necessary, not only locating in places where a point is previously calculated and saved into the fingerprint database, centroid algorithm is a better solution. Centroid algorithm is able to locate in hallways or courtyard, even in the entrance of each room.

## III. Hardware Used

The system is composed of a WSN to acquire RSSI and a PC system where the location estimation is made via a positioning desktop application. This system is based on low-power, embedded wireless devices, MeshNetics´s sensor "motes" called MeshBean2.

The advantage of this platform over other motes is that it's equipped with leds, buttons, sensors, and extra sensors could be easily connected that might be used for different purpose applications for indoor position system, ambient living or smart buildings, if the application requires them, so for prototyping, those boards works quite well. They also have a USART accessible by a USB connector, so a PC can be connected via USB port, emulating a COM port, for both, programming and receiving information, in this case, beacons and sensor values.

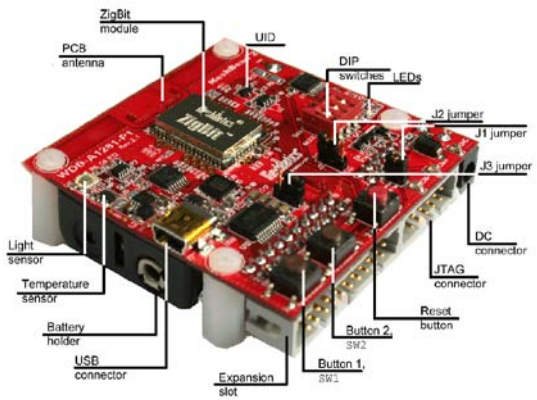



Figure 2. Meshbean development board.

Other advantage of this mote is that the supplier has developed the ZigBee RFC4 stack architecture [14] in a software pack called BitCloud Stack and also an IEEE 802.15.4 MAC level implementation, in a software pack called OpenMAC.

A MeshNetics´s mote is shown in Fig. 2, in this case, it has got an integrated PCB antenna, but we have used others that haven't, it affects only the range of coverage. This mote has a MCU wireless, called ZigBit, a compact 802.15.4/ZigBee module featuring record-breaking range performance and exceptional ease of integration. It integrates both the ATmega1281 microcontroller and AT86RF212 transceiver of ATMEL (www.atmel.com) so the AVR tools are necessary for programming purposes.

## IV. BitCloud Implementation

BitCloud is a full-featured ZigBee PRO stack that supports a reliable and scalable wireless applications running on Atmel wireless platforms. In ZigBee there are three kinds of devices, each one having its own purpose:

1. Coordinator (C): A full function device (FFD) that it is in charge of creating the PAN (Personal Area Network) and typically is the point of the WSN (Wireless Sensor Network) to acquire all sensors information from all the other motes to be shown in a computer. The used icon which represents this device is a filled circle, Fig. 1 shown one.

2. Router (R): A FFD that it is in charge of routing when the range of coverage requires this capability, so it is possible to have dynamic topologies. The used icon which represents this device is a small filled circle inside a circle, Fig. 1 shown six ones.

3. End device (ED): A reduced function device (RFD) that is always slept (to reduce consumption) and only wakes up to do a specific task, for instance, to send sensor information to the WSN, typically directed toward C. The used icon is a not filled circle; this is, like the R icon in Fig. 1, but no filled circle inside.

So a ZigBee WSN is composed of one C, many EDs and many Rs. Each kind of devices can receive what the other transmit if they are in the same range of coverage, because the transmission media is shared by all one, but not all the received information is processed (the explanation of why this is that way is out of the scope of this paper).

As explained in the previous section, to determinate the position, we require two kinds of beacons, beacon1 and beacon2. Beacon1 is used to inform other devices that a mobile mote is present and beacon2 is used to inform C the RSSI value that fixed mote has received from a mobile one for getting location estimation.

To send both beacons in BitCloud Stack, the information saved in a table, called neighbour table, at the network layer of a certain mote is used. This table has registered all the FFD, this is, C or R that are in the same range of coverage of the mote and, for each one, it registers the RSSI value of the received signal from the mentioned mote. Periodically, a FFD device sends a Zigbee Network layer message to inform other that is in the PAN, so that message is used by neighbor motes to measure the RSSI value of the received signal and to save it in their own neighbor table. So beacon1 is sent automatically by the protocol stack. As only FFD sends this kind of message the mobile motes have to be R, as shown in Fig. 1.

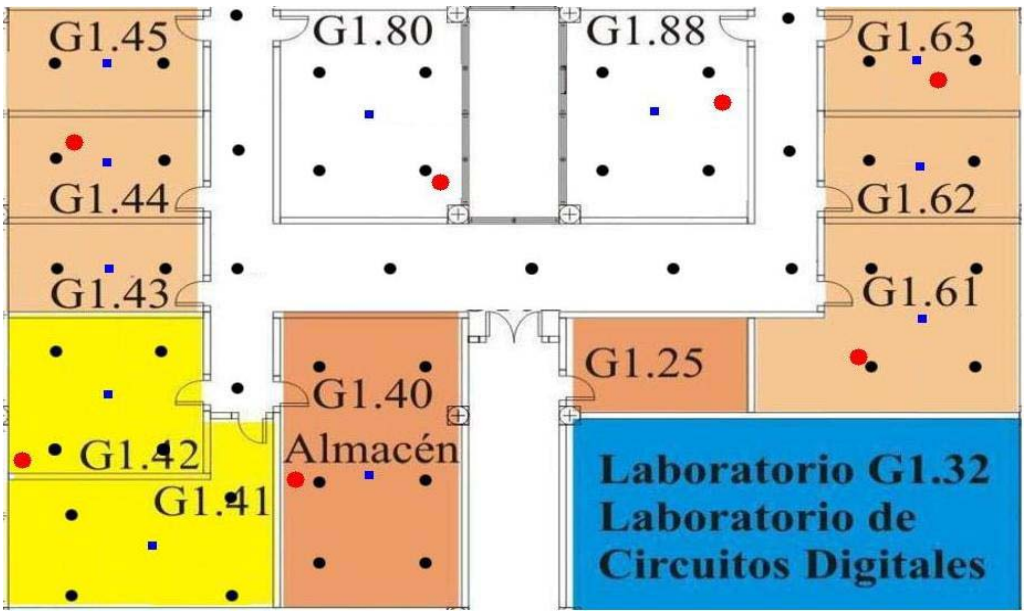



Figure 3. Fixed Motes Location and signature points saved in the database

To send periodically beacon2 messages, each fixed motes search in its neighbour table to find out if the mobile mote is in its range of coverage, if so, the beacon2 is sent to C with the information required as explained in section 3. The beacon 2 message is a Zigbee Application message provided by the APS (APplication Service). As neighbour table is only in FFD, fixed motes have also to be R.

We deployed the BitCloud solution over half floor of our Department Area, measuring roughly 225 m2. To cover all this area we required 7 fixed motes strategically placed, shown as red points in Fig. 3. An off-line phase was required for each algorithm to fill in the signature database, once it was full, the system was ready to be tested.

Fig. 4 shows the PC interface to presents the mobile mote position, four in this case. It also shows mobile mote sensors information.

### *A. BitCloud Test*

We have tested the system to check if it can determine the location of the mobile mote using both algorithms. It must be noticed that we pretend to locate in rooms, hallways or courtyard, but with room-level accuracy. We considered it doesn´t matter exactly where it is inside de room. This has been this way, because the kind of applications for whom our indoor position solution is going to be used, doesn´t require more precision.

### *B. Simple Algorithm*

For single fingerprint algorithm, 11 points were collected and saved in the fingerprint database, one point of each room, shown as blue points in Fig. 3. Through this algorithm is pretended to locate only in rooms, not halls or courtyards. In this case, we placed the mobile mote on each room (11 times) and nearly all of the position measurements performed good results. 82% of success was obtained. Only two rooms were unallocated. Fig. 5 shows the results. Empirical tests have demonstrated that 11 points were enough to conclude, because tests made on the same room, had behaved the same way.

### *C. Centroid Algorithm*

However, centroid fingerprint algorithm is also intended to locate in corridors too, even in the entrance of each room. In this case, 46 points were used to fill the fingerprint database, distributed by rooms and hallways. See Fig. 3 black points

We decided, also based on empirical measurements, to test 30 points, to check how the positioning centroid algorithm worked. It was determined that the algorithm presented the right position in 23 points but in 7 points, it made a bad position determination. Fig. 5 shows those points. Therefore our precision was about 77%.

Although results were as expected, as shown in Motetrack this solution has two drawbacks:

The mobile node has to be FFD because it has to use the neighbour table to get the RSSI so the power consumption is very high and it causes consumption problems since mobile node is battery powered.

The periodicity of beacon1 messages can´t be controlled, and we can't find out the age of the RSSI values as it is a Zigbee Network parameter not accessible by BitCloud because it is an Application Level Stack.

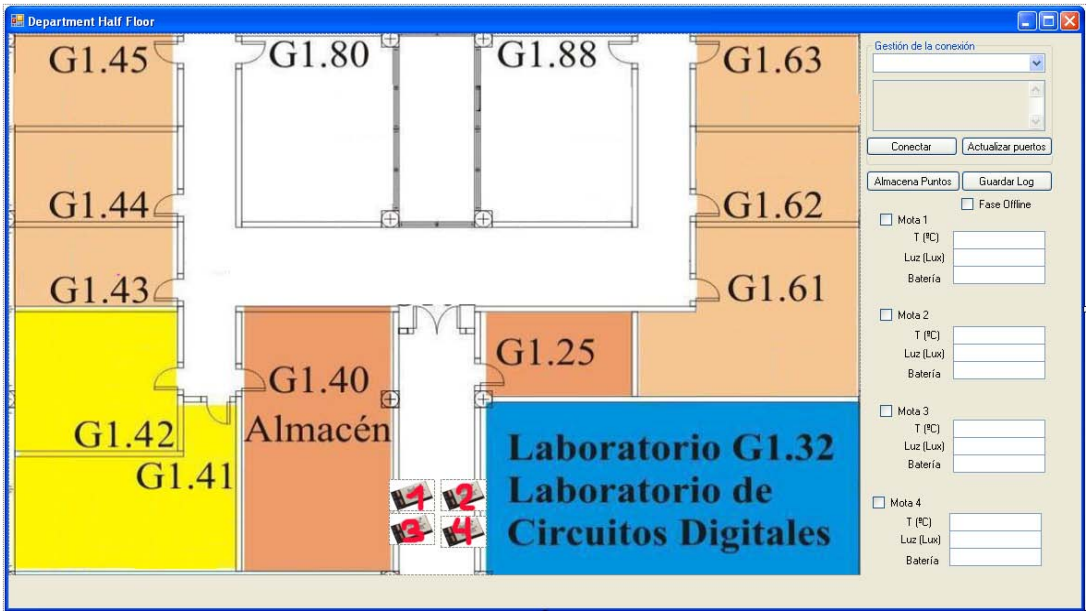



Figure 4. Position System Interface

Further work is done to optimize the system WSN stack, in order to fix the found drawbacks and to try to get more accuracy.

## V. OPENMAC IMPLEMENTACION

OpenMAC is an open source implementation of IEEE802.15.4 Media Access Control (MAC) layer. It's an embedded software that provides basic networking functionality, only star and peer-to-peer topologies. Because of it, only Coordinator (C) and End Devices (ED) examples are implemented. To create all the PAN devices, the MAC services implemented in OpenMAC for doing so has been used.

It has some advantage over using BitCloud Stack:

1. Enables users, who do not require full functionality of BitCloud Stack, to develop custom WSN applications.

2. Enabled advanced users to modify OpenMAC internals to suit specific application needs.

3. Jump start application development on top of MAC with thoroughly documented sample applications.

4. Provide a convenient C API to developers not familiar with TinyOS or nesC programming language (technologies at the core of OpenMAC).

5. Provide a reference design to be ported to analogous hardware platforms.

To deploy the same indoor solution as in BitCloud Stack, a PAN like the one shown in Fig. 1 has been created, where all types of ZigBee devices, C, R and ED

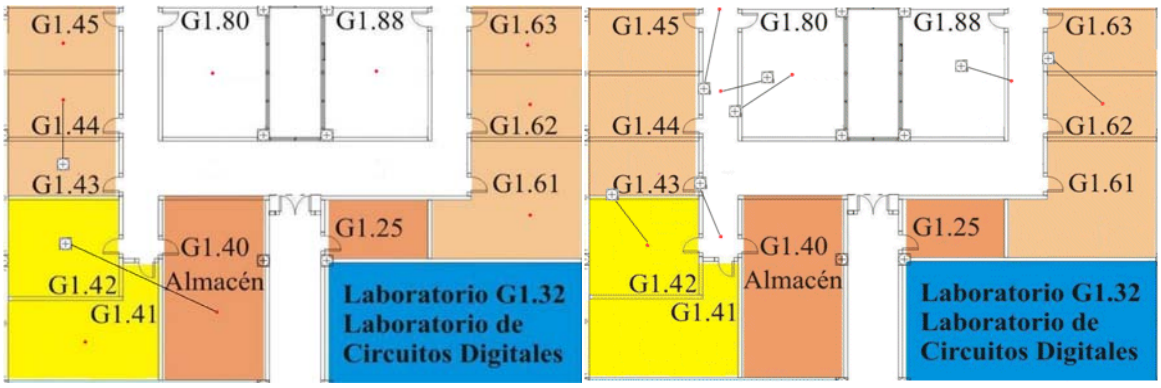



Figure 5. Fingerprint and Centroid Tests Points and Errors using BitCloud Stack

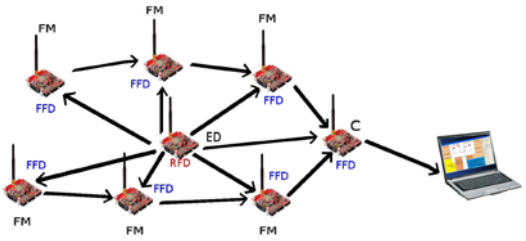



Figure 6. OpenMAC System Overview

are created and all of them implement the corresponding functionally. So, C creates the PAN and R and ED connect to the PAN via C or other R that is already in the PAN. See Fig. 6.

All data flow is towards C, so Rs are in charge of forwarding packets when other Rs or EDs requires it, this happens when the first ones are closer to C than the second ones.

The ED is in charge to send beacon1 messages. This message is broadcasted, so all its neighbours are able to calculate the RSSI value of the received message and send this information directly or indirectly to C. When an ED message is received, R sends beacon2 message to his father. This message is unicast, so a R that receives one, has to forward it if the source R is one of its child.

The OpenMAC PAN works correctly and is set to deploy it over the half flour of our Department too.

The PC system is nearly the same used with BitCloud stack, only some changes were necessary to collect data from beacon2 messages. Obviously OpenMAC beacon2 messages structure is different from BitCloud one.

### A. *OpenMAC tests*

*Simple Algorithm*

For single fingerprint algorithm, 11 points were collected too, using the new protocol stack, one point of each room, as equal to BitCloud test, and the same results of BitCloud were obtained, two errors over 11 measures, one per room, 82% of success. Fig. 7 shows the test

*Centroid Algorithm*

Again, the same 46 points we used in BitCloud tests, were saved into the OpenMAC Centroid fingerprint database.

The same 30 points tested in BitCloud has been used to check how the positioning centroid algorithm works using the OpenMAC stack. It is determined that the algorithm gives the right position in 26 points, and in 4 points it presents a wrong position determination. Making numbers again our precision is about 87%. Better results than the first BitCloud tests. The test can be seen in Fig. 7

However, the two main drawbacks of BitCloud stack are solved this way. In OpenMAC, mobile motes are RFD (Fig. 6), so they are slept all the time and are only woken up when they have to send broadcast beacon1 message, the periodicity of the messages are controlled and updated every time, so the age of the RSSI value is now controlled. Using OpenMAC, it can be considered that the accuracy of the obtained off-line fingerprint database is better than the BitCloud. RSSI values are new on each time and the average is made using a realistic sample. In BitCloud, as said before, the age of the RSSI measure is unknown and if it isn't updated, the same value is always given. The fingerprint database made using this sample must be worse than the one made by OpenMAC.

## VI. Conclusion

We have presented an indoor position system based on WSN using the RSSI. BitCloud stack has been used to implement the network functionality.

Some drawbacks have been found by using BitCloud stack and were considered they could affect positioning accuracy so another implementation has been done using another stack, OpenMAC. A PAN infrastructure for indoor position has been developed and used in a lab environment.

By checking the OpenMAC implementation improvements and compare it to the previous one, some off-line phase has been required to fill in the different signatures database and some tests have been done using

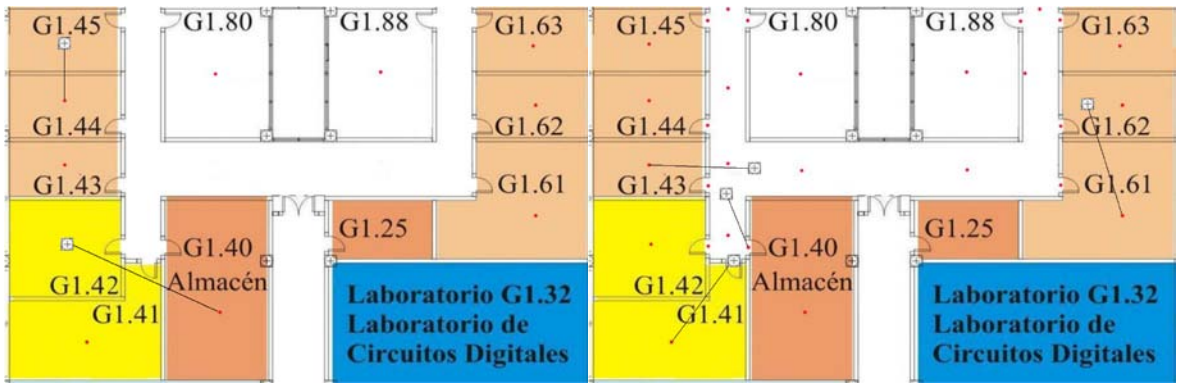



Figure 7. OpenMAC Fingerprint an Centroid Tests Points and Errors

different positioning algorithms, Simple (if only room positioning is required) and Centroid (if more complex positioning estimation is needed), both getting good results. Accuracy is increased 10% using Centroid algorithm. Using Simple algorithm we got the same results.

## ACKNOWLEDGMENT

This work has been carried out within the framework of two research programs: (P08-TIC-3631) – Multimodal Wireless interface (IMI) funded by the Regional Government of Andalusia and Efficient and Health Intelligent Technologies Oriented to Health and comfort in Interior Environments (TECNO-CAI) approved project at the fifth call of CENIT program by the Innovation Science Ministry of Spain (CDTI and Ingenio 2010 Program).

## REFERENCES

[1] Hristova, A.; Bernardos, A.M.; Casar, J.R.2008.Context-aware services for ambient assisted living: A case-study. First International Symposium on Applied Sciences on Biomedical and Communication Technologies, 2008. ISABEL '08.

[2] Figueiredo, C.P.; Gama, O.S.; Pereira, C.M.; Mendes, P.M.; Silva, S.; Domingues, L.; Hoffmann, K.-P. 2010 Autonomy Suitability of Wireless Modules for Ambient Assisted Living Applications: WiFi, ZigBee, and Proprietary Devices.*Sensor Technologies and Applications (SENSORCOMM*).

[3] Hong Sun; De Florio, V.; Ning Gui; Blondia, C., Towards. 2008. Building Virtual Community for Ambient Assisted Living. 16th Euromicro Conference on Parallel, Distributed and Network-Based Processing.

[4] Sun, H.; De Florio, V.; Gui, N.; Blondia, C., 2009. PRomises and Challenges of Ambient Assisted Living Systems. Information Technology: New Generations, 2009. ITNG '09.Sixth International Conference on. Page(s): 1201 - 1207.

[5] Martin, H.; Bernardos, A.M.; Bergesio, L.; Tarrio, P.,2009. Analysis of key aspects to manage wireless sensor networks in ambient assisted living environments. Applied Sciences in Biomedical and Communication Technologies, 2009. ISABEL 2009. 2nd International Symposium. Page(s): 1 - 8.

[6] Dietrich, D.; Bruckner, D.; Zucker, G.; Palensky, P. 2010.Communication and Computation in Buildings: A Short Introduction and Overview. IEEE Transaction on Industrial ElectronicVolume: 57 , Issue:11.

[7] Chen, Po-Wei; Ou, Kuang-Shun; Chen, Kuo-Shen. 2010.IR indoor localization and wireless transmission for motion control in smart building applications based on Wiimote technology.SICE Annual Conference 2010.

[8] Han Chen; Chou, P.; Duri, S.; Hui Lei; Reason, J. 2009.The Design and Implementation of a Smart Building Control . Page(s): 255 - 262.

[9] Snoonian, D , Smart buildings.2003. Spectrum, IEEE Volume: 40 , Issue: 8 Digital Object Identifier: 10.1109/MSPEC. Page(s): 18 - 23.

[10] Bahl P., Padmanabhan VN. 2000. RADAR: an in-building RF-based user location and tracking system. In: INFOCOM, pp 775–784.

[11] Konrad L. and Matt W. 2005. MoteTrack: A Robust, Decentralized Approach to RF-Based Location Tracking. In Proceedings of the International Workshop on Location and Context-Awareness (LoCA 2005) at Pervasive May 2005.

[12] Konrad L., Matt W. 2006. MoteTrack: A Robust, Decentralized Approach to RF-Based Location Tracking. To Appear in Springer Personal and Ubiquitous Computing, Special Issue on Location and Context-Awareness. . ISSN: 1617-4909 (Print) 1617-4917.

[13] Medina A. V., Gómez I., Romera M., Gómez J.A. and Dorronzoro E. 2011.Indoor Position System based on BitCloud Stack for Ambient Living and Smart Buildings. In 3rd International ICST Conference on IT Revolutions. Córdoba, Spain.

[14] ZigBee, 2009 RF4CE Specification. Version 1.00. ZigBee Document 094945r00ZB, March 17th, 2009.

[15] Medina A.V., Gómez J.A. Rivera O. Dorronzoro E. and Merino M. 2011. Fingerprint Indoor Position System based on OpenMAC. In International Conference on Wireless Information Networks and Systems, Sevilla Spain

**Jose A. Gómez** was born in Seville, Spain on 14 February 1983. He received the computer engineer degree from the Seville University in 2009 and network and computing engineering master in Computer Engineering and Network from the University of Seville in 2011

He has been worked as a Teacher in higher grade formative cycles. Currently he is a PhD student in Department of Electronic Technology where also works as a teacher. His research interest includes WSN and Indoor Localization Systems. He has published some technological papers.

**A.Verónica Medina** (IEEE Member'08) received her degree in computer engineering in 1993 and her doctorate (Ph. D.) in computer engineering in 1999 from the University of Seville (Spain). Since 1990, she has been working on the formal description techniques and their application to the development of communication protocols. In 1992, she was awarded a prize by the Andalusia government for her study "Application of Computer Sciences to the Control of Electric Power Networks." Presently, she is a Lecturer of electronic engineering at the University of Seville and a researcher in TAIS research group (Technologies for Care, Inclusion and Health, http://matrix.dte.us.es/grupotais/).

**Octavio Rivera** received the Ph.D. degree from the University of Seville, Seville, Spain, 2010.

His research interests include Assistive technology, Augmentative and Alternative Communication, HCI, Serious games and rehabilitation engineering.

He is currently a Lecturer and Researcher at University of Seville, Seville, Spain.

He has been a member of a working group in the communication field since 2001. He worked in some communication projects involving power-line communications systems, digital-subscriber-line technologies, indoor location systems and telecontrol protocols.

**Enrique Dorronzoro** received his master degree of Computer Science Engineer at University of Sevilla, Spain, in 2007.

He has worked as Instructor at the Cisco Networking academy program. He was on 2010 in a 6 months research visit at Norut in Tromsø, Norway. Nowadays he works as researcher at Electronic & Technology Department of the same University where he has worked in several research paper. His interest include wireless networks sensors systems applied to telemedicine.

**Sergio Martín** received his degree in Computer Engineering in 1994. Since 1990, he has been working on the Formal Description Techniques, their application to the development and also in the application of Expert System to the Industry. Now his research is focusing on adding expert rules to the management objects in TMN (Telecommunication Management Network). Presently he is an Associate Professor of Computer Engineering at the University of Seville.

# Localization method for low-power wireless sensor networks

Diego Fco. Larios, Julio Barbancho, Fco. Javier Molina and Carlos León *Senior Member, IEEE*.
Department of Electronic Technology, "Escuela Politécnica Superior"
University of Seville, Seville, Spain,
Phone: +34 954 55 28 38
Fax: +34 954 55 28 33
Email: dflarios@dte.us.es, jbarbancho@us.es, fjmolina@us.es, cleon@us.es

***Abstract*— Context awareness is an important issue in ambient intelligence to anticipate the desire of the user and, in consequence, to adapt the system. In context awareness, localization is very important to enable a responsive environment for the users.**

**Focusing on this issue, this paper presents a localization system based on the use of Wireless Sensor Networks devices. In contrast to a traditional RFID, these devices offer the possibility of a collaborative sensing and processing of environmental information.**

**The proposed system is a range-free localization algorithm that uses fuzzy inference to process the RSSI measurement and to estimate the position of mobile devices. The main goal of the algorithm is to reduce the power consumption and the cost of the devices, especially for the mobiles ones, maintaining the accuracy of the inferred position.**

***Index Terms*— Fuzzy system, WSN, localization, RSSI, centroid, AmI**

## I. Introduction

Ambient Intelligence (AmI) refers to electronic environments that improve the user experience of the environment by responding to them intelligently. It implies the use of advanced networking computing technology that is aware of human presence, personalities and needs. In ambient intelligence, devices may work to help people in their everyday activities. These devices would both sense and react based on the environmental context.

Intelligent homes are a typical AmI application area, because they increase security and comfort for the users [1]. In the notion of ambient intelligence, the main keys [2] are the followings (figure 1):

- Embedded. Many networked devices are integrated into the environment for sensing and controlling it. Ideally, these devices are deployed throughout the environment, but their presence must be undetected for the users. The devices share information to increase user's comfort. Communication is essential to obtain an adequate response in computing, where control intelligence is spread all over the network. WSN is an example of technology that complies with this key.
- Context awareness. The system would recognize the situational context of the users. This implies that the devices can recognize the user [3], his position and the positions of those devices, the users can interact with.
- Personalized. The system must adapt to the needs of the users. The system must be intelligent enough to adapt to the desires of the user in function of his natural reactions.
- Adaptive. The system must learn the user's needs. For example, learning the temperature or the illumination that users feel like having in each room, adjusting them automatically.
- Anticipatory. The system must anticipate user's needs without conscious mediation. Anticipatory is related to the user interfaces in the ambient intelligence and it searches for an easy interaction between machines and users [4]. Ideally, these interfaces would not be perceived by the users [5].

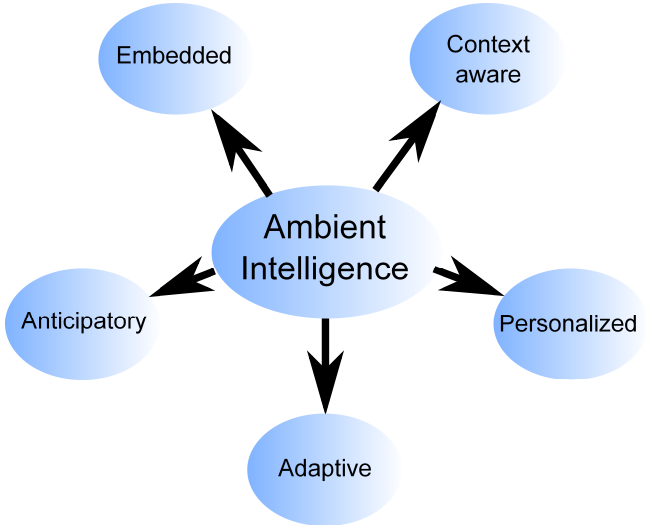



Figure 1. Ambient Intelligence keys.

This paper is based on "Locating sensor with fuzzy logic algorithms," by D. F. Larios, J. Barbancho, F. J. Molina and C. León, which appeared in the Proceedings of the 24th IEEE Symposium on Proceedings of the 8th International Joint Conference on e-Business and Telecomunications (ICETE), Seville, Spain, July 2011.

This research has been supported by the "Consejería de Innovación, Ciencia y Empresa", "Junta de Andalucía", Spain, through the excellence project ARTICA (reference number: P07-TIC-02476) and by the "Cátedra de Telefónica, Inteligencia en la Red", Seville, Spain, through the project ICARO.


doi: 10.4304/jnw.8.1.45-58

The ambient intelligence paradigm is often built on pervasive computing, ubiquitous computing and computational intelligence [6]. AmI paradigm started to be developed with the challenge launched by the European Commission in 2001, aiming that the researchers investigate in this area [7].

In context awareness, localization system is essential to enable a more reactive useraware responsive environments [8]. Tracking users offers basic information useful to learn about their habits, or to predict following displacements [9]. For example, before a user enters a room, the air conditioning would be adjusted to his comfort automatically.

This paper faces the problem of locating users based in the AmI paradigm, reducing the power consumption and the cost of the devices. Energy efficiency is especially relevant for mobile devices. It improves battery autonomy and increases the user comfort. The proposed method is based on a Wireless Sensor Network (WSN) that can monitor and control the environment intelligently in function of the user desires.

The rest of this paper is organized as follows: section II sums up the state of the art about WSN applied to ambient intelligence. Section III describes the localization algorithm. Section IV evaluates the advantages of the proposed localization algorithm over the power consumption. The outcome of the localization algorithm performance is developed by simulations, and the results are shown in section V. Finally, in Section VI we present concluding remarks.

## II. Localization techniques

A wireless sensor network (WSN) consists of lots of small devices deployed in a physical environment for its study. Each node has special capabilities, such as wireless communications with its neighbours, sensing, data storage and processing.

WSN has been widely used in many areas [10], such as environmental monitoring [11] and control [12], healthcare and medical research [13], national defense and military affairs [14] [15], etc.

The use of WSN in ambient intelligence offers many advantages [16]. It uses has been proposed by many authors [17] and rejected by others because of the cost and battery consumption that these devices present [18]. But nowadays, the cost has been reduced significantly, and the energy consumption can be reduced with suitable software and hardware design. In this way, it is possible to get years of battery autonomy for the mobile nodes.

Using the microcontroller and external sensors and actuators, the current nodes used on WSNs can control the comfort and safety of the users. For example, a device can make emergency calls in case of an accident [19].

Traditionally, in ambient intelligence, the localization solution is based on passive or active RFID technology, but RFID has a very limited capacity to monitor the environment compared to WSNs. Moreover, the prices of RFID tags and some WSNs nodes are currently similar (20-30 $ [20]). Moreover, the in WSN tags and readers share the same economic hardware, but in RFID a reader is much more expensive than a tag (around 1500 $) and more expensive than a WSN device.

In this journal, we propose a system to control the localization using WSN devices only. This reduces the cost and the complexity of the system and increases its capabilities and functionalities.

*1) Localization Techniques for WSN:* For this applications, we are going to consider the following types of nodes:

- **Anchor nodes:** located on a fixed position in the house. Theses devices are used to route the information to the Base Station and for the localization algorithm. These devices act as the readers of the RFID technology.
- **Tags:** are the small devices deployed with the object or users to locate. Their positions are initially unknown and the algorithm would find them. These devices can obtain environmental information through sensors, such as temperature or accelerometers. These devices are also called non-anchor nodes.
- **Base Station:** This is a special anchor node that acts routing the WSN information from the network to a PC. This PC provides the information acquired by the network to the rest of the AmI devices.

Localization algorithms presented in the literature can be classified into two categories:

- **Range-based:** These techniques estimate, point-to-point, the distance between all the nodes using sensors such as ultrasound [21]. With this information, using techniques such as triangulation, the absolute position of the non-anchor nodes can be estimated. Generally, these techniques require additional hardware. The most common ones are Received Signal Strength Indication (RSSI) [22], Time Of Arrival (TOA) [23] and angle of arrival (AOA) [24].
- **Range-free:** In these techniques, the position of non-anchor nodes is obtained according to implicit information provided by anchor nodes, usually based on messages exchanged, commonly called beacons. This information is usually made up of different aspects, such as radio coverage membership or number of hops between devices. The most common ones are Centroid (CL) [25] and DV-Hop [26].

In general, the range-based ones offer good accuracy, but additional hardware is often needed. Therefore, the weight, the cost and the power consumption of node devices increase, and make these sort of techniques unsuitable. RSSI range-based techniques are an exception to this because most of the current transceivers provide this measure by default. However, RSSI techniques are very sensitive to noise and interferences. Figure 2 shows experiments realized by the authors to evaluate the relationship between RSSI and the distance in different situations: free-space without obstacles and long urban area with obstacles.

The results do not match with any known models, such as the Friis equation, but in very ideal conditions. In fact,

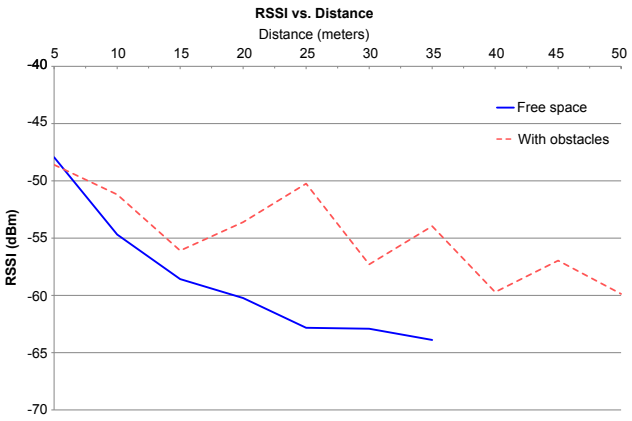



Figure 2. RSSI vs. distance.

in common scenarios, with obstacles and/or inside the buildings, there is no direct relationship between distance and RSSI. In fact different distances produce the same RSSI value.

Instead of using a mathematical model, the proposed solution uses a fuzzy-logic-based system to derive the distance from RSSI level. This is more robust in noisy and complex scenarios. This idea is not new. Computational Intelligence has been proposed for localization in several papers, such as [27] that uses probabilistic neuronal networks, [28] that applies a fuzzy system and [29] that uses fuzzy neurons. In general, all these algorithms track down current positions based on estimate position changes. But none of them consider the problem of power consumption.

## III. Description of the Proposal Solution

To locate the important elements (devices, users, etc) over a WSN, we proposed LIS (Localization Based on Intelligent Systems). LIS is a technique that determines the localization using a fuzzy system. LIS use RSSI, but not associating it to a distance estimation. Due to it, it is a range-free technique, such as weighted centroid.

The inputs of the fuzzy system are the RSSI measurements related to the non-anchor node. These RSSI values are measured when anchor nodes receive a message sent from the tag.

LIS was designed to be developed in a building. In fact, a prototype is currently deploying in the "*Red building*" of the University of Seville. This building is in the "Campus de Reina Mercedes" and has 12 classrooms and 38 offices spread over three floors and a basement. This prototype will consist of a mesh of fixed anchor nodes spread throughout the building (figure 3), and any small non-anchor nodes situated together to the elements to locate. The aim of this network is to control the position of any resources, such as laptops, monitors, printers, projectors, etc, to prevent theft.

With LIS, all the information is stored and timestamped in a PC attached to the Base Station. This PC stores two different classes of information gathered from the network:

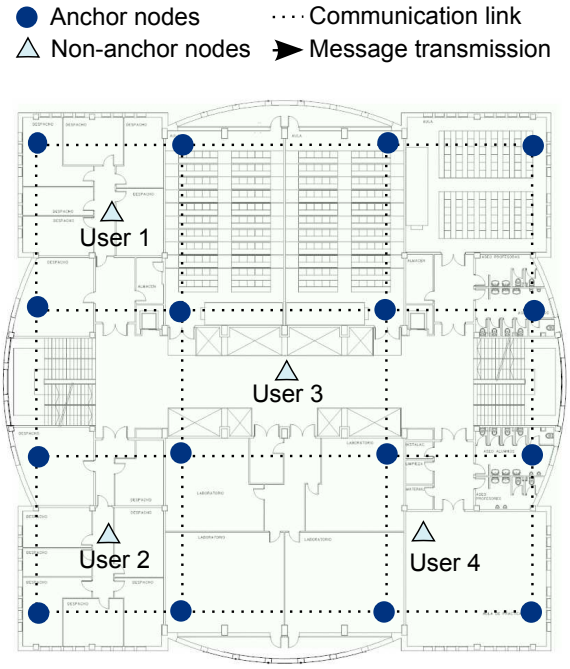



Figure 3. Infrastructure of LIS

- **Target location:** This information describes the position.
- **Sensors information:** This information is the environmental measurement, such as the temperature, humidity, etc.

Despite that range-free and range-based techniques have been extensively studied, nowadays there are some aspects that continue to be a challenge:

- The use of additional hardware or lots of beacons increases power consumption.
- Centralized processing (i.e. on Base Stations) requires a large amount of messages. Conversely, processing in tags nodes reduces the battery of these devices significantly.
- Scalability. Many of range-based algorithms are hard to extend to big sensor networks.

LIS has been especially designed to face all of the above problems. As a result, the proposed algorithm is scalable and the power consumption and network autonomy are optimized. LIS combines: (I) a fuzzy system to estimate (actually to qualify) the distance between transmitter and receiver from RSSI measurements, (II) a ubiquitous algorithm executed in anchor nodes that receive the beacon of the non-anchor nodes, to determine relative positions to them, and (III) a cooperative algorithm to derive the most likely location running at the Base Station.

LIS algorithm consists of four stages:

S1: Anchor nodes wait for non-anchor nodes (tags) beacons.
S2: The tag node broadcasts a beacon.
S3: Receiver anchor nodes measure RSSI, and execute the Ubiquitous Processing (UP) for relative and

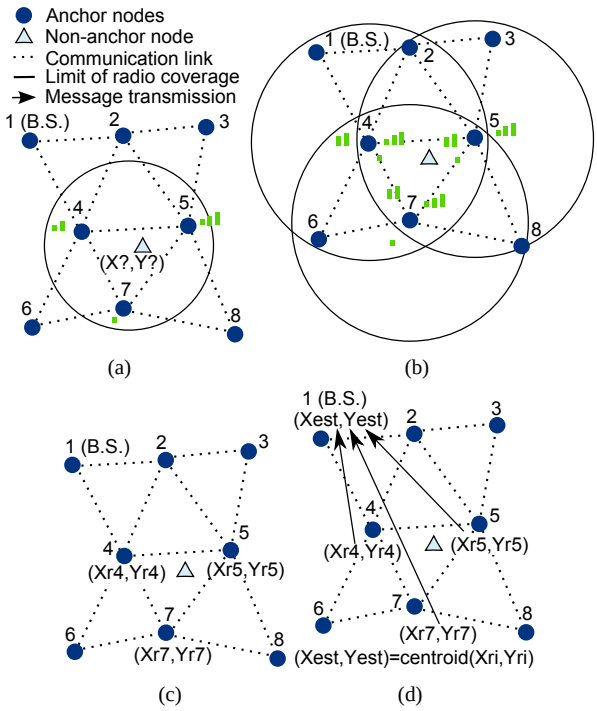



Figure 4. Steps of LIS algorithm: (a) S1; (b) S2; (c) S3; (d) S4

partial positioning.

S4: Anchor nodes send partial solutions to the Base Station, where the location is finally determined with the Cooperative Processing (CP).

Figure 4 illustrates these stages. When a non-anchor node broadcasts a beacon or any other sort of message, the localization process starts. Just anchor receiver nodes participate in the process. The rest of the nodes can switch off the radio transceiver or hold in a low power state.

For localization, the tags periodically send broadcast messages to the anchor nodes. The frequency of the messages would be adjusted in function of the characteristics of the system.

### *A. Ubiquitous Processing (UP)*

LIS uses the measured RSSI of a node and its neighbours to determine the area where the non-anchor node could be located. This algorithm is based on a fuzzy system distributed on every anchor node of the network.

According to the algorithm stages, once an anchor node receives a beacon, it estimates the position of the non-anchor nodes. The localization algorithm has been designed to distribute the computation consumption over the network. The area where the non-anchor node could be located with a certain probability is called Representative Area (RA). A *Sector* is the minimum area formed by three anchor-node neighbours. A Representative Area (RA) represent the estimation, in terms of sectors, where a node estimate a tag would be. as described below, RA can be made up of one or more sectors with an UP processing higher than a certain threshold.

Anchor nodes must execute the distributed Fuzzyfication Algorithm (FA) measurement for every surrounding sector. Figure 5 shows an example with five sectors in which, the fuzzy processing are executed five times.

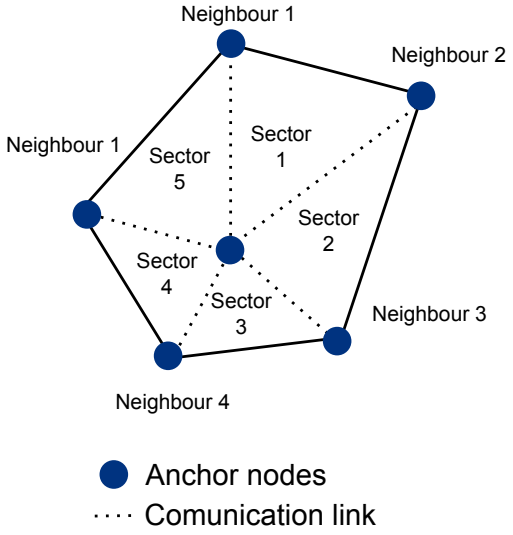



Figure 5. Example of node with 5 neighbours.

Every node that receives a beacon, it measures and broadcasts the RSSI level to its neighbours (figure 4.b). In this way, the closest anchor nodes elaborate a table with the RSSI measured by them and their neighbours.

The RSSI table is processed by FA (figure 6) to evaluate the representative area no matter the number of sectors. This area can be formed by the union of one of more sectors (figure 4.c). A sector is considered part of the representative area if its membership degree is higher than the threshold. This value is adjusted experimentally.

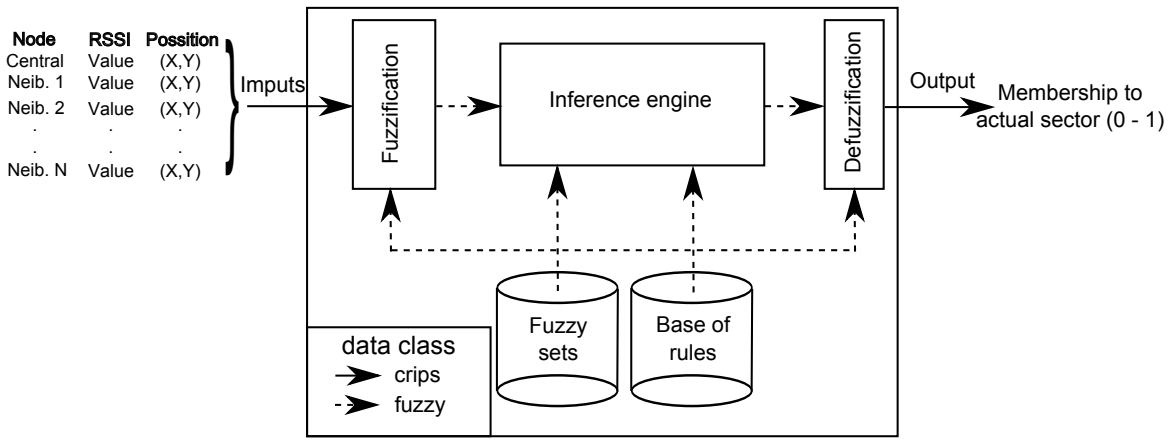



Figure 6. Inference fuzzy system.

Our simulations show that a threshold of 0.1 manages a good trade-off between noise immunity and localization performance. The results of Representative Areas (RA) are sent from the anchor nodes to the Base Station to compute the final solution (figure 4.d).

An RA is empty if it does not contain significant sectors, this is if the membership degree for all of them is lower than the threshold. In this case, to save energy, the result is discarded and the algorithm will finish until the next beacon arrives (figure 8). This is especially important in huge networks, where the energy needed for multi-hop transmissions is high. This issue is discussed in detail in section IV.

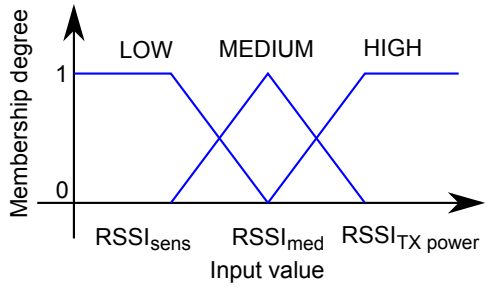



Figure 7. Sets of the fuzzy imputs.

*1) The Fuzzy System Inputs:* RSSI tables represent the signal level received in either local and neighbour nodes. Three fuzzy sets qualify the RSSI as High, Medium and Low for each input, as appear in figure 7. This figure is a simplification. The area of the fuzzy imput sets varies in function of the estimated RSSI. Due to it, it may be not symmetrical.

The LOW RSSI fuzzy set is represented by a trapezoid. Maximum membership degree (1) is assigned if power falls down the sensibility threshold of the emitter node ($RSSI_{sens}$). As the power increases, membership degree decreases linearly until zero. The following equation defines this fuzzy set:

$$\mu(x) = \max\left(\min\left(1, \frac{RSSI_{med} - x}{RSSI_{med} - RSSI_{sens}}\right), 0\right) \quad (1)$$

The MEDIUM RSSI fuzzy set is represented by a triangle where maximum membership degree corresponds to the medium RSSI value ($RSSI_{med}$). Zero membership is reached for power RSSI values lower than the sensibility threshold or close to the maximum transmission ($RSSI_{TXpower}$). In this paper, medium RSSI value must be computed for every sector using the Friis model equation (equation 2) and assuming the emitter tag is located at the centre. This computation only needs to be executed once because anchor nodes are located at fixed positions.

More complex models, such as Two-Ray ground model or ITU indoor model are not considered in this paper for minimizing the complexity of the algorithm, But in can be used to increase the accuracy of the system. The error with led to the use a simplified model is compensated with the redundancy of information and the use of FA and UP Algorithms.

$$\frac{P_{RX}}{P_{TX}} = G_{TX} \cdot G_{RX} \cdot \left(\frac{\lambda}{4\pi R}\right)^2 \quad (2)$$

Where $G_{TX}$ and $G_{RX}$ are the gain of TX and RX antennas, $R$ is the distance between transmitter and receiver and $\lambda$ the wavelength.

Next expression defines the fuzzy set for MEDIUM RSSI:

$$\mu(x) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (3)$$

Where $a = RSSI_{sens}$, $b = RSSI_{med}$ and $c = RSSI_{TXpower}$.

Fuzzy set for HIGH RSSI values is a trapezoid with a lineal increasing from 0 to 1 for RSSI power values ranging between $RSSI_{med}$ and $RSSI_{TXpower}$. This fuzzy set is defined by the next expression:

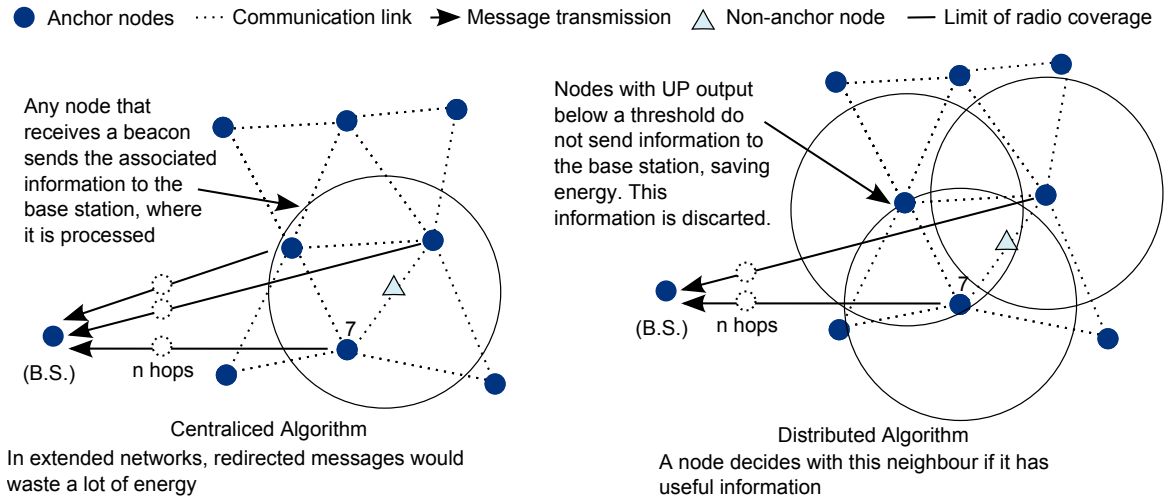



Figure 8. Distributed algorithm would save Power Energy on extended networks.

$$\mu(x) = \max\left(\min\left(\frac{x - RSSI_{med}}{RSSI_{TXpower} - RSSI_{med}}, 1\right), 0\right) \quad (4)$$

*2) The Fuzzy System Outputs:* The Fuzzy System offers an output for each and every sector. The output associated to a sector is a [0, 1] ranged value that represents the confidence degree that the tag is actually located in that sector.

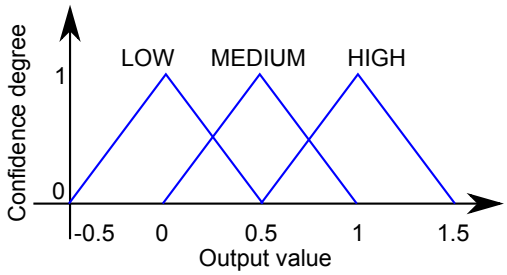



Figure 9. Sets of the fuzzy output.

As figure 9 shows, the LOW output fuzzy set is a triangle with the central point at zero and the corners at -0.5 and 0.5. The next expression defines this fuzzy set:

$$\mu(x) = \max\left(\min\left(\frac{x + 0.5}{0.5}, \frac{0.5 - x}{0.5}\right), 0\right) \quad (5)$$

MEDIUM output is represented by a triangle with the central point at 0.5 and corners at 0 and 1. Mathematically it can be expressed by the following equation:

$$\mu(x) = \max\left(\min\left(\frac{x}{0.5}, \frac{1 - x}{0.5}\right), 0\right) \quad (6)$$

HIGH output qualifier is also defined by a triangle with the central point at 1 and the corners at 0.5 and 1.5. This fuzzy set is defined by the next expression:

$$\mu(x) = \max\left(\min\left(\frac{x - 0.5}{0.5}, \frac{1.5 - x}{0.5}\right), 0\right) \quad (7)$$

*3) Inference Engine:* The inference engine is a Mandani's rules based one with a centroid defuzzification method and a singleton input fuzzificator. The fuzzy engine evaluates the antecedent of every rule by the intersection of the fuzzy inputs, using the minimum function for the AND operator (Eq. 8), and the maximum function for the OR operator (Eq. 9). The implication between inputs and outputs applies the minimum function.

$$AND(a, b) = min(\mu(a), \mu(b)) \quad (8)$$

$$OR(a, b) = max(\mu(a), \mu(b)) \quad (9)$$

As mentioned, the rules must be evaluated for every single sector to estimate the confidence degree, taking into account the fuzzy qualifications of RSSI values of either the current sector nodes and the surrounding ones. The rules summed up in table I have been derived from multiple simulations in order to obtain the best trade-off between precision and noise immunity.

### *B. Cooperative Processing (CP)*

The Base Station collects the partial solutions from the anchor nodes, and it processes them cyclically as follows:

CP-S1: The Base Station waits to receive the first partial solution.

CP-S2: On arrival, the partial solution is saved and a timer starts running.

CP-S3: While the timer is running, the next partial solutions are saved in a table as they were received.

CP-S4: When the timer expires, the system will compute the final position as the centroid of all these partial solutions (triangle sectors). The centroid computation of a finite set of points $\vec{P_1}, \vec{P_2}, \cdots \vec{P_N}$ can be simplified as:

TABLE I.
RULES OF THE INFERENCE ENGINE.

| RSSI node | RSSI Neighbours | Output |
|---|---|---|
| High. | All medium. | High |
| Low. | All low. | Low |
| Medium. | All medium. | High |
| Medium. | All low. | Low |
| High. | All high. | Medium |
| Medium. | Medium in current sector.<br>Low in the rest. | High |
| Medium. | High in any sector except the current one.<br>Low in the rest. | Low |
| High. | High in a neighbour of the current sector.<br>Low in the rest. | Medium |
| High. | High in a neighbour, except on the current sector.<br>Low in the rest. | Low |
| Medium. | Medium in a neighbour of the current sector.<br>Low in the rest. | Medium |
| Medium. | Medium in a neighbour, except on the current sector.<br>Low in the rest. | Low |

$$Position = \frac{\sum_{i=1}^{N} \vec{P}_i}{N} \quad (10)$$

Previous algorithms can be easily extended to locating multiple tags, by simply associating a tag identifier to the transmitted beacons.

The final estimated position is timestamped and saved in the Base Station to make it accessible through Internet.

## IV. POWER CONSUMPTION OF LIS

Generally, power consumption is a strong constraint in a WSN application, especially for non-anchor nodes (tags), where mobility and additional constraints like size and weight do not allow use of high capacity batteries.

Most node power consumption is caused by radio transmissions. As an example, Telosb platform consumes 41 mW in active mode ($P_{On}$). The microcontroller consumes only $3\ mW$ and the remainder power consumption is caused by the radio transceiver that requires $38\ mW$ in reception mode and $35\ mW$ in transmission mode [30].

Figure 10 represents a localization algorithm computed in the non-anchor node. This is the typical execution phase of range-based or range-free algorithm, such as centroid.

As it can be observed, after the tag node broadcasts a beacon (figure 10.a), it waits for the response of all the anchor nodes placed in the radio range (figure 10.b). This phase takes a long time because of the number of surrounding nodes and because of the collisions. After that, the tag node executes the localization algorithm and delivers the result to the Base Station (figure 10.c). During all of this time, the radio transceiver must be in the active state. It wastes a lot of energy and its autonomy is considerably reduced for the non-anchor node. It is in fact, the device with the highest energy constraints. Figure 11 depictd this information. It shows the energy consumption in a generic tag that computes a localization algorithm,

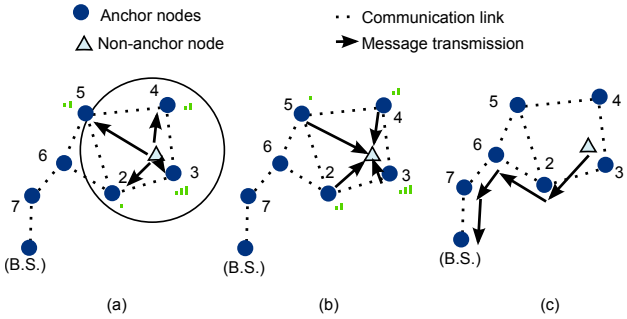



Figure 10. Localization algorithm using the non-anchor node for estimating it position.

considering the power consumption of TelosB node [30]. In this simulation the radio transceiver is maintained in the [10 s - 60 s] range, sending messages in the [1 - 60] per hour range. In these graphics, Tx, Rx and idle power consumption of the node are considered.

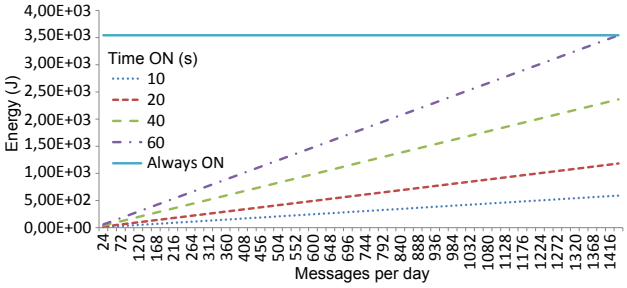



Figure 11. Energy consumption in non-anchor nodes executing localization algorithms.

As this figure shows, the energy consumption of the tag's processor in idle mode is negligible in comparison with the power consumption of the radio transceiver. Moreover, the power consumption of the node with the radio transceiver in transmission mode ($35\ mW$) is similar to the power consumption of the node with the transceiver in reception mode ($38\ mW$). Therefore, the power consumption in a WSN node would be estimated as equation 11 shows.

$$E_{node,day} \cong T_{On,day} \cdot P_{On} \quad (11)$$

It is important to point out that the power consumption is very high either in transmission and also in reception mode. Because of that, to reduce the power consumption in tags it is necessary to reduce the number of exchange messages, but it is also necessary to stop all the node activity enabling low power modes and switching off the radio transceiver. Therefore, a suitable activity manager is needed.

LIS takes this issue into account, also that anchor nodes have more power supply resources than the tags. The algorithm has been designed to be executed mainly in the anchor nodes. Furthermore, the radio transceiver of the tag is activated for a short time, just enough to broadcast the beacon. In the remaining period of time, the tag will

be in a idle state and its radio transceiver off. The time with the radio transceiver on ($T_{on}$) can be estimated, in a worst case, attending equation 12.

$$T_{on} = T_{act} + T_{msg} + T_{deact} \leq 17\ ms \quad (12)$$

Where $T_{msg}$ is the time needed to send a beacon ($T_{msg} \leq 5\ ms$, considering the preamble and a 128 bytes message length). $T_{act}$ and $T_{deact}$ are the times for activating and deactivating the radio transceiver ($T_{act} = T_{deact} = 6\ ms$ in telosB nodes). Considering this timing, the energy consumption of a tag sending a message ($E_{msg}$) can be obtained attending equation 13.

$$E_{msg} = T_{act} \cdot P_{Rx} + T_{msg} \cdot P_{Tx} + T_{deact} \cdot P_{Rx} \quad (13)$$

Where $P_{Rx}$ is the power consumption of the microcontroller with the radio transceiver in reception mode ($P_{Rx} = 41\ mW$) and $P_{Tx}$ is the power consumption of the microcontroller with the radio transceiver in transmission mode ($P_{Tx} = 38\ mW$) [30].

As it was explained in section III, with LIS non-anchor node only sends a beacon and receive nothing. This is why the energy consumption of reception is not considered.

Figure 12 shows these results, and depicts the energy consumption of LIS in a non-anchor node, sending a message in a [1 - 60] per hour range. As it can be seen, its power consumption is very low, practically equal to the power consumption of maintaining always the tags in sleep mode. This represents the best case for saving energy, where $T_{off} >> T_{on}$.

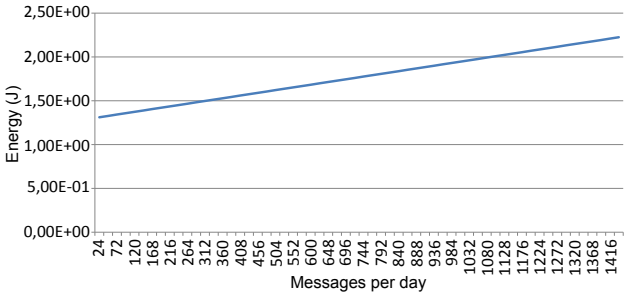



Figure 12. Energy consumption in non-anchor nodes executing LIS.

However LIS also reduces the power consumption in anchor nodes. It implements a ubiquitous and distributed algorithm that spreads the localization processing amongst the nodes surrounding the tag. In a centralized algorithm, all the information received by the anchor nodes must be delivered to the Base Station (figure 13). By contrast, our proposed algorithm saves power energy because only significant information is delivered (figure 14).

In the worst case, LIS delivers practically the same number of messages than a centralized algorithm. But for low dense deployments, for example when medium number of nodes that a beacon receives is lower than the medium number of hops necessary to reach the Base Station, the saved energy is significant.

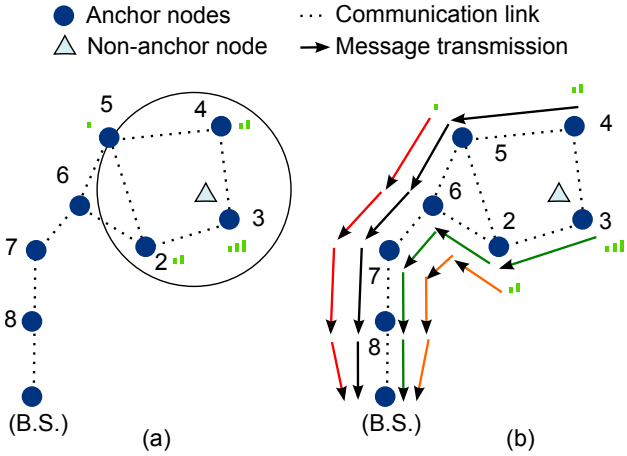



Figure 13. Example of centralized algorithm.

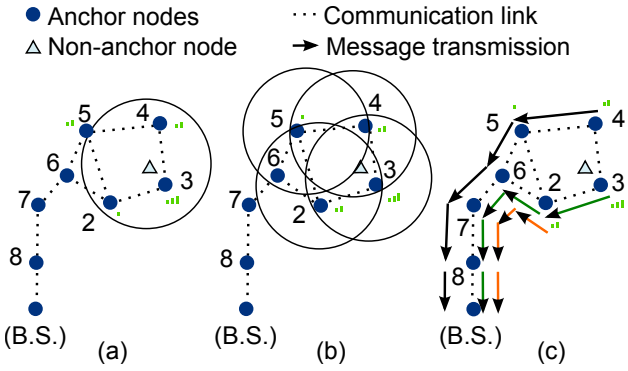


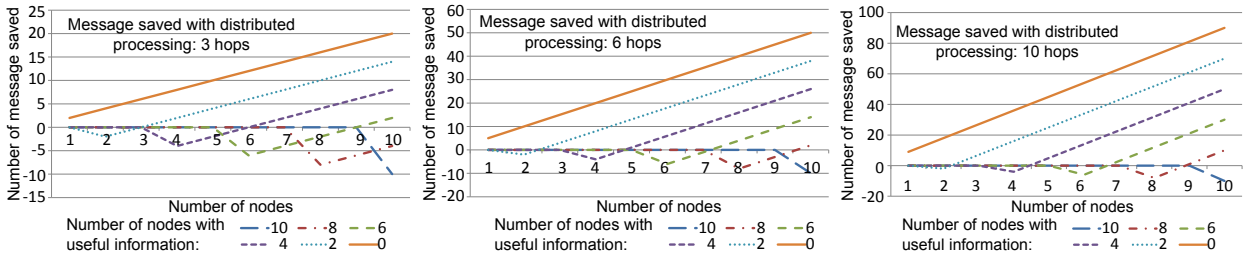

Figure 14. Example of distributed algorithm.

As a consequence, the energy saved with the distributed algorithm varies with the density and complexity of the networks. Figure 15 shows a study about the number of saved messages in function of the number nodes with useful information. From this, it can be derived that in case all the information obtained by the anchor nodes were useful, both methods send practically the same number of messages. But the more number of nodes with useless information, the energy saving performance of the distributed processing increases drastically, due to the reduction of transmissions.

Additional savings can be managed clustering the networks, and using the clusterheads as Base Stations. This is, receiving an processing partial estimations from its cluster nodes. Figure 16 shows this idea.

For this case, the algorithms must be modified as follows:

- S1: Anchor nodes wait for non-anchor node beacons.
- S2: The tag node broadcasts a beacon.
- S3: Receiver anchor nodes measure RSSI, and execute both, fuzzification algorithm and ubiquitous processing for relative and partial positioning.
- S4: Anchor nodes send partial solutions to the clusterhead, where the location is finally determined.
- S5: Clusterhead node executes the cooperative positioning algorithms and delivers the final position to the Base Station.




(a) (b) (c)

Figure 15. Messages saved versus the number of nodes: a) Three hops; b) Six hops; c) Ten hops.

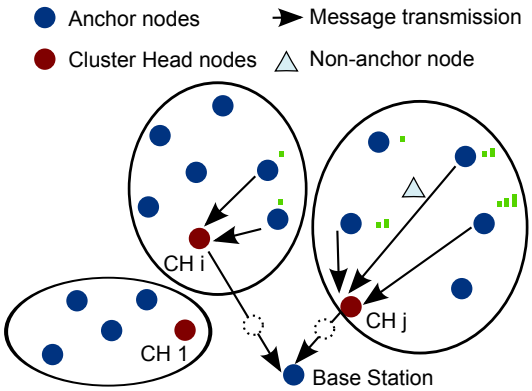



Figure 16. Example of the use of clusters.

S6: Base Station executes the same cooperative positioning algorithm than the clusterhead nodes, but using the information delivered from these clusterheads. In this way, if the tag positioning comes from just one clusterhead, this position will be considered as the final solution. But, if it is received from more than one clusterhead, the centroid estimation is applied to all of them.

Determining when the use of clustering saves more energy is not trivial. It depends on the size and complexity of the network. But in general, it is reasonable to think that clustering techniques are better for wide and complex networks.

Additionally, a distributed processing such as the one proposed in this paper, increases throughput and reduces the response delay, because traffic bottleneck and collisions close to the Base Station are avoided. Distributed processing also spreads computational load over the network. This is especially important for wide networks or with multiple non-anchor nodes.

## V. Accuracy of LIS

We have compared the accuracy of LIS versus the classic CL algorithm [25] using different simulations. The tested network was made up of 25 anchor devices with a radio range of 200 meters and separated also by about 200 meters. Anisotropic radiation pattern is assumed. The simulator has been developed in C++. It allows the selection of radio range, radiation pattern, noise, sensibility, network deployment and anchor location. Friis equation is used to calculate the received RSSI level in every node.

All parameters, in the tests have been selected to model Telosb devices.

The results of the simulations are presented in the next subsection.

### A. Error vs. Position

The following experiments include a moving tag. The noise has been neglected and the error is expressed in meters. Figures 17 and 18 show the position error in front of the position of the non-anchor node. Maximum and medium errors of LIS algorithm are considerably smaller than the ones estimated with the (CL) Centroid classic algorithm.

As we can see in figure 18, with LIS the errors always remain less than a 25 % of the distance between anchor nodes.

### B. Error vs. Coverage

In this test, radio range is increased while the distance between nodes remains constant.

Figure 19 shows the influence of radio range. As the radio range increases, the number of non-anchor nodes that receive a beacon also increases, and the error decreases. In all the cases, LIS gets smaller errors than centroid algorithm.

Similar results are obtained fixing the coverage area of the non-anchor nodes and reducing the distance between anchor devices.

### C. Error vs. Noise

In this test we have analysed the error obtained in a fixed position of the non anchor nodes when adding Gaussian noise to the system. All analyses use a coverage area of 200 m for the tag. We have made 1000 analysis at every point for the simulations.

The number of errors (figure 21) represents the number of absolute errors bigger than 100 m (1/2 of the coverage area) obtained in the localization. As it can be seen, LIS always has less number of errors bigger than 100 m. than the classic CL algorithm.

This result shows that despite the interferences, LIS continues to offer good accuracy for localization.

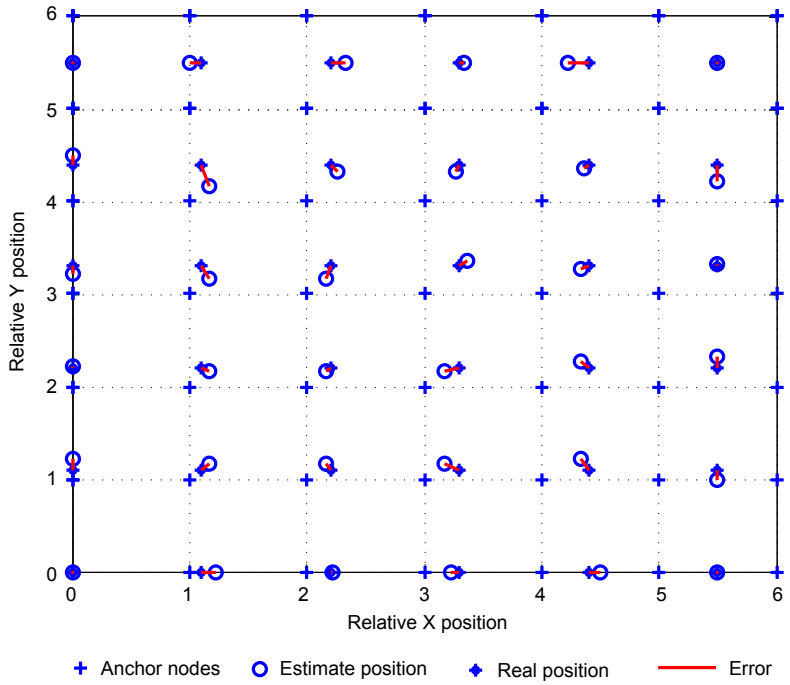



Figure 17. Position error of LIS algorithm.

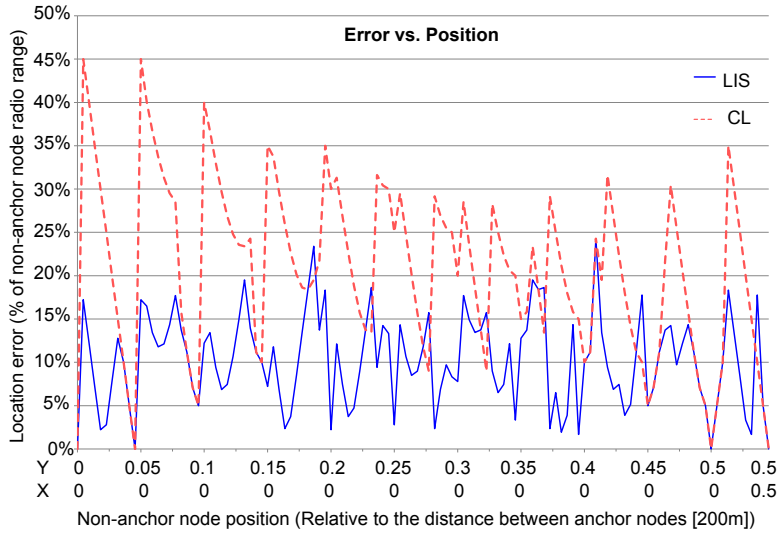



Figure 18. Localization error vs. the position of the non-anchor node.

It is important to consider that to obtain this result, there is no use of any class of filter with the results. To estimate the position, the system only used the current information and not the past estimations. Filtering the results, the accuracy of the system versus the noise would be improved.

### *D. Error vs. Battery discharging*

As described in [31], the power of transmission messages decreases as the battery is discharging. This paper shows that during the lifetime of the battery of a node, the power transmission drops practically linearity up to 10 $bB$ (figure 20). Its attenuation will cause many problems with localization algorithms, but this constraint is frequently not considered in the design of the localization algorithms. Commonly, RFID localization techniques uses a calibration, and a localization based in a footprint matching from the training patterns. When the energy in the battery decreases, these techniques requires a recalibration of the environment.

In WSN, RSSI range-based techniques, the reduction in the power transmission may cause that the algorithm not obtain an estimation of the position. For example, using triangulation and a estimation of the distance with the Friis model. If the power transmissions drops, the circles of distances obtained with Friis do not intercept in a point,

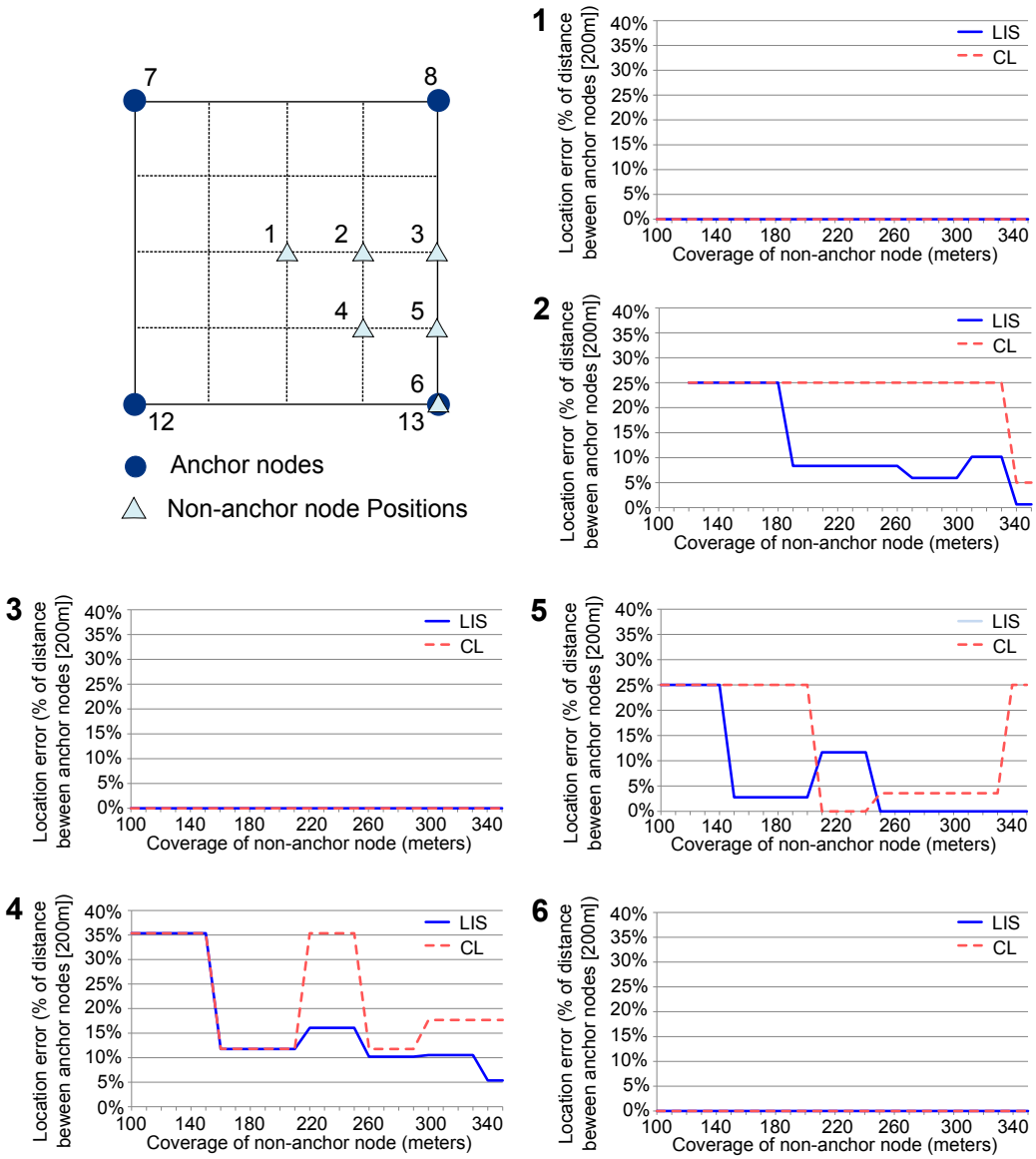



Figure 19. Localization error vs. the coverage radio area of non-anchor node.

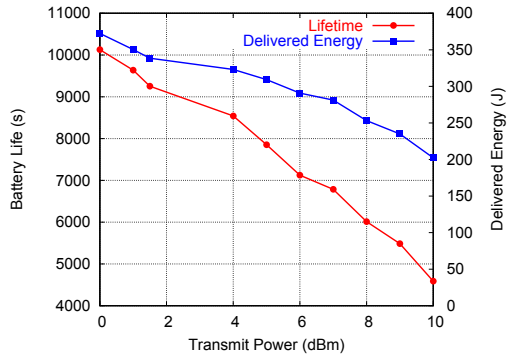



Figure 20. Impact of transmission power level on battery life and delivered energy [31].

an area of inaccuracy appearing where the non-anchor node would be.

Moreover, the range-free algorithms, as the presented system, are more robust versus a variation of the power TX transmission than the range-based ones. This is because to these techniques do not use the RSSI, or use RSSI without a mathematical model that determine a distance in function of a power TX transmission.

The simulator permits simulating these conditions, assuming a coverage area of 300 meters and a attenuation of 10 $dB$ during all the lifetime of the battery. These results are depicted in figure 22.

These graphs represent the estimated relative error position along the duration of the battery. These results show that LIS offers good accuracy versus the lack of power transmission due to the discharge of the battery. LIS offers better results than the classical centroid algorithms. These results are similar to the ones obtained with other range-free algorithms, typically better in this sense than range-based algorithms.

With LIS, the robustness versus the battery discharging

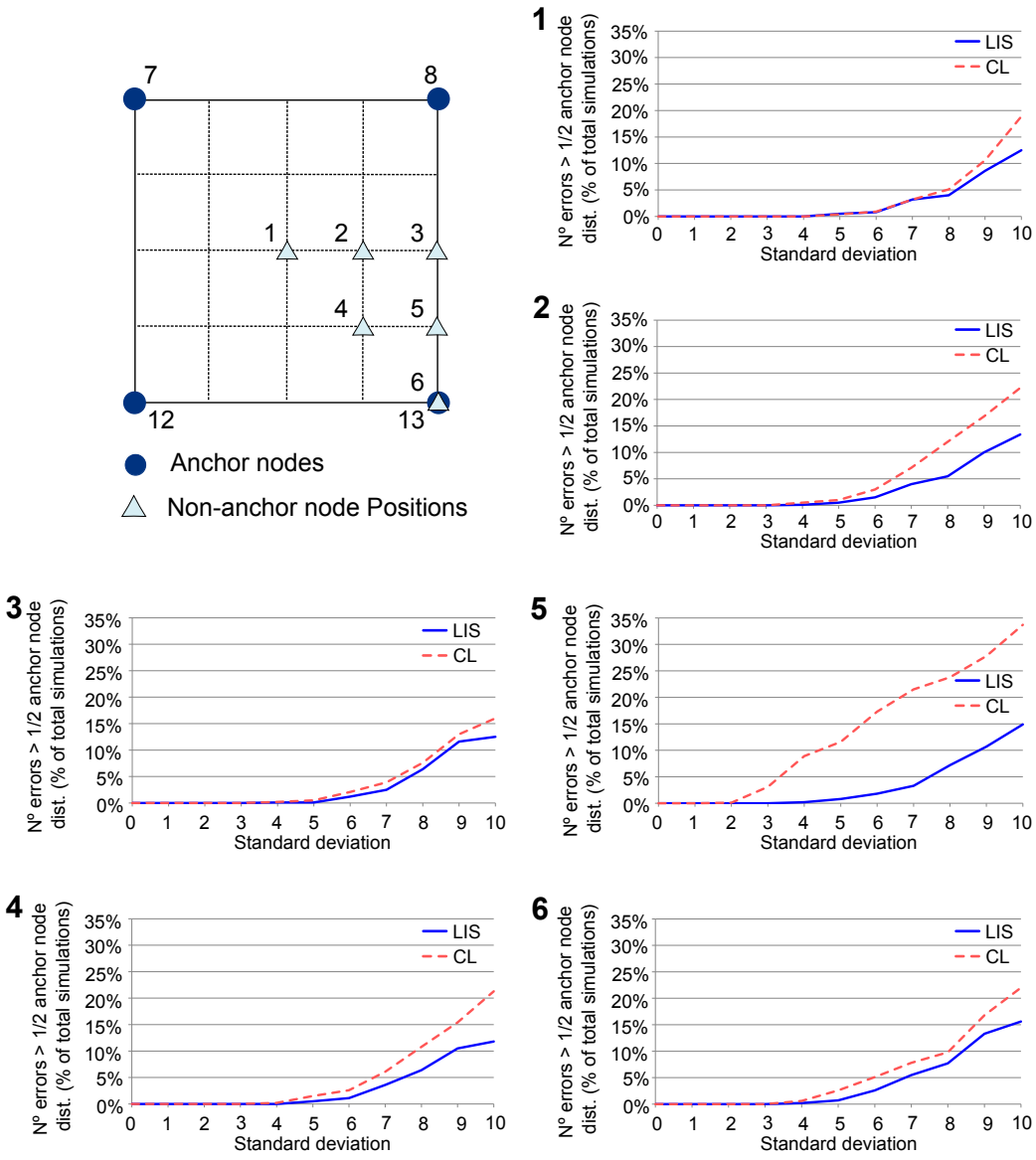



Figure 21. Number of error bigger than 100 m.

would be improved using the measurement capabilities of the WSN nodes. The tag can measure its battery charge level, and transmit it in the beacon transmission. The nodes would use these value to calibrate the high, medium an low RSSI level of the fuzzy algorithm. It is not possible with classical RFID tags, and its one of the biggest advantages of WSN in front of active RFID: WSNs permit monitoring and measuring additional variables of the environment.

## VI. Conclusions

LIS is a new localization system for Ambient Intelligence designed to reduce power consumption, especially but not limited to, the tag nodes where power constraints are higher. LIS filters the useless information after being processed in the anchor nodes. It also implements a hibernation mechanism. All these mechanisms increase battery autonomy.

LIS has been tested by simulations. They show that the proposed method obtains less localization errors than the well known CL algorithm without higher computation requirements or an extensive use of radio.

It is important to consider that with LIS the localization process starts when the tag sends a message. I.e. the tags control the refresh time of the localization information. In the proposed system, we considered a fixed time between message transmission, but it can work correctly even with a variable activation time.

The tag can improve its power consumption in two ways: (I) sending the beacons only if it detects a change (e.g. using an accelerometer), or (II) increasing the time between messages when the battery level decreases.

## Acknowledgment

This research has been supported by the “Consejería de Innovación, Ciencia y Empresa”, “Junta de An-

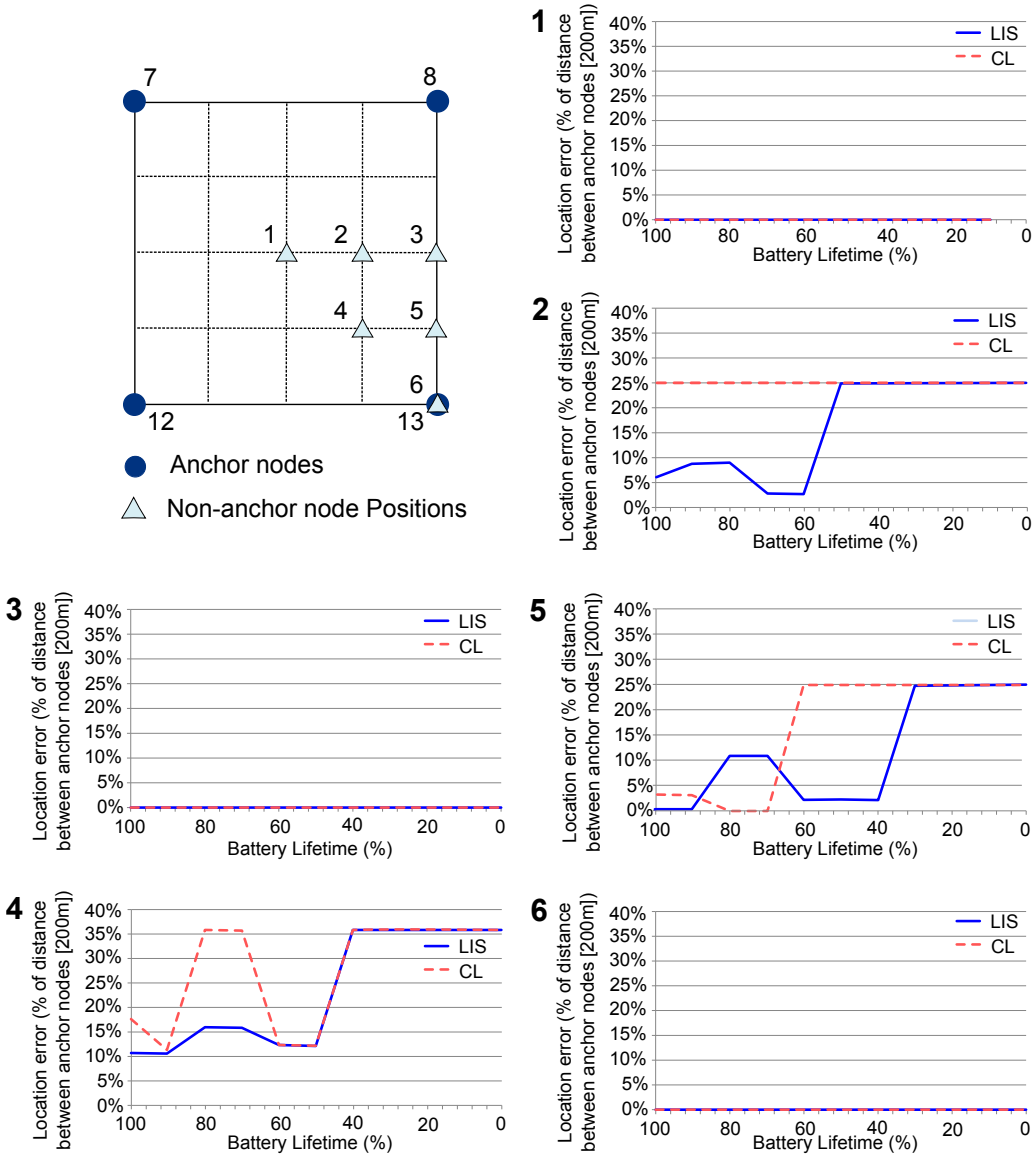



Figure 22. Error versus the battery lifetime.

dalucía", Spain, through the exellence project ARTICA (reference number: P07-TIC-02476) and by the "Cátedra de Telefónica, Inteligencia en la Red", Seville, Spain, through the project ICARO.

## REFERENCES

[1] F Michael, O Da Costa, Y Punie, P Alahuhta, S Heinonen. Perspectives of ambient intelligence in the home environment, Telematics and Informatics, Volume 22, Issue 3, (2005), Pages 221-238, ISSN 0736-5853.

[2] E Aarts. Ambient intelligence: a multimedia perspective, on Multimedia, IEEE , vol.11, no.1, pp. 12- 19, (2004).

[3] V Menon, B Jayaraman and V Govindaraju. Multimodal identification and tracking in smart environments. Personal and Ubiquitous Computing (2010), Springer London, vol. 14, pp. 685-694.

[4] P Ross, D Keyson. The case of sculpting atmospheres: towards design principles for expressive tangible interaction in control of ambient systems. Personal and Ubiquitous Computing (2007). Springer London, vol 11.

[5] N Streitz and P Nixon. The disappearing computer. Communications of the ACM, (2005), Vol. 48 Issue 3.

[6] C Ramos, J C Augusto, D Shapiro. Ambient Intelligence: the Next Step for Artificial Intelligence, Intelligent Systems (2008), IEEE , vol.23, no.2, pp.15-18.

[7] Remagnino, P.; Foresti, G.L.; ,Ambient Intelligence: A New Multidisciplinary Paradigm," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on (2005), vol.35, no.1, pp. 1- 6.

[8] J Bravo, L Fuentes, D de Ipiña. Theme issue: ubiquitous computing and ambient intelligence. Personal and Ubiquitous Computing (2011), Springer London. pp. 1-2

[9] J Lindenberg, W Pasman, K Kranenborg, J Stegeman and M Neerincx. Improving service matching and selection in ubiquitous computing environments: a user study. Personal and Ubiquitous Computing (2007). Springer London, pp. 59-68, vol, 11.

[10] IF Akyildiz, W Su, Y Sankarasubramaniam, E Cayirci. Wireless sensor networks: A survey, Comput.Networks. 38 (2002) 393-422.

[11] J Yick, B Mukherjee, D Ghosal. Wireless sensor network survey, Comput.Networks. 52 (2008) 2292-2330.

[12] JA López Riquelme, F Soto, J Suardíaz, P Sánchez, A Iborra, JA Vera. Wireless Sensor Networks for precision horticulture in Southern Spain, Comput. Electron. Agric. 68 (2009) 25-35.
[13] W Chung, Y Lee, S Jung, A wireless sensor network compatible wearable U-healthcare monitoring system using integrated ECG, accelerometer and SpO2, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., EMBC - Pers. Healthc. through Technol. (2008) 1529-1532.
[14] T He, S Krishnamurthy, JA Stankovic, T Abdelzaher, L Luo, R Stoleru, et al. Energy-efficient surveillance system using wireless sensor networks, MobiSys Second Int.Conf.Mobile Syst .Appl. Serv. (2004) 270-283.
[15] A Boukerche, HABF Oliveira, EF Nakamura, AAF Loureiro. Secure localization algorithms for wireless sensor networks, IEEE Commun Mag. 46 (2008) 96-101.
[16] L Benini, E Farella, C Guiducci. Wireless sensor networks: Enabling technology for ambient intelligence, (2006) Microelectronics Journal, 37 (12), pp. 1639-1649.
[17] Chatzigiannakis, I Kinalis, A Nikoletseas, S Sink. Mobility protocols for data collection in wireless sensor networks (2006) MobiWAC 2006 - Proceedings of the 2006 ACM International Workshop on Mobility Management and Wireless Access, (2006), pp. 52-59.
[18] D J Cook, J C Augusto, V R Jakkula. Ambient intelligence: Technologies, applications, and opportunities, Pervasive and Mobile Computing, Volume 5, Issue 4, (2009), Pages 277-298, ISSN 1574-1192.
[19] C Stephanidis, T Kleinberger, M Becker, E Ras, A Holzinger and P Müller. Ambient Intelligence in Assisted Living: Enable Elderly People to Handle Future Interfaces. Universal Access in Human-Computer Interaction. Ambient Interaction. Lecture Notes in Computer Science (2007), Springer Berlin / Heidelberg, pp. 103-112, vol. 4555.
[20] dizic 802.15.4 ARM modules, available on `http://www.dizic.com`. Last visited: July 2012.
[21] H Piontek, M Seyffer and J Kaiser. Improving the accuracy of ultrasound-based localisation systems. Personal and Ubiquitous Computing, Springer London (2007), vol. 11 pp. 439-449.
[22] A Awad, T Frunzke, F Dressler, Adaptive distance estimation and localization in WSN using RSSI measures, Proc. - Euromicro Conf. Digit. Syst. Des. Archit., Methods Tools, DSD. (2007) 471-478.
[23] S Wu, N Zhang. Two-step TOA estimation method for UWB based wireless sensor networks, Ruan Jian Xue Bao. 18 (2007) 1164-1172.
[24] P Rong, ML Sichitiu. Angle of Arrival Localization for Wireless Sensor Networks, Sensor and Ad Hoc Communications and Networks, 2006. SECON '06. 2006 3rd Annual IEEE Communications Society on. 1 (2006) 374-382.
[25] N Bulusu, J Heidemann, D Estrin. GPS-less low-cost outdoor localization for very small devices, IEEE Pers Commun. 7 (2000) 28-34.
[26] GQ Gao, L Lei. An improved node localization algorithm based on DV-HOP in WSN, Proc. - IEEE Int. Conf. Adv. Comput. Control, ICACC. 4 (2010) 321-324.
[27] S Rajaee, SMT Almodarresi, MH Sadeghi, M Aghabozorgi. Energy efficient localization in wireless ad-hoc sensor networks using probabilistic neural network and independent component analysis, Int. Symp. Telecommun., IST. (2008) 365-370.
[28] F Xiufang, G Zhanqiang, Y Mian, X Shibo. Fuzzy distance measuring based on RSSI in Wireless Sensor Network, Proc. Int. Conf. Intell. Syst. Knowl. Eng., ISKE. (2008) 395-400.
[29] S- Chiang, J- Wang. Localization in wireless sensor networks by fuzzy logic system, Lect. Notes Comput. Sci. 5712 LNAI (2009) 721-728.
[30] J Polastre, R Szewczyk, D Culler. Telos: Enabling ultra-low power wireless research, (2005) 4th International Symposium on Information Processing in Sensor Networks, IPSN 2005, 2005, art. no. 1440950, pp. 364-369.
[31] C Park and K Lahiri. Battery discharge characteristics of wireless sensor nodes: An experimental analysis, In Proceedings of the IEEE Conf. on Sensor and Ad-hoc Communications and Networks (SECON), 2005.

**Diego Fco. Larios** received his degree in Industrial Electronics and Automatic Control Engineering in 2009. Both of then in the University of seville, Spain.

Currently, he is pursuing his Ph.D. in Industrial Informatic at the same University.

He is a member of the research group "Tecnología Electrónica e Informática Industrial" at the University of Seville.

His research interests are in the area of low-power wireless sensor networks, computational intelligence, automatic control and industrial automation.

**Julio Barbancho** received his degree in telecommunication engineering in 2002 from the University of Seville, Spain.

He obtained in 2008 his Computer Science Doctoral Degree at the same university, where he works as a professor at the Department of Electronic Technology, since 2002.

His research interests are in the area of analysis and design of wireless sensor networks, and control applications based on expert systems, fuzzy logic and neural networks techniques.

**Fco. Javier Molina** received his Physical Electronics degree in 1987 from the University of Seville, Spain. Since 1989 he works as a professor at the same university.

He is a member of the research group "Tecnología Electrónica e Informática Industrial" at the University of Seville.

His research interests are related with wireless sensor networks and emerging technologies for applications in automation.

**Carlos León** received the B.Sc. degree in electronic physics in 1991 and the Ph.D. degree in computer science in 1995, both from the University of Seville, Seville, Spain. He is a professor of Electronic Engineering at the University of Seville since 1991 and currently CIO of the University of Seville. His research areas include knowledge based systems and computational intelligence focus on Utilities System Management. He is a Member of the IEEE Power Engineering Society.

# Development Tools for Context Aware and Secure Pervasive Computing in Embedded Systems (PECES) Middleware

Ran Zhao, Neil Speirs
School of Computing Science, Newcastle University, Newcastle NE1 7RU, United Kingdom
{ran.zhao1, neil.speirs}@ncl.ac.uk

Kirusnapillai Selvarajah
Department of Electrical and Electronic Engineering, The University of Nottingham, Nottingham NE7 2RD, United Kingdom
kirusnapillai.selvarajah@nottingham.ac.uk

***Abstract* — The main objective of the PECES project is the development of system software to enable the communication among heterogeneous devices across multiple smart spaces, breaking the traditional barrier of "smart islands" where only the services offered in a nearby spatial area can be used easily. PECES development tools help the application developer to build and test the PECES middleware based applications. This paper presents a set of tools, namely Peces Project, Peces Device Definition, Peces Ontology Instantiation, Peces Security Configuration, Peces Service Definition, Peces Role Specification Definition, Peces Hierarchical Role Specification Definition, Peces Event Editor, Peces Event Diagram and Peces Testing which enable application developers to build, model and test the PECES middleware based smart space application using the novel concepts such as role assignment, context ontologies and security.**

***Index Terms*—smart space, middleware, pervasive computing, wireless networking, context ontologies, modelling, testing, dynamic addressing, eclipse plugins, communication gateway, registry interface.**

## I. PECES PROJECT

The objective of the PECES project [1] is the creation of a comprehensive software layer to enable the seamless cooperation of embedded devices across various smart spaces on a global scale in a context dependent, secure and trustworthy manner. The increasing number of devices that are invisibly embedded into our surrounding environment as well as the proliferation of wireless communication and sensing technologies are the basis for visions like ambient intelligence, ubiquitous and pervasive computing. The benefits of these visions and their undeniable impact on the economy and society have led to a number of research and development efforts. These include various European projects such as EMMA [2], [3] that develop specialized middleware abstractions for different application areas such as automotive and traffic control systems or home automation. Middleware for pervasive environments primarily manages the stationary infrastructure in the environment. Usually, this infrastructure consists of stationary devices deployed within a predefined physical location such as meeting room or car parking. There are several other middleware systems have been developed for this purpose such as Aura [19], Gaia [12], [13] and IROS [20]. Reliable service management framework is proposed in [21] by formally defining a message-oriented service application model and protocols that facilitate autonomous composition, failure detection and recovery of services.

These efforts have enabled smart spaces that integrate embedded devices in such a way that they interact with a user as a coherent system. However, they fall short of addressing the cooperation of devices across different environments. This results in isolated "islands of integration" with clearly defined boundaries such as the smart home or office. For many future applications, the integration of embedded systems from multiple smart spaces is a primary key to providing a truly seamless user experience. Nomadic users that move through different environments will need to access information provided by systems embedded in their surroundings as well as systems embedded in other smart spaces. The PECES project is committed to developing the technological basis to enable the global cooperation of embedded devices residing in different smart spaces in a context-dependent, secure, and trustworthy manner. The most innovative features of the PECES middleware are to enable the communication among heterogeneous devices across the different smart spaces using dynamic addressing, security and context ontologies.

The PECES project has the following scientific and technical objectives:

1. Development of a flexible ontology to capture the context of cooperating objects and to specify groups of cooperating objects in an abstract manner.
2. Development of a middleware – i.e. a set of application independent services that enable the dynamic


doi: 10.4304/jnw.8.1.59-70

and context aware formation of a secure execution environment from a set of cooperating objects.

3. Development of a set of application development tools that simplify the formation of groups and the description of the context of cooperating objects.

4. Validation of the abstractions using lab tests and prototype applications.

In this paper, the PECES development tools are discussed. Section II describes the PECES project and its novel features such as role specification, dynamic addressing, gateway concepts and registry concepts. Section III provides an introduction to the development environments. Section IV describes the PECES developments tools that have been developed by the PECES consortium for PECES middleware based application development. The PECES development tools evaluation methods and findings are given in Section V. Future work and conclusions are then presented in Section VI.

## II. PECES MIDDLEWARE

The PECES project consortium built the targeted cooperation layer on top of BASE middleware [9]. This enables the project consortium to focus the development efforts on the novel and innovative features of the PECES middleware. BASE is freely available as open source under BSD license which facilitates the necessary modifications and extensions and enables the free reuse – even for commercial exploitation.

### A. BASE Middleware

The overall architecture of BASE is divided into three layers. At the highest layer – the application layer, local and remote application objects and system services interact with each other. This layer relies on the functionality offered by the middleware core which is represented by the micro-broker layer. The micro-broker layer, in turn, uses the capabilities of the plug-in layer to discover remote devices and to communicate with them. Figure 1 shows the BASE three layer architecture and its main components.

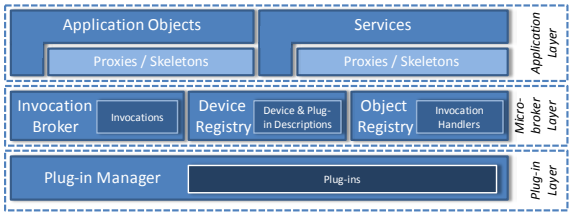



Figure 1: BASE Architecture

### B. PECES Middleware

The BASE middleware enables the communication between devices that are within communication range. Yet, in order to achieve the goal of providing cooperation layer that enables the seamless interaction within and across the boundaries of a single smart space, it is necessary to extend the basic concepts. The extension of the BASE middleware focused communication gateway concepts, addressing concepts and smart space concepts.

#### 1. Communication Gateway

Due to the heterogeneity of devices and communication technologies, it is not safe to assume all future devices will be equipped with the same set of communication technologies. As an example consider that a sensor node might only be equipped with ZigBee (based on IEEE 802.15.4) but not with Bluetooth in order enable energy-efficient communication. Thus, in order to enable a Bluetooth device to communicate with such sensor nodes, it is necessary to use a device that is equipped with both technologies as a local gateway. Similarly, due to the associated costs and other factors, not all devices will have a direct connection to a global interconnection network like the Internet. In order to enable the communication between devices that are not directly connected to the Internet, it is necessary to enable some devices to act as remote gateways for others.

PECES middleware support local gateways as well as remote gateways. The main difference between these two types of gateways is that the local gateway locally shares the required knowledge. In the remote case, the knowledge sharing should be restricted to a minimum in order to avoid the costly distribution of frequently changing information. The remote gateways need to be realized differently in that they require an external entity to distribute the information that is distributed by means of device discovery in the local case. This information will be distributed by means of the registry that is specified in the PECES Communication Mechanism and Registry Interface Specification [8].

#### 2. Generic Role Assignment

Due to the continuous changes in context and due to the mobility of devices the underlying systems can be highly dynamic and the network topology can change frequently in the pervasive computing environments. So that it is vital to enable pervasive computing applications such as PECES prototype applications to adapt to the continuous changes in context and device availability. The responsibility for adaptation can be shifted between different entities. In cases where changes are infrequent, a user may manually configure and adapt the system. However, if changes are frequent, manual configuration and adaptation are clearly not a viable approach as they conflict with the goal of distraction free support for tasks. In order to mitigate this, the adaptation can be automated through the application. This approach relieves the user from performing manual adaptation but it complicates the development of applications and it may result in inefficiencies in cases where multiple applications implement and use similar adaptation mechanisms. As a result, the PECES middleware is aiming at automating the initial configuration and the continuous adaptation to changes in order to shield the user and the application developer from the accompanied complications.

In order to be suitable for a broad range of different systems and in order to minimize the utilization of resources that are required for automation, PECES middleware provides configuration and adaptation support by means of a uniform abstraction. To create a

uniform abstraction that is suitable for a broad range of different configuration tasks, it is necessary to introduce a clear separation between the result of a configuration, the computations that need to be done to produce it and the utilization of this result. This enables the reuse of the same basic mechanisms for different tasks. Generic role assignment provides such a uniform abstraction. More detailed information about the role assignment concepts can be found in [7].

A role can be assigned to any device as long as there are no further constraints that limit the assignment. To enable the automated computation of an assignment that reflects a particular goal of a configuration task, generic role assignment introduces rules. Rules define contextual constraints on the assignment of roles to devices. The simplest form of contextual constraint that is generally useful for all configuration tasks is a simple filter. An example of such a filter is to demand that all devices should be at a certain location. Another form of contextual constraint that is particularly relevant for PECES middleware are so called reference rules. Reference rules refer to a set of devices that has been assigned a particular role.

The set of rules together with their corresponding roles form a role specification. Given that the necessary contextual information can be captured by sensors or other types of information sources, one can use an algorithm to automatically assign roles to the devices whose context satisfies the constraints specified by their rules.

### *3. Smart Space Concept*

A smart space can be defined as a group of networked devices that cooperate to support their users. The boundaries of a smart space are typically defined on the basis of a geographic location, e.g. a room or a building. However, such narrow definitions are not flexible enough to support the application prototypes in the PECES project. Obviously, these smart spaces cannot be defined on the basis of a single location. For example in applications based upon a car, the whole car, i.e. the smart space itself, is mobile.

In order to extend the definition, the addressing and grouping scheme can be used to support the formation of smart spaces based on arbitrary contextual properties. However, as explained earlier, the resulting definition will be automatically restricted to devices that are residing in the same local network. This is a result of the fact that the formation process of basic groups is limited to a local network. Yet, for typical smart spaces local connectivity is guaranteed.

To support smart space formation, the PECES middleware introduces three additional components which are coordinator, member and gateway. These components can be easily motivated by looking at the anatomy of the smart spaces that are identified in the PECES Use-Case Specification [4]:

- Coordinator: A smart space consists of at least one coordinator device. This device is responsible for identifying members of the smart space based on role specification.
- Member: In addition to coordinator device, a smart space may contain additional devices that are dynamically entering or leaving the local network. Depending on the context, a member device might either be integrated or not. Currently, a member device can only be integrated at most into one smart space at a time.
- Gateway: Some devices that are part of a smart space may also be able to communicate with other devices through an Internet connection. Examples for such devices are smart phones or residential gateways as well as laptops that are equipped with a UMTS modem. In these scenarios, the PECES middleware gateway functionality provides connectivity for other devices in the smart space.

### *4. Registry Interface Concept*

The PECES middleware provides a collaboration mechanism that enables communication between devices in and across different smart spaces in a context dependent manner. This requires a mechanism that allows devices to discover and access information about each other. The BASE middleware provides an internal service and device registries to maintain the access information of the locally available services. Since the services are accessed not only from inside but also outside the smart space, the PECES middleware provides a distributed registry mechanism that can further be extended for remote group formation. The distributed registry can facilitate the information distribution of the assigned roles in the smart spaces coupled with necessary context information for forming an overarching environment. The PECES cooperation layer can thus use the distributed registry to lookup for devices based on the role and retrieve necessary plug-in information to access the devices. Hence, the distributed registry enables cooperation between heterogeneous devices by identifying the relevant devices that may provide services or can be used to form an overarching smart space on top of existing smart spaces.

The spontaneous appearance and disappearance of the devices in a typical smart space naturally requires a registry infrastructure where information about the services and roles are easily but securely accessible. The accessibility and security trade-off impose a natural scoping on the service availability. From a device perspective, the required services for an application may reside on the same device, or may be available on a remote device that may or may not be the part of the same smart space. This clearly outlines three different scopes for available services namely; “Device”, “Space” and “Internet”. More detailed information about the PECES registry concepts can be found in [8]

### *5. Security Concept*

The PECES consortium extends the PECES middleware to derive a secure middleware. For this, PECES consortium introduced a basic trust model that is used as basis for the concepts and mechanisms of the middleware. These mechanisms enable the secure interaction of devices. To enable this, they span the

management of cryptographic keys, the authentication of information – specifically context information and role assignments, the secure data and service centric communication as well as role based access control. Although they do not introduce additional interaction features, together they span the whole set of security related requirements that have been identified in the PECES Requirements Specification [5] and thus, they are sufficient to be applicable to a broad range of scenarios.

The PECES security mechanisms are modular and they introduce a certain degree of configurability that can be leveraged by application developers for optimization purposes. This enables them to define application specific trade-offs between security and application performance. In order to simplify the configuration of these mechanisms, the PECES consortium provides set of tools (which will be discussed in Section IV) that simplify basic security related tasks such as the distribution of keys and certificates during application development.

## III. Development Environments

The PECES project provides a set of development tools that can be used by the application developers to develop, test, and analyse their applications. The Development tools also provide an environment to simulate/emulate applications. These types of development tools are economical because developers can carry out experiments without the actual hardware and it is a feasible way to test scalability of any proposed applications. For wired/wireless networked based application and protocol development, there are many simulators have been proposed.

Pervasive computing application developers need development tools for rapid development and evaluation of novel smart space systems application. This is especially true for dealing with heterogeneous device environments with context based smart space formation as proposed in the PECES prototype applications [4], [5]. To the best of our knowledge, not many frameworks are available for effective simulation, emulation and testing of smart space system applications. However, there are some general purpose development tools that are available to test/simulate pervasive computing based applications.

A Middleware based application framework for Active Space applications is proposed in [12]. The Active Space consists of the Gaia middleware OS [13] managing a distributed system composed of plasma displays, a video wall, audio system, touch screens, IR beacons, badge detectors, wireless and wired networks connecting several Windows 2000 and PDAs. The framework focuses on providing an application framework that leverages the functionality provided by the Gaia middleware OS to assist developers in the construction of Active Space application. The application framework defines an application model that accommodates the requirements of Active Spaces including dynamically changing the cardinality, location, input, output and processing devices used by an application. Then the application framework provides a mapping mechanism to define applications requirements and automatically mapping them to the resources present in a particular Active Space. Finally, the framework implements a flexible policy driven application management interface that allows customising applications to the dynamic behaviour of Active Spaces.

The Distributed Trust Toolkit (DTT) [11] proposed a framework for implementing and evaluating trust mechanisms in pervasive computing systems and introduced two new abstractions: trust groups and trust blocks. Trust groups allow associated application devices to share recorded trust data and trust computations. Trust blocks makes policy decisions based on data gathered by the computation component which implements network based trust protocols and allows the DTT to interoperate with legacy trust systems. The Distributed Trust Toolkit facilitates the extension and adaptation of trust mechanisms by abstracting trust mechanisms into interchangeable components. Furthermore, the DTT provides a set of tools and interfaces to ease implementation of trust mechanisms and facilitates their execution on a variety of platforms and networks.

UbiWise, a simulator for ubiquitous computing system was proposed in [15]. UbiWise concentrates on computation and communication devices situated within their physical environments. Multiple users can attach to the same server to create interactive ubiquitous computing scenarios. The devices are specified through a combination of a device-description file in XML and Java. UbiREAL simulator [16] was proposed for realistic smart space systematic testing. UbiREAL facilitates reliable and inexpensive development of ubiquitous applications where application software controls a lot of information appliances based on the state of external environment, user's contexts information. The simulator realistically reproduces behaviour of application software on virtual devices in a virtual 3D space. Interestingly, it provides mechanisms to simulate virtual devices with real devices.

The tools discussed here provide limited support for application developers, as they have been designed for different goals and concepts. Although these tools are available to support pervasive computing environment application development, they only offer little methodological support for role assignment, context awareness and security, which are the core features of the PECES middleware. For example, Protégé [22] may provide support for context ontology instantiation for the devices but it may be difficult to integrate with other tools the way application developers wanted for smart space application development. The following sections present set of tools that have been developed for building novel smart space applications using the PECES middleware.

## IV. PECES Development Tools

PECES development tools focus on configuring devices, modelling smart spaces and context dynamics and testing the novel concepts provided by the PECES middleware. The tools provide support for application developers to build PECES middleware application and

simulate and analyse the smart space behaviours with respect to the context changes and network changes. Instead of running PECES application on real devices, application developers are able to test the features of the PECES middleware in a development PC for any specific application. This provides the opportunity for the application developers to test and analyse their application in a controlled and repeatable environment which enable them to optimise certain parameters which may be necessary for the best performance of any smart space applications.

The PECES development tools provide a set of tools which are integrated into the Eclipse development environment (as Eclipse Plugins). This way, the usual development assistance provided by the Eclipse IDE can also be used for PECES focused development support.

The following sections explain how to use the PECES development tools to build middleware applications. The tools are implemented as an Eclipse plugins and use a wizard and multipage editor approach. The development tools provide several tools for different purposes during application development. Earlier version of this tool was presented in [23]. The following sections explain the tools (listed in Figure 2) for configuration of the application such as Peces Project, Peces Device Definition, Peces Ontology Instantiation, Peces Security Configuration, Peces Service Definition, Peces Role Specification Definition, Hierarchical Role Specification Definition. For modelling purposes, Peces Event Editor and Peces Event Diagram tools are discussed. Finally for testing purposes, Peces Testing tool is explained. The following subsections are arranged according to the development sequence that the developers would typically follow during the smart space application development.

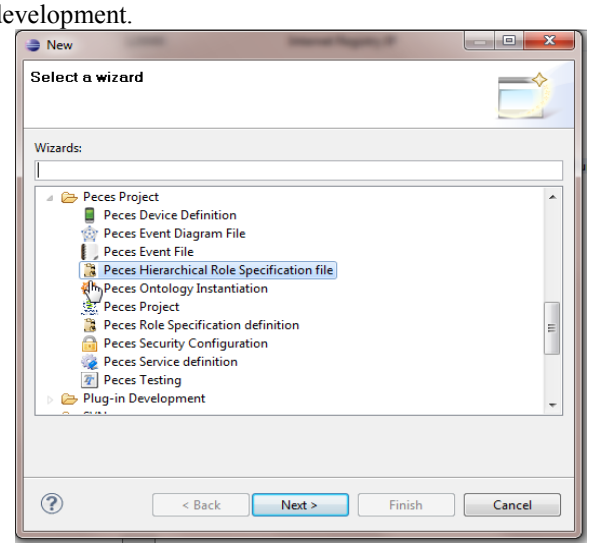



Figure 2: Screenshot of the Peces Project Tool

### A. Peces Project Tool

This is the first part of the application development process. Using this tool, a project can be generated in the Eclipse workspace with three different folders to keep different configuration, modelling and testing related files which will be generated by other tools. For example, a *PrototypeDemo* project was created by the Peces Project tool with *ConfigurationTool*, *ModellingTool* and *TestingTool* folders (as shown in Figure 4).

### B. Peces Device Definition Tool

The Peces Device Definition tool can be used to define communication plugins such as IP, Bluetooth, ZigBee (e.g. *MxIPBroadcastTransceiver, MxIPMuticastTransceiver, MxSpotTransceiver*), and device functionalities (*Coordinator, Gateway, Coordinator&Gateway, Member*) and also device name. All smart space applications should define one device with coordinator functionality which is responsible of the coordination of the smart space. The devices can be placed using drag and drop method. Different colours will be shown according to the selected device functionality (e.g., Coordinator is red).

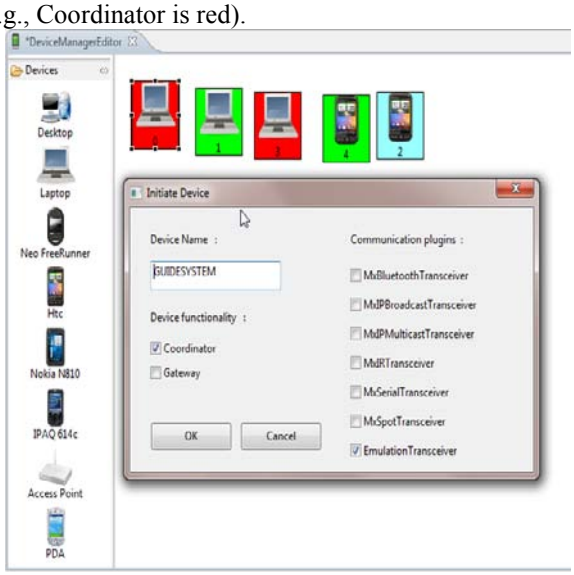



Figure 3: Screenshot of the Peces Device Definition Tool

After placing the selected devices in the workspace, device IDs are automatically generated according to the order of the placement (e.g first device placed in the workspace will be give to 1, next device will be give 2 and so on). This device ID will not be used by the middleware as device middleware address will be uniquely generated by the middleware based on the role specification mechanism but these IDs will be used to visualise the smart spaces based on the test log data. Application developers may change the device name and device communication features as well as device functionalities such as coordinator and gateway and member.

In this example application where five devices are defined namely *GUIDESYSTEM, LOCATIONSYSTEM, VISITOR_HTC, TAXISYSTEM and TAXI*. The *GUIDESYSTEM* and the *TAXISYSTEM* devices are defined as the coordinators of the network (shown as red in Figure 3), the *LOCATIONSYSTEM and TAXI are* defined as gateway (shown as green in Figure 3) and *VISITOR_HTC* is defined as member devices (shown as blue in Figure 3).

By completing the device definition process, five Java projects with the device name (*GUIDESYSTEM,*

*LOCATIONSYSTEM, VISITOR_HTC, TAXISYSTEM and TAXI*) are generated for each device by the tool (as shown in Figure 4). These Java projects contain PECES middleware libraries (peces-2.0.jar) and necessary Java files under the *src* folder. These Java projects are configured with the additional PECES Nature which automatically generates Java files from context definition (*.peces.ctx ) file and SPARQL queries (*.peces.query ) file.

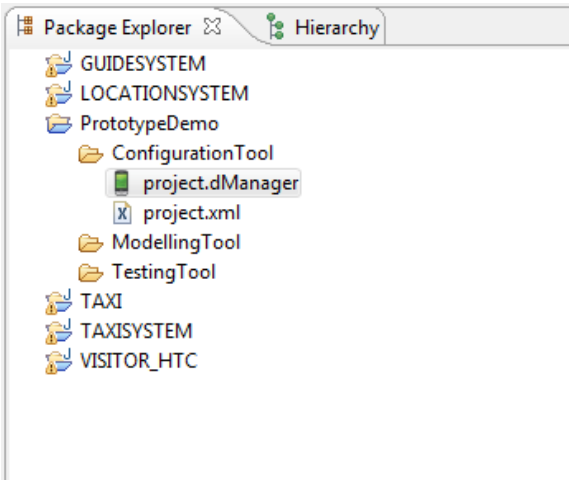



Figure 4: Screenshot of the Eclipse Package Explorer

### *C. Peces Ontology Instantiation Tool*

The PECES context ontologies are composed by the SmartSpace, Measurement, Device profile, User Profile and Event ontologies and these ontologies are available at the PECES project website to download [17]. The document [6] clearly explains the dependencies among them, as well as the external ontologies which provide a basis for the PECES concepts and properties. The core ontology for representing contextual information of a smart space is the *SmartSpace* ontology. The basic concepts to model the contextual information of a smart space are Device, Context, Location and Service.

The Device profile ontology provides vocabularies to model specification of devices inside smart spaces. There are three categories of devices defined in the PECES prototype application which are *PECESEmbeddedDevice*, *Accessory* and *SensorDevice*. *PECESEmbeddedDevice* represents those embedded devices that deploy the PECES middleware. There are three categories of embedded devices according to their role inside a smart space, namely gateway, coordinator and member. In order to specify which kind of accessories an embedded device has, the property *hasAccessory* can be used to link a *PECESEmbeddedDevice* instance to an Accessory instance. Accessory instances are Keyboard, Touch Screen, Speaker, Screen and Microphone. In addition to this, *SensorDevice* has two sub-concepts: *Detector* and *MeasuringSensor*. A *MeasuringSensor* instance represents a sensor which can measure a measurement such as light, noise, temperature, etc.

Peces Ontology Instantiation tool enables application developer to instantiate the devices. This tool supports all PECES ontologies (e.g., http://www.ict-peces.eu/ont/device.owl) as well as other custom ontologies (e.g., http://www.daml.org/services/owl-s/1.1/Service.owl) which application developers may wish to use for instantiation of the devices. The Ontology Instantiation tool automatically loads the participating device name and its functionality information from the project.xml (Figure 4) file which is generated by the Peces Device Definition tool. The Peces Ontology Instantiation tool provides GUI where application developers can add instances and link properties. When the instantiation process is completed, the tool creates a RDF file (project.owl) in the *PrototypeDemo* project *ConfigurationTool* folder and also creates *.peces.ctx files which contains the device context information for each devices. Those device *.peces.ctx files are placed in the appropriate java project and once the is placed in the Java project, the PECES Nature will automatically create necessary Java files for the middleware from the *.peces.ctx and *.peces.query files. The *.peces.ctx files are used to provide local context information of the devices. The *.peces.query files are used to provide context queries which enable the smart space coordinator to defined roles to the smart spaces.

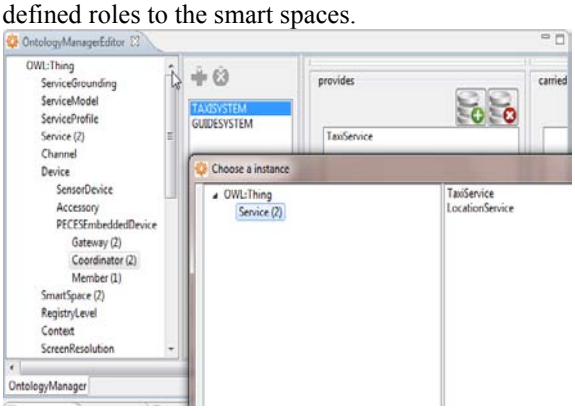



Figure 5: Screenshot of the Peces Ontology Instantiation Tool

Figure 5 shows that five devices defined in the Peces Device Definition tool are automatically loaded by the Peces Ontology Instantiation tool. Two new services (*TaxiService* and *LocationService*) are defined here with the tool. The *LocationService* linked (via provides) with *GUIDESYSTEM* device and *TaxiService* linked (via provides) with *TAXISYSTEM* device. Also *LOCATIONSYSTEM* and *VISTOR_HTC* are linked (via consumes) with *LocationService* and *TAXI* is linked (via consumes) with *TaxiService*.

### *D. Peces Security Configuration Tool*

The PECES middleware uses the OpenSSL library to create necessary certificates and keys. As a result, the Peces Security Configuration tool integrates the OpenSSL toolkit to enable application developers to generate keys and certificates for smart space application. The tool provides an interface to gather necessary information for root certificate, intermediate certificate (trust chain) and client certificate. The necessary information gathered from the Java interface is passed to

the OpenSSL command line interface with the use of script files.

Developers should first generate a root certificate (as shown in Figure 6) and then are able to generate necessary trust chain and client certificates (as in shown Figure 7). Once necessary certificate chains are created, they appear as trees in the Certificates area. To generate a Client certificate, first, developers must select the appropriate trust chain in the tree, and then click on the "Client. Cert" button to generate client certificate. The tool also provides a mechanism for a device to deploy certificates to other devices which are to be trusted.

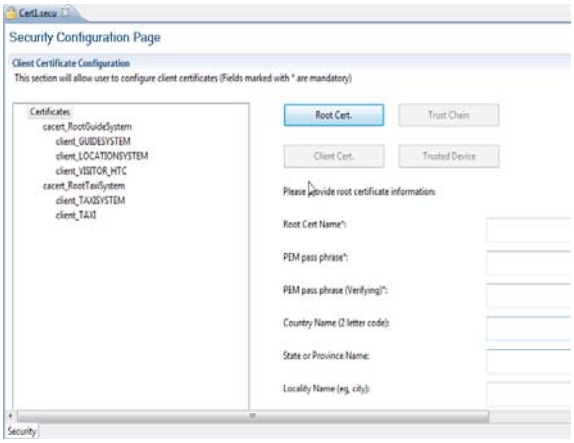



Figure 6: Screenshot of the Peces Security Configuration Tool – Root Certificate Creation

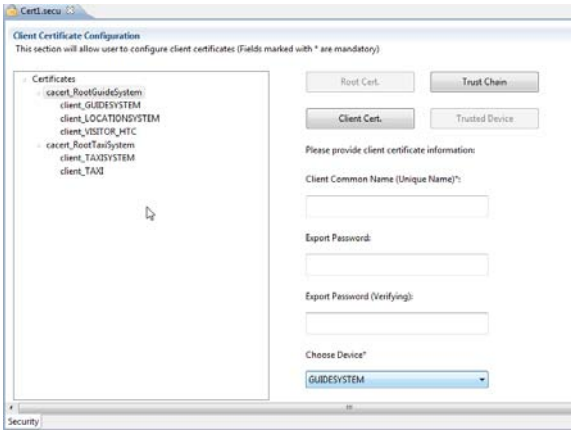



Figure 7: Screenshot of the Peces Security Configuration Tool – Client Certificate Creation

### *E. Peces Service Definition Tool*

The Peces Service Definition tool provides a simple interface to the developers that allows the automatic generation of necessary Java code needed to instantiate and make use of a PECES middleware based service.

The services defined (*LocationService*, *TaxiService*) in the Peces Ontology Instantiation tool is automatically loaded in the Peces Service Definition tool wizard page. The developer should select one service from the list of services to make use of the Peces Service Definition tool to further defined services. Once this process is completed, the Peces Service Definition tool main window will show several options to configure the service such as Device, Scope and Implemented Functions.

The "Device" option shows that which device will be implementing the selected service. The list of possible devices is shown to the developers, based on the definition in the Peces Ontology Instantiation tool. The Peces Service Definition tool automatically infers which the possible candidate is, but the developers can always change this default configuration.

The "Scope" option determines at which scope the service will be published. According to the PECES middleware Registry Interface Specification [8], the possible scopes are "Device" (available only to clients on the same device), "Space" (available to devices within the same smart space) and "Internet" (available to all smart spaces). The Peces Service Definition tool permits the developer to define the interface that the service will offer to its clients (i.e. the functions that will be available to them). These definitions follow a format that is similar to any Java function.

After completing this process, all necessary code is generated where needed. In particular, a Java file with the name of the service will be created in the project corresponding to the selected device. That file will contain an empty implementation of the service, which the developer will have to complete with the actual implementation.

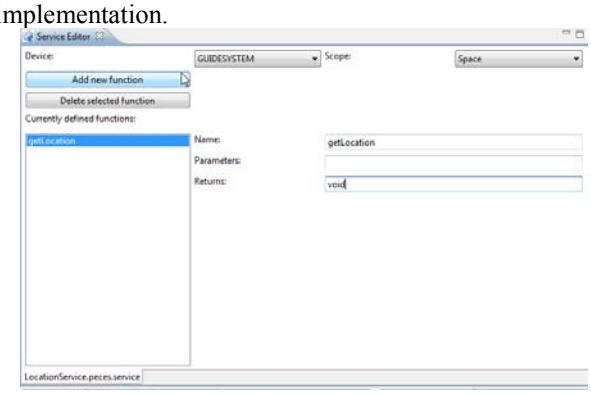



Figure 8: Screenshot of the Service Definition Tool

### *F. Peces Role Specification Definition Tool*

The Peces Role Specification tool provides an interface where developers can define the different rules that the application will use to dynamically form groups of collaborative devices.

In the PECES middleware application, rules are written essentially as constrained queries over the context properties of the devices. For that reason, the Peces Role Specification Definition tool loads the results of the Peces Ontology Instantiation tool, showing on a tree-shaped diagram all the devices that have been defined in the project, and their properties (upon which the rules will be defined).

Using the tool, a new Role Specification can be defined. Application developers should select a device for Role Specification. The selected device functionality should be a coordinator. For these reason, a combo box with all coordinators defined for the project is presented, where developers can choose the proper one. There are three Registry Interface Specification, "Device" (available only to clients on the same device), "Space" (available to devices within the same smart space) and

"Internet" (available to all smart spaces). Developers should select one of the three available scopes depending on their application. Next step, application developers should append *Rulesets* to the Role Specification. A *Ruleset* is a constrained query over the context properties of a device. A device fulfills a *Ruleset* when ALL the conditions defined there are fulfilled (AND conditions). On the other hand, a device fulfills a Role Specification when at least one of its *Rulesets* is fulfilled (OR conditions). Therefore, by combining several *Rulesets* in a single Role Specifications, reasonably complex conditions can be applied to the group formation process.

When a *Ruleset* is selected, its definition can be altered using the two tree-shaped property diagrams and the right hand window editor. By double clicking a property in the devices tree, a constraint over that property is added to the *Ruleset*. For instance, a constraint "?device provides ?service" would mean that "any device providing any service" fulfills the *Ruleset*. The window showing the defined constraints allow to change the third part of the constraints ("?service" in the example) by any possible value actually defined (possible values are shown in a combobox). For instance, a constraint "?device provides LocationService" would mean that "any device providing LocationService" fulfills the *Ruleset*. If the third part of a constraint is left undefined (i.e. beginning with an interrogation mark), it will appear in the variables tree. By appending a property of that variable to the *Ruleset*, composed constraints can be designed.

During the whole definition process, the bottom-left window with the title "Preliminary members of this smartspace" shows a prediction of the devices that fulfill the Role Specification, according to the static and initial context properties of all the devices defined within the project. By completing role specification process, the Peces Role Specification definition tool automatically generates all necessary Java code in the required projects to define and instantiate using the middleware.

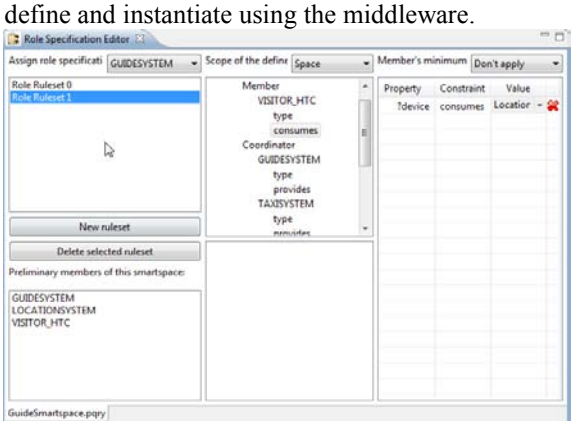



Figure 9: Screenshot of the Role Specification Definition Tool

### *G. Peces Hierarchical Role Specification Tool*

The Peces Hierarchical Role Specification tool (as shown in

Figure 10 ) provides an easy method to create all the code necessary to instantiate this kind of "composed" smart spaces. The tool presents a list with all the smart spaces included in the project.

Firstly, the application developers must choose the coordinator device which will in charge of the hierarchical smart space. A combo box listing all coordinators defined in the project is shown to the developers. Then application developers can choose which smart spaces will compose the hierarchical smart space. The developers can easily compose the list of "selected smart spaces" by using the appropriate buttons. By saving the Peces Hierarchical Role Specification, necessary files will be automatically created for smart space initializations and instantiations.

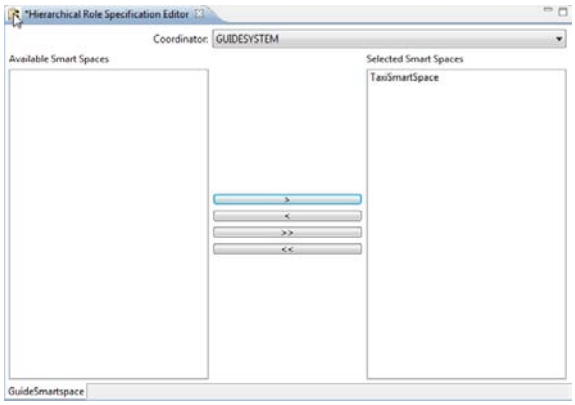



Figure 10: Screenshot of the Hierarchical Role Specification Definition Tool

### *H. Peces Event Editor Tool*

The Peces Event Editor tool is used to define single event definition. Five different events types can be defined which are Delay Events, Device SwitchON Event, Device SwitchOFF Event, Context Event and Connection Event.

The Event Editor is a multipage editor and in the second page (the Context Page), the context of the corresponding device can be changed if the event's type is Device Context Change. The third page (Connection Editor Page) can be used for connection related changes. The developers can only switch to this third page if the event's type is Connection Link Change. The developers can create several *.peces.event files necessary for their test.

Figure 11 shows a connection event which generated for the example devices defined in the Peces Device Definition tool.

Figure 11: Screenshot of the Event Editor Tool – Connection Page

### *I. Peces Event Diagram Tool*

When the developers have defined the needed events using the Peces Event Editor tool, the sequence of the events can be easily defined with the Event Diagram Editor. The Peces Event Diagram tool enables application developer to model smart spaces, dynamics connections and dynamic contexts for testing. The export of the events which can be done with the icon next to the "+" icon will generate an events.xml file with all the information needed by the Peces Testing tool.

Figure 12 shows Peces Event Diagram tool with the events generated for the example application described here.

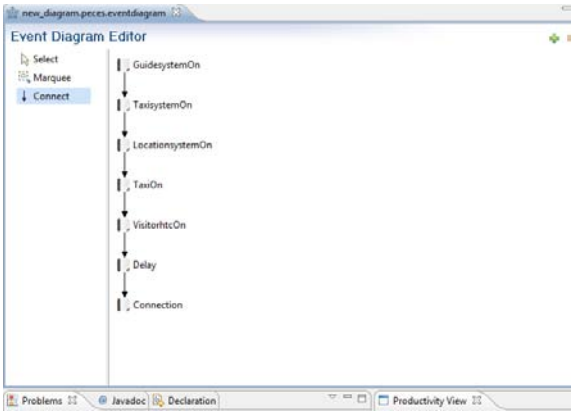



Figure 12: Screenshot of the Peces Event Diagram Tool

### *J. Peces Testing Tool*

The Peces Testing tool enables application developers to test the modeled application defined by the Peces Event Editor tool and Peces Event Diagram tool. For this purpose the Peces Testing tool generates another new Java project (emulator project) to centrally control other device related events. In addition to the necessary plugins, this new Java project should install the *EmulationTransceiver* plugin. This emulator Java project (run as new JVM process) is used to control connections and context dynamics (add and remove connections, add and remove context) between devices which is defined by the events.xml.

The Testing tool provides support for executing and analysing the smart space applications (each device application is considered as a separate JVM) built by the previous tools. The Testing tool loads and parses necessary device related information from the *project.xml* file and modelling information such as context changes connection changes from the *events.xml* file and also relevant device java project path information to execute the configured devices. The Testing tool wizard generates additional emulator Java project with the same name as the Testing tool editor to control the dynamic events of application. Application developers are able to define the required time to test the application specified by the previous tools. Developers can also provide Internet Registry IP and port information if they want to test that application with the internet registry. Figure 13 shows the defined application device status before the test was executed. The "Execute" button enables developers to run the application. As seen in above, the status of each device is shown during test (three devices are "ON" and one device is "OFF" at a particular time). All middleware and application related information is logged to a single log file with the specific device and absolute time of the system. Using this absolute time, the relevant time between the devices (based on test start time) are calculated for further analysis. The figure below shows the defined application device status while the test is running.

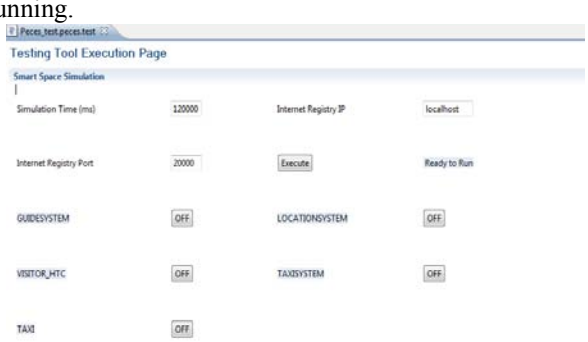



Figure 13: Screenshot of the Peces Testing Tool

The Peces Testing tool provides features to visualise the smart space network status based on the test log data events with relative time. The Peces Testing tool analyses important events occurred during the test from the test log data and displays as a "List of Events". The "List of Events" contain event information such Device Switch ON, Device Switch OFF, Connection, Disconnection and Smart Space Establish, Smart Space Join, etc. The status of the system can be viewed by double clicking on the name of the specific event. For example, by double clicking on the fourth event (*TaxiBooking Establish*) in the list in Figure 14 shows the visualisation of the smart space system at time 3050 ms (after test started). Figure 15 displays the two different smart spaces (*BoothNavigation and TaxiBooking* ) with its coordinator devices and it also displays the PECES Internet Registry availability but not connected with the smart spaces at this time event. The smart spaces are expected to form a hierarchical smart space defined by the Peces Role Specification tool.

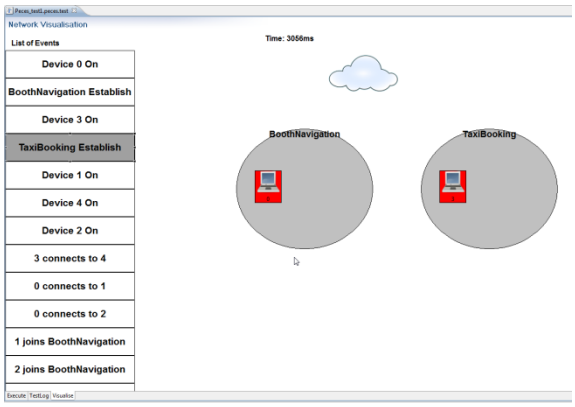



Figure 14: Screenshot of the Testing Tool Visualisation with two smart spaces

Figure 15 shows that *BoothNavigation* smart sapce and *TaxiBooking* smart space are formed a hierarchical smart space using the PECES Internet Registry when it is necessary. This clearly visualise the PECES middleware enables communication between devices in and across different smart spaces in a context dependent and secure manner.

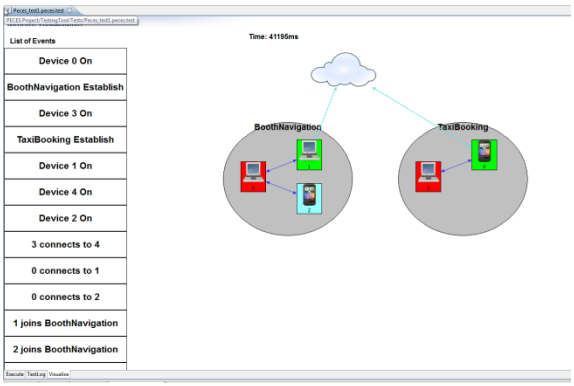



Figure 15: Screenshot of the Testing Tool Visualisation with a hierarchical smart space

## V. Evaluation

The PECES development tools have been evaluated by 20 evaluators from Germany, Spain and the UK. The evaluators were all Java programmers with varying degrees of experience ranging from Undergraduate Students thought to Post-Doctoral researchers and programmers from industry. Evaluators were given a short tutorial on PECES middleware and the development tools before testing the tools. They were asked to develop a simple service which required using the configuration tools (except for the Hierarchical Role Specification tool).

The evaluators completed a questionnaire and their development was measured by an Eclipse Productivity Plugin.

There were two main purposes of this evaluation. The first was to establish that user found the PECES development tools useful application development. Secondly, the tool developers wanted to obtain some useful suggestions as to how to improve the tools in the future.

### *A. Result Analysis*

Figure 16 shows how difficult it was found to develop PECES middleware application without the PECES Development Tools. The range is 1-5 where 1 indicated very difficult and 5 indicated very easy. Developing applications on PECES middleware is thought to be difficult without development tools.

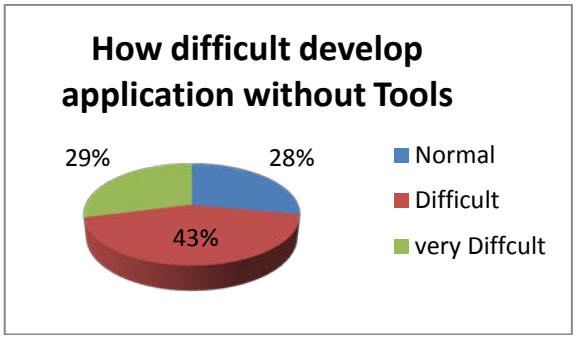



Figure 16: Difficulty of development without tools

The following figure indicates the responses to the question "what is the general impression you have for the PECES Development Tools". Range 1 indicated very impressive and 5 indicated very unimpressive. Over half delegates (55%) reported that the development tools are either impressive or very impressive. Only 25% people from the survey do not agree.

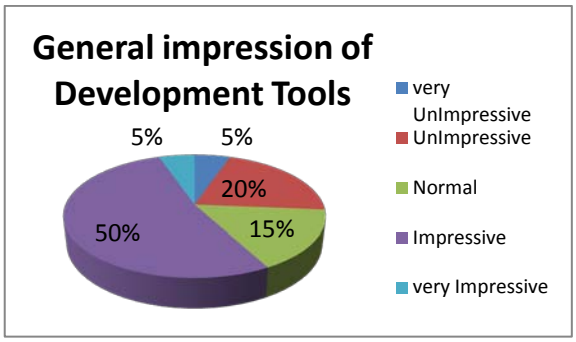



Figure 17: General impression of Development Tools

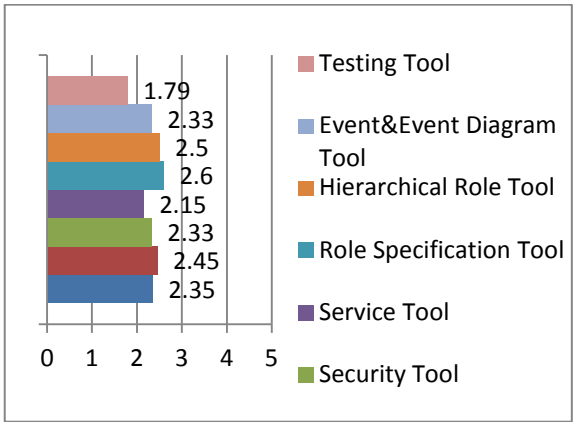



Figure 18: Mean of how easy to use development tools

Figure 18 compares the ease of use of the different tools used. Every tool that was used is shown in the charts. The Range of responses was 1-5 where 1 indicated very easy to use and 5 indicated very difficult to use. Some tools are not considered easy to use because they require extra background knowledge or concepts from the middleware. The Ontology Definition tool, Role Specification tool and Hierarchical Role Specification tool were in this category.

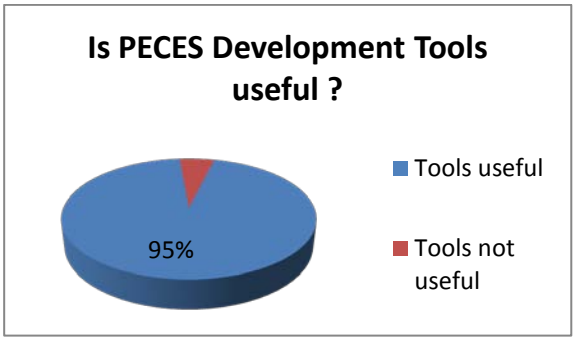



Figure 19: Are PECES Development Tools Useful

Figure 19 depicts the percentage of users who agree that the PECES development tools are useful. 95% delegates believe development tools really useful for developing PECES middleware based application. Only one person thought that the tool set is unsuitable for application development.

Finally, 55% of the evaluators believe a high level of training is needed whereas 40% think they just need a normal amount of training.

These results suggest that the PECES Development Tools are useful for experienced developers of PECES application rather than general users without any background knowledge of the middleware.

### *B. Productivity Plugin*

The aim of the productivity plug-in connected to Eclipse is to continuously track the different tasks developers work on, the time they individually require for carrying out those tasks, and the file operations needed to be performed within a given development environment. During the monitoring process the plug-in saves which file the developer edited and saved, when they started an application server, and which project or sub-project the user was involved in, and in which perspective.

The plugin measures the user interaction and hence the difficulty of use of each tool. The Role Specification tool was found to be the hardest to use, requiring almost twice as many interactions as the other tools.

## VI. Future Work and Conclusions

One of the main objectives of the PECES middleware is to provide a cooperation layer that enables seamless interaction and coordination among devices in and across smart spaces in a secure manner. This paper presented a set of tools which provide support for PECES middleware based application development. The tools provide support for device configuration, ontology instantiation, security configuration and role specification. The tools also enable dynamic modelling of the network connections and context changes. Finally, the tools provide support to test the smart space application performance and visualise the test results. The Evaluation of the PECES Development Tools suggests that tools are useful for smart space application development. Tools are presented here already integrated with the evaluator's comments.

There are still some room for improvements for the current version of the development tools. In this version, each device application is created as java project. There is no automatic support available to deploy the java applications on actually devices. We plan to improve the development tools and provide features to support deployment on different mobile platforms (e.g. Android).

PECES middleware supports building applications with multiple numbers of smart spaces. Even though all current the Peces Development Tools provide support for building application with multiple smart spaces except the Peces Testing Tool's Visualization Page only supports two smart spaces visualisation at the moment. We are currently looking at the different options to provide support for visualizing multiple smart spaces.

## Acknowledgements

The work presented here is sponsored by EC under FP7 programme (FP7-224342-ICT-2007-2) and authors also would like to thank all the project partners for **their** contributions.

## References

[1] PECES Project, http://www.ict-peces.eu, last accessed June 2012

[2] EMMA Project, http://www.emmaproject.eu, last accessed December 2010

[3] K. Selvarajah, C. Shooter, L. Liotti and A. Tully: *Heterogeneous Wireless Sensor Networks for Transportation Applications*, International Journal of Vehicular Technology (Special Issue on Vehicular Ad Hoc Networks), Feb 2011

[4] PECES Consortium, *PECES Use-Case Specification*, Deliverable D 1.2, PAS, http://www.ict-peces.eu, last accessed June 2012

[5] PECES Consortium, *PECES Requirements Specification*, Deliverable D 1.1, PAS, http://www.ict-peces.eu, last accessed June 2012

[6] PECES Consortium, *PECES Context Ontology and Query Specification*, Deliverable D 2.1, PAS, http://www.ict-peces.eu, last accessed June 2012

[7] PECES Consortium, *PECES Addressing Scheme Specification*, Deliverable D.3.1, PAS, http://www.ict-peces.eu, last accessed June 2012

[8] PECES Consortium, *PECES Communication Mechanisms and Registry Interface Specification,* Deliverable D.3.2, PAS, http://www.ict-peces.eu, last accessed June 2012

[9] C. Becker, G. Schiele, H. Gubbels, K. Rothermel: *BASE - A Micro-broker based Middleware For Pervasive Computing,* In Proceedings of the 1st IEEE International Conference on Pervasive Computing and Communications, pp. 443-451, Fort Worth, USA, March 2003

[10] Eclipse IDE, *Eclipse Website*, http://www.eclipse.org/, June 2012

[11] B. Lagesse, M. Kumar, J. M. Paluska and M. Wright, *DTT: A Distributed Trust Toolkit for Pervasive Systems*, http://www.ioc.ornl.gov/publications/lagesseDTT.pdf, last accessed June 2012

[12] M. Roman and R. H. Campbell, *A Middleware-based Application Framework for Active Space Applications,* Proceedings of the ACM/IFIP/USENIX International Conference on Middleware, 2003

[13] M. Roman and R. Campbell, *Gaia: Enabling Active Spaces, 9th ACM SIGOPS European Workshop*, pp.229-234, September 2000
[14] PECES Consortium, *PECES Secure Middleware Specification*, Deliverable D 4.1, PAS, http://www.ict-peces.eu, last accessed June 2012
[15] J. Barton and V. Vijayaraghavan, UBIWISE, A *Ubiquitous Wireless Infrastructure Simulation Environment,* http://www.hpl.hp.com/techreports/2002/HPL-2002-303.html, last accessed June 2012
[16] H. Nishikawa, S. Yamamoto, M. Tamai, K. Nishigaki, T. Kitani, N. Shibata, K. Yasumoto and M. Ito, UbiREAL: Realistic Smartspace Simulator for Systematic Testing, Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp2006), LNCS4206, pp. 459-476, Sep. 2006.
[17] PECES Ontologies, http://www.ict-peces.eu/ont/, last accessed June 2012
[18] Productivity plug-in version 3 http://www.ict-peces.eu, last accessed June 2012
[19] D. Garlan, D. Siewiorek, A. Smailagic, P. Steenkiste, "*Project Aura: Towards Distraction-Free Pervasive Computing*", IEEE Pervasive Computing, vol. 1, no. 2, pp. 22-31, April-June 2002.
[20] B. Johanson, A. Fox, and T. Winograd, "*The Interactive Workspaces Project: Experiences with Ubiquitous Computing Rooms*", IEEE Pervasive Computing, pp. 67-74, April-June, 2002
[21] Chun-Feng Liao, Ya-Wen Jong and Li-Chen Fu, *Toward Reliable Service Management in Message-Oriented Pervasive Systems*, IEEE Transactions on Services Computing, July-Sept. 2011, Volume: 4 Issue: 3, pp. 183 - 195.
[22] Protégé website: http://protege.stanford.edu/, last accessed June 2012
[23] R. Zhao, K. Selvarajah and N. A. Speirs, " *Development Tools for Pervasive Computing in Embedded Systems (PECES) Middle ware*", Proceedings of the International Conference on Wireless Information Networks and Systems (WINSYS), 2011

**Ran Zhao** is a doctoral student in Computing Science at Newcastle University. Before he started his Ph.D. study, he completed his M.Sc. in Computing Science from Newcastle University in 2007. Before that, he received his B.Eng. in Software Engineering from South China University of Technology, China.

**Kirusnapillai Selvarajah** completed his Ph.D. in Automatic Control and Systems Engineering at Sheffield University in August 2006. He obtained his B.Sc (Eng) in Electrical Engineering with 1st class honours from the University of Moratuwa, Sri Lanka in 2001. His research interests are Wireless Sensor Networks, Pervasive Computing, Embedded Systems, Swarm Intelligence and Optimisation. He worked as a Research Associate on the EU funded PECES and EMMA projects and currently working as a Research Fellow in the EPSRC funded MATCH project at The University of Nottingham.

**Neil Speirs** obtained a 1st class Honours degree in Mathematics from Newcastle University in 1980 and a doctorate in Theoretical Physics from the University of Durham in 1985. Since 1987, he has been at the University of Newcastle upon Tyne where he is currently a Senior Lecturer in Computing Science. His main research interests are in fault-tolerance, reliability and mobile distributed systems. He is the Newcastle University Project Manager on the EU PECES Project.

# Performance of OpenDPI in Identifying Sampled Network Traffic

Jawad Khalife and Amjad Hajjar
Lebanese University/Faculty of Engineering, IT department, Beirut, Lebanon
Email: jawad_khalife@hotmail.com, arhajjar@idm.net.lb

Jesús Díaz-Verdejo
University of Granada/Department of Signal Processing, Telematics and Communication, Granada, Spain
Email: jedv@ugr.es

***Abstract*—The identification of the nature of the traffic flowing through a TCP/IP network is a relevant target for traffic engineering and security related tasks. Despite the privacy concerns it arises, Deep Packet Inspection (DPI) is one of the most successful current techniques. Nevertheless, the performance of DPI is strongly limited by computational issues related to the huge amount of data it needs to handle, both in terms of number of packets and the length of the packets. One way to reduce the computational overhead with identification techniques is to sample the traffic being monitored. This paper addresses the sensitivity of OpenDPI, one of the most powerful freely available DPI systems, with sampled network traffic. Two sampling techniques are applied and compared: the per-packet payload sampling, and the per-flow packet sampling. Based on the obtained results, some conclusions are drawn to show how far DPI methods could be optimised through traffic sampling.**

***Index Terms*— network traffic identification, deep packet inspection, optimisation, payload truncation, flow truncation, traffic sampling**

## I. INTRODUCTION

Network traffic identification aims to classify packets (packet-based identification) or flows (flow-based identification) in a given network according to the associated application protocol. Traditionally, this task has been considered quite simple as ports were assigned for many application protocols. In this scenario a simple inspection of transport layer header suffices to identify the underlying protocol. Nevertheless, this situation is changing, thus making traffic identification a hot research topic, as some Internet applications, such as P2P, are becoming more and more challenging to identification techniques by using port obfuscation, encryption, and/or tunnelling [1]. One of the most successful methods currently available to identify traffic is based on the examination of the payloads to find known protocol patterns or signatures (e.g. "GET * HTTP"). This is the so-called DPI (Deep Packet Inspection) [2].

However, in today's networks, performance and privacy issues are two important factors that are considered some of the weaknesses of DPI. On the other hand, DPI is not able to inspect ciphered payloads. This fact is pushing researchers for alternate solutions in which P2P identification is still considered a complex task, especially when DPI is not involved at all.

As such, one of the current research trends is to optimise current DPI based identification methods characterised by their high accuracy, while keeping at the same time an acceptable level of user privacy and performance.

One of the DPI optimisation means is to reduce the input size through traffic sampling. Although different sampling policies exist [3], in this work, we applied sampling techniques at two different levels:

- Per-packet sampling: (or payload truncation) this is performed on the packet level, through partially inspecting the payload of each packet.
- Per-flow sampling: (or flow truncation) this is performed on the flow level, through inspecting the full payloads of a subset of packets per flow.

While sampling obviously provides a significant impact on the processing times by reducing the size of the input to process, it may have an unexpected impact on the traffic classification.

However, what impact the traffic sampling process would have on DPI classification accuracy, and which is the preferred sampling technique to use in optimising DPI, are important questions that we try to answer through this work.

In an attempt to answer these questions, we present in this paper, a study on the effect of traffic sampling on identification accuracy by using one of the best DPI-based tools: OpenDPI [4]. Our conducted identification experiments were based on full payload dataset traffic as captured through an institution's Internet link. We tested OpenDPI accuracy with per-packet sampling (using incremental payload truncation lengths) and with Per-flow sampling (using different number of sampled packets per flow), keeping three goals in mind:

- To provide protocol oriented results for classification accuracy.
- To compare the effect of both traffic sampling techniques on OpenDPI accuracy and their required input.
- To draw conclusions on how far combined DPI methods could be optimised through traffic sampling.

The remaining of this paper is organised as follows. Section 2 provides an overview of payload based


doi: 10.4304/jnw.8.1.71-81

identification tools, methods and optimisations. Section 3 describes OpenDPI tool in the way it analyses and labels packets and flows. Section 4 provides a description of the testbed we used for the experiments. Our conducted experiments and the obtained results in running the OpenDPI tool, both with per-packet and per-flow sampling techniques, are shown in Sections 5 and 6. Section 7 compares the obtained results. Finally, Section 8 presents some conclusions and future work.

## II. Identification Of Flows Based On Payloads

Deep Packet Inspection" (DPI) is defined in [2] as being "*…a computer networking term that refers to devices and technologies that inspect and take action based on the contents of the packet (commonly called the "payload") rather than just the packet header.*"

The most important parts of DPI are regular expression matching and signature based scanning. In this technique the payload of all the packets is checked against the set of known protocol signatures.

Some well-known DPI technology based tools are OpenDPI [4], an open source traffic classification tool, L7-filter [5], an open source application layer classifier for Linux's Netfilter, and Snort [6], an open source network intrusion prevention and detection system. In this paper, our choice was to use the OpenDPI tool since it includes the latest DPI technology combined with other techniques making it one of the most accurate classifiers.

Many authors attempted to enhance DPI accuracy by combining it with other methods, such as behavioural [7], statistical [8], port based [9] and DFI (Deep Flow Identification) based methods [10].

On the other hand, many recent works attempted to optimise DPI performance for high link speeds. Some of them apply software based optimisation focused on enhancing DPI algorithms, e.g. [11][12][13], while others use hardware based optimisation e.g. [14].

In this paper, we will focus on a software optimisation which consists on reducing the size of DPI input through partial payload inspection. In this context, different methods were proposed in the literature. For instance, ML (Machine Learning) identification methods [1] use the feature selection algorithm. On the other hand, sampling techniques are more general and easy to implement as they just try to reduce the size of the input data by simply taking samples or parts from the data according to a given criteria. This later approach could be jointly applied with DPI. In fact, this is the scenario considered in this work.

Sampling network traffic is the process of taking partial observations from the monitored traffic, and drawing conclusions about the behaviour of the system from these sampled observations. They are mainly used for network management and monitoring [15] although may also be used in classification tasks e.g. [9][16][17]. As many works [3][18][19] show, sampling techniques can be integrated within the traffic classification process.

Apparently, few works apply sampling to network traffic classification. A detailed taxonomy of sampling techniques according to the used method is provided in [15]. Another way of categorising sampling techniques is related to the target considered by the method. From this point of view, they can be classified as per-packet payload sampling [9][16], i.e. sampling bytes from within the packet payload, per-flow packet sampling [3][15][20][21], i.e. sampling a subset of packets from within the whole traffic flow, or a combination of both [17].

Per-packet sampling was shown in [16], where authors proposed a novel approach that brings the sampling idea to the regular expression field. Their approach, called payload sampling, allows skipping a large portion of the text in the payload, thus processing less bytes. Their results show that the sampling approach is faster than previous advanced solutions. However, the price to pay is a slight number of false alarms which require a confirmation stage.

Another example of per-packet byte sampling was shown in [9] which also combined the port-based method with the DPI approach. Using L7-Filter [5] DPI tool, one of the paper's targets was to study the amount of payload information actually relevant in successful DPI matches.

For each session, L7-Filter attempts to match its regular expression rules against the stream of payload every time a new packet is seen. Their experimental results showed that 72% of the total attempts happen at the first packet of a flow. Moreover, they computed the offset of matching regular expression's first character and last character from the beginning of the packets respectively containing them. They showed that almost all matching strings start (99.98%) and finish (90.77%) in the first 32 bytes of payload.

Per-flow sampling [3] for DPI classification is shown in many papers using different sampling techniques such as: sampled NetFlow [20], related sampling [21], Bloom filters [22][23], k-ary sketch [24], and mask-match sampling [25]. In [27], we studied the effect of per-flow sampling on DPI classification accuracy and showed that more than 90% of OpenDPI classification accuracy is maintained by sampling the first ten packets of each flow.

The combination of per-flow and per-packet sampling is addressed in [17]. In this work the authors combined both sampling methods through the so-called LW-DPI. Results showed that most flows can be classified with only their first 7 packets or a fraction of their payload.

Rather than presenting an exhaustive list of comparisons of existing per-packet or per-flow sampling policies, we preferred to compare at higher level, that is, by choosing one representative technique from each category for comparison purposes. The chosen techniques were designed to focus on sampling the first payload chunks: the first bytes of each packet's payload and the first packets of each flow. This is supposed to be an efficient yet distinguished sampling method yielding up to increased computational gain especially for large flows. In fact, most works in the literature used continuous sampling rates, which implies that the number of sampled packets will increase as long as the flow is under course, while it is fixed to a predefined number with the per-flow

sampling approach used in this work (as detailed in Section 6).

Comparison is based on two main criterions: the effect of sampling on the classification accuracy and the required input size. We consider both sampling techniques as eventual means of DPI optimisation as the size of input will be reduced by only inspecting the truncated part of the packet payload, with per-packet sampling, and a subset of packets per flow, with per-flow sampling.

## III. OPENDPI

As previously stated, the tool of choice for the classification of traffic is openDPI [4], which is derived from the commercial PACE product from Ipoque [26]. On 2009, Ipoque announced that it succeeded to win a test of deep packet inspection (DPI) bandwidth management solutions for monitoring and regulating peer-to-peer (P2P) traffic conducted by the European Advanced Networking Tester Centre (EANTC). Test results yield up to 99% detection and regulation accuracy for all popular P2P protocols.

The core of OpenDPI is a software library designed to classify internet traffic according to application protocols. In [4] the authors explain that OpenDPI incorporates different techniques such as behavioural (by searching for known behavioural patterns of an application in the monitored traffic) and statistical analysis (by calculating some statistical indicators that can be used to identify transmission types, as mean, median and variation of values used in behavioural analysis and the entropy of a flow).

Therefore, OpenDPI is not a pure-DPI product as it is not only signature-based but also incorporates information from other sources. This way, the classification accuracy is improved (no false classification according to Ipoque's claims), although some packets and flows still remain unclassified. This, together with the availability and quality of the signatures, made us to select OpenDPI instead of any other similar product.

In its current version, up to 101 different protocols can be identified, including the most common ones as SIP (Session initiation protocol), DNS (Domain Name Service), HTTP (Hypertext Transfer Protocol), HTTPS (Secure HTTP), FTP (File Transfer Protocol), and P2P protocols such as eDonkey, DirectConnect, Bittorrent etc.

Nevertheless, and according to its functioning, the capabilities of OpenDPI are mainly limited by the need to analyse the whole payload of all the packets in a flow in search of signatures (DPI behaviour) and to extract the behavioural and statistical information from the flows. Therefore, it is a basically full payload / full flow analysis which imply a high computational cost. This way, it would be desirable to reduce the size of the explored data in order to reduce this computational cost, but without degrading the performance of the classifier.

In this context, the target of this paper can be stated as analysing how sensitive are the mechanisms involved in OpenDPI to the per-packet and per-flow sampling techniques.

## IV. TESTBED

In order to evaluate the effects of truncating the payloads in the traffic identification task, we have developed an experimental setup built from two main components. These components are a database of real traffic captured in an academic network, and a tool to automatically classify packets and flows according to their payloads by primarily using Deep Packet Inspection (DPI) which is based in OpenDPI.

Therefore, we have built a tool based on the OpenDPI library which is able not only to identify the application protocols but also to follow and differentiate the packets in each flow. To be able to handle UDP packets, we have generalized the concept of flow through the use of sessions. Sessions are considered as defined by the exchange of information associated to a tuple (IP addresses, ports and transport protocol) [4]. Nevertheless, throughout this paper, we will use the term flow to refer to a session, unless explicitly stated. It was convenient not only to apply sampling techniques on TCP sessions, but on UDP sessions as well. In fact, experiments show that application signatures are detected within UDP flows and that the classification accuracy is affected accordingly. Similarly to TCP based applications, classification results for UDP based applications (such as DNS and SIP) are protocol dependant, as shown in section 6.

As the output of the tool, two levels of classifications are provided: flow-based (each flow is labelled) and packet-based (each packet is also labelled). The tool operates in batch mode.

On the other hand, the traffic database contains the data captured during 3 working days at the access link of a medium size institution. The network consists of one head office to which more than 50 branches are connected according to a star topology. Local application services such as Email, Web and DNS are hosted in the head office which aggregates and controls all traffic flows, generated by different branches, through one central firewall. Through this network, around 4000 concurrent users are usually connected and generating approximately 40000 concurrent flows.

The data acquisition was carried out at a border router in the head office in order to be able to monitor all incoming and outgoing traffic. Therefore, apart from the boundaries of the caption, flows are captured complete and in both directions. Table I highlights some figures of the database.

By using the customized OpenDPI tool over the whole database we have built the "ground truth", that is, the set of correctly labelled flows and packets that will be used as the reference when evaluating traffic sampling techniques, in the following sections, where the evaluation of the identification is measured in terms of accuracy [1], that is, the percentage of detected packets/flows in regard to the full payload case. This procedure is adopted under the assumption that DPI is the

best currently available method for traffic classification and that the number of errors is negligible. This is a common approach in the traffic identification field, the number of packets and flows that DPI is not able to classify being its major limitation. In fact, some flows are not classified by OpenDPI (labelled as *unknown*), even when inspecting complete flows with full packet payloads, i.e. without sampling. Evidently, when sampling techniques are applied, these flows will remain unclassified by OpenDPI. Nevertheless, when sampling is applied, some of the flows, classified at the ground truth, become unclassified. Consequently, in order to highlight the effect of sampling on classification accuracy, unknown flows at the ground truth level are not counted when evaluating the flow accuracy under sampling.

The results provided by the classification tool show up to 42 protocols, including known web protocols such as HTTP and FTP, voice over IP protocols, such as SIP, and P2P applications such as Bittorrent, etc. Most of these protocols have been identified in the database, being HTTP the most frequent one, while an important part of the flows and packets remain unclassified. The relative distributions of flows and packets for most relevant protocols are shown in Fig. 1. A first inspection evidences big differences among the properties or frequencies at flow and packet levels. Therefore, the results can be different depending on whether we focus at flow or packet levels.

In this work, we will evaluate sampling techniques and their accuracy from the point of view of flow classification, not individual packets. This is because classifying flows is semantically more significant and more adequate to most traffic engineering tasks. Furthermore, flow classification is more efficient as all the packets in a flow will be classified including even those that do not contain any application-specific signatures or patterns.

## V. Truncation Of The Payloads

In this section, we will show the conducted experiments in running OpenDPI on partially truncated packet payloads using the per-packet sampling technique.

Our main targets at this level are, as mentioned in Section 1: *To provide protocol oriented results for accuracy as a function of the sampled input (truncation length), and to show to what extent could the payload truncation affect OpenDPI accuracy.*

Through packet truncation, we intend to partially inspect each packet's payload. The studied sampling technique is very simple: the classifier must parse only a specified length of bytes (called payload truncation length or S) within each packet's payload. This is supposed to decrease the global classification time for the whole traffic.

As such, the per-packet sampling scheme we used is defined as: "*First S Bytes per packet*"

### A. Methodology for Truncation Experiments

To achieve our targets, we customized OpenDPI tool to be able to parse only a specified length of bytes within each packet's payload. In order to obtain granular results, our choice was to iterate with incremental truncation length values with step of D Bytes, ranging from 0 Bytes (no payload) to 1500 Bytes (full payload). We have chosen D=128 Bytes.

TABLE I.
FIGURES FOR THE CAPTURED TRAFFIC DATABASE.

| | |
|---|---|
| Size of the database | ~180 GB |
| Number of IP packet | 278 Mpackets |
| Number of different IPs | 822519 |
| Number of flows | 6.3 Mflows |
| Number of identified protocols | 42 |

The complete dataset we captured is very huge (63 pcap files totalling 177 GB) to use for the classification experiments, joint with sampling. In fact, we needed to run the customized OpenDPI up to 15 times on the same set of capture files. Since the customized OpenDPI requires around 50 minutes in classifying 1GB of capture data, the tool was run on a subset of only 17 randomly selected files (totalling 45 GB, i.e. 25% of the number of pcap files) due to time constraint.

On the other hand, those packets and flows that were not classified by OpenDPI when using the whole payload are dismissed and not considered in the figures and percentages that will be shown.

For this section, accuracy results are shown as a function of the truncation lengths and grouped according to three different sets: per protocol, per protocol group, and for all the protocols. For this purpose, the protocols were categorised into 12 groups that were defined according to [26].

### B. Global results

The results obtained for all the protocols are shown in Fig. 2, where we show the number of successfully classified packet (in red) and flows (in blue) as a function of the length of the sample from each packet payload. At packet level, a sudden drop in the accuracy for truncation lengths lower than 1408 Bytes is observed. For 1280 Bytes, 47% of the packets were correctly classified, while for 1408 Bytes, 99% of all the packets were identified. On the other hand, the results at flow level show that for truncation length equal to 512 Bytes, 57% of total flows

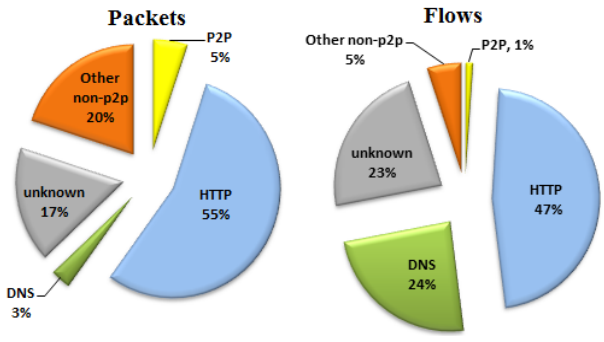



Figure 1. Distribution of packets (left) and flows (right) for most relevant protocols or groups of protocols.

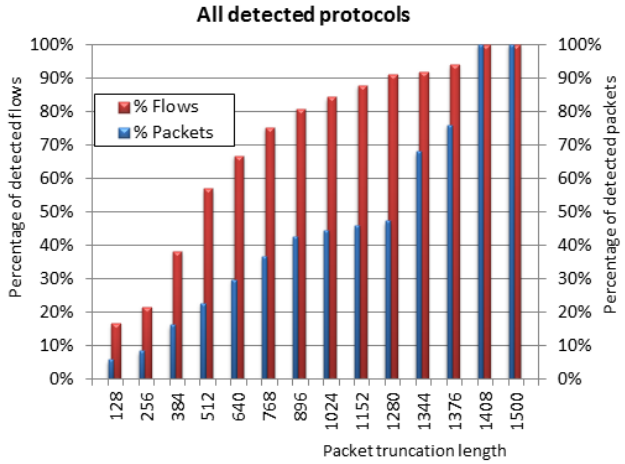



Figure 2. Global results for classification accuracy as a function of the truncation length of the payloads.

were detected, while for 1280 Bytes, 91% of flows were detected. Therefore, the analysis is more tolerant to payload truncation at flow levels than at packet levels.

Thus, truncation length must be at least 1280 Bytes to reach 50% of both flow and packet accuracy. This is not a very encouraging result for DPI optimisation through payload truncation as reducing only 15% of payload input would lead to a 50% drop in OpenDPI packet accuracy. However, results are encouraging if only flow accuracy is the main concern since still 57% of flows can be detected for 512 bytes of truncation.

From a macroscopic point of view, OpenDPI showed a common behaviour for all protocols:

- The number of detected packets/flows is increasing as the truncation length increases.
- For truncation length equal to 512 Bytes, 57% of flows were detected while only 22% of packets were detected.

### *C. Results per protocol group*

When varying the truncation length, OpenDPI shows different behaviour for different protocol groups.

As an example, results for web group packets and flows are shown in Fig. 3.a. Web group results show that truncation, though differently, is affecting both packet and flow accuracy. In addition, web packet accuracy seems to be more affected by truncation than flow accuracy. It's noticeable that packet classification accuracy drops to around 50% for 1280 Bytes while for flow accuracy it drops to 50% only if less than 512 Bytes are truncated.

A different behaviour is observed for other groups. For example, if we consider the IM (Internet Messaging protocols) group –Fig. 3.b– or DNS group –Fig. 3.c– the classification accuracy is only slightly affected by truncation. In fact, for a truncation length equal to 256 Bytes, more than 50% of both packets and flows are detected. The same applies for DNS packets and flows.

The results for P2P protocols exhibit a mixed behaviour –Fig. 3.d– as they are similar to those from the web group at packet level and to those from IM and DNS groups at flow level. In fact, packet accuracy drops to around 50% for 1280 Bytes while flow accuracy stays above 92% even for 128 Bytes only.

In summary, at a granular level, the experimental results showed different behaviour for OpenDPI with truncation for different protocols. This in fact could be based on two main factors: the stateful behaviour of some protocols combined with the detection algorithm used by OpenDPI which considers some behavioural and statistical information for the whole flow.

We can evidence this assertion if we examine the obtained results for the web and DNS protocol groups. Since DNS is a stateless protocol, flows with truncated packets can still be detected. On the other hand, as web is a stateful protocol, the detection of web flows drops for truncated packets. Though not shown, FTP results also were different since FTP protocol has a special behaviour.

Therefore, we can conclude that stateless protocols are less sensitive to payload truncation than stateful ones. Thus, optimising DPI/DFI methods through payload truncation could be more effective for stateless and P2P protocols.

For interpreting the differences between flow and packet results for the same protocol, flow results are considered more significant since undetected flows may contain a huge number of packets thus affecting packet accuracy. We also noticed that flows detected at higher truncation length mostly contain a huge number of packets.

As a result for per-packet sampling, studied in this section, unless just a few bytes (not more than 128 Bytes) were omitted from the end of the packet payload, payload truncation with combined DPI/DFI will lead to many unknown flows and packets. For instance, by inspecting the full packet payload and omitting the last 512 bytes, only 57% of flow accuracy can be maintained.

## VI. Truncation of The Flows

In this section, we will show the conducted experiments in running OpenDPI on sampled flows using the per-flow sampling technique.

Our main targets at this level are, as mentioned in Section: *To provide protocol oriented results for accuracy as a function of the sampled input (number of inspected packets per flow) and to show to what extent could the flow truncation affect OpenDPI accuracy.*

For comparison purposes with per-packet sampling, we conducted per-flow sampling experiments to obtain results for the same protocol groups, shown in Figures 2 and 3 of the previous section.

The methodology we used for flow truncation is described in [27], where we intend only to inspect, within each flow, the packets whose ordinal number inside the flow is lower than a predefined threshold ($N_{min}$). With this sampling scheme, while inspecting only the first $N_{min}$ packets of the flow for the purpose of classification, the classifier will still handle the remaining packets for the purpose of assigning them to the flow. The difference is that for these packets the inspection part is to be omitted, and this is where the concept of optimisation comes: Through flow-sampling, we emphasize on the inspection time as we consider it to be the only sensitive term to flow truncation. In other words, the sampling speed-up in terms of CPU processing time comes only from speeding-up the flow classification itself, but there is no gain in the operation of mapping packets to flows, as this operation is independent and untouched.

As such, the per-flow sampling scheme we used is defined as: "*$N_{min}$ packets per flow*"

### A. Methodology for Truncation Experiments

In this Section, we used the same OpenDPI customization we performed in [27], on which we run on a subset of randomly selected files from our original dataset, since the main dataset is very huge. This customization allowed us to output the packet ordinal number inside the flow the packet belongs to at which

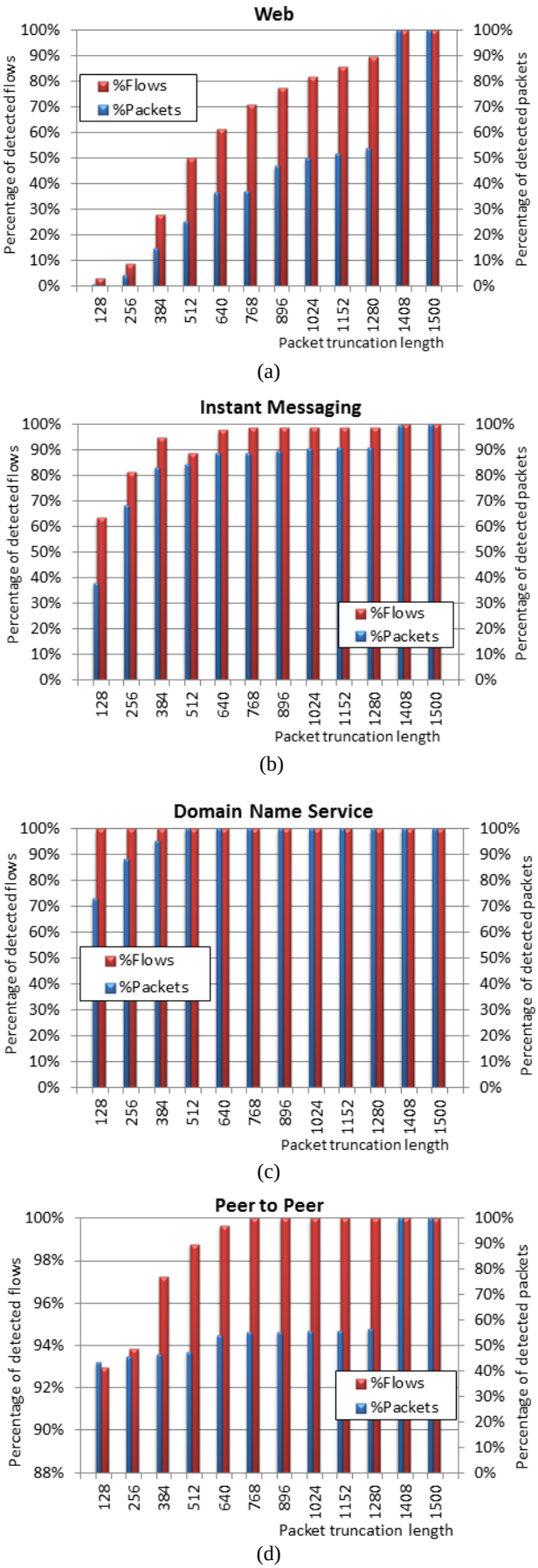



Figure 3. Results for various protocols/groups as a function of the truncation length for packets and flows. a) Web; b) Instant messaging; c) DNS; and d) P2P.

detection is achieved, referred to as *packet detection number* or *flow detection number*. In addition, we were able to generate accuracy results for different numbers of sampled packets per flow ($N_{min}$), without effectively truncating the flows.

As in the previous section, flow accuracy results are shown in terms of number of successfully classified flows as a function of the number of sampled packets from the beginning of the flow. Again, these results are grouped according to three different sets: per protocol, per protocol group, and for all the protocols.

### B. Global Results

Fig. 4 shows the percentage of flows that have been classified vs. the number of sampled packets. As we can see, most flows are detected by inspecting the few first packets. Specifically, within the first ten packets ($N_{min}$=10), most protocols are being detected with an accuracy value of 99%. As also depicted in Fig. 4, flow accuracy is near 90% for $N_{min}$=4. Only a slight increase in accuracy is obtained for $N_{min}$ values greater than 10. For these reasons, $N_{min}$=4 or $N_{min}$=10 could be considered critical values for the per-flow sampling scheme, according to the required level of accuracy and the required level of classification speed-up.

### C. Results per Protocol Group

Results for the same protocol groups tested in the previous section are now shown in Fig. 5. The DNS group in Fig. 5.c seems to be the less sensitive protocol to flow truncation, as it's being classified by OpenDPI by inspecting solely the first packet with a 99% of classification accuracy. Other protocol groups are shown as well, like Web in Fig. 5.a, and Instant Messaging in Fig. 5.b. The same result as seen globally for all protocols persists: at least four packets are required to be inspected to reach an accuracy level of 90% and above. Results for P2P are shown in Fig. 5.d where 84% of accuracy can still be reached for $N_{min}$ =4. If the classifier inspects the first ten packets of a P2P flow, 99.15% of classification accuracy can be reached as well.

### D. Results per Protocol

The average packet detection number in the dataset is

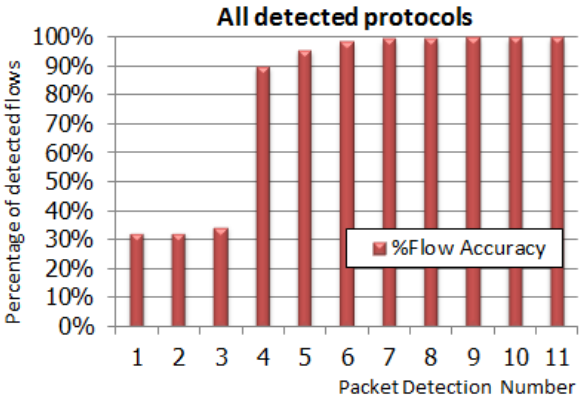
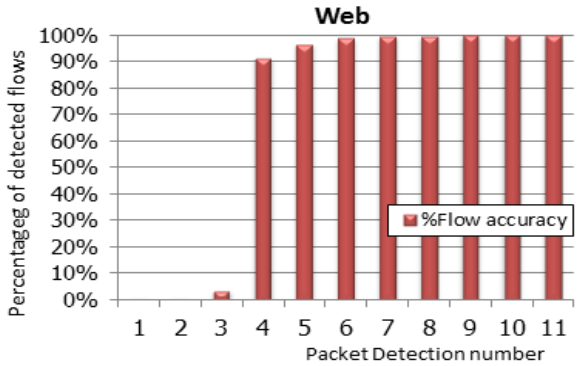
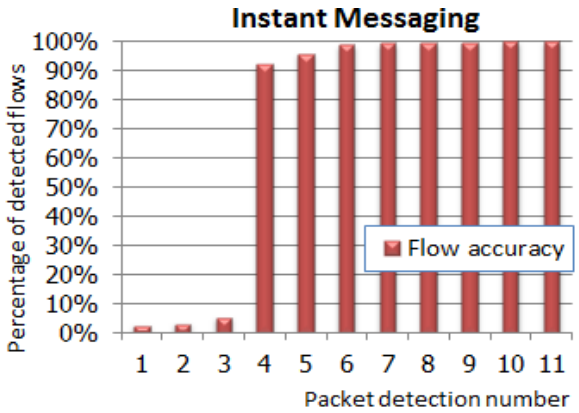


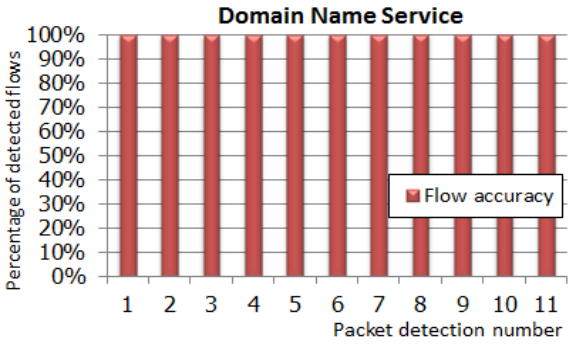

Figure 4. Global results for flow accuracy as a function of the packet detection number.



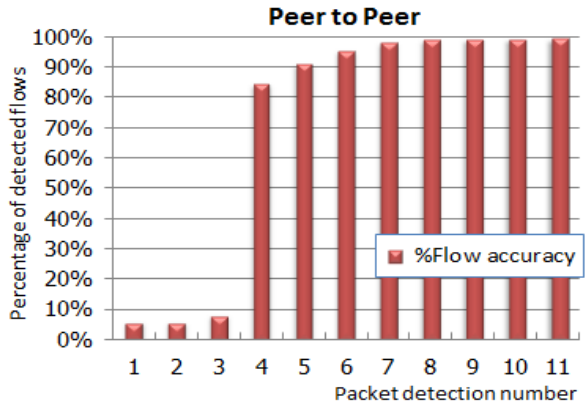

(a)




(b)




(c)




(d)

Figure 5. Flow accuracy results for various protocols/groups as a function of the packet detection number. a) Web; b) Instant messaging; c) DNS; and d) P2P.

shown in Fig. 6 for most common protocols.

Some protocols like iMESH and Bittorrent, show higher values than other protocols. We validated the fact that the presence of most deviation is due to flows that were under course during the start of the capture. Most protocols averages were below 10 packets. To validate this fact, we generated results per individual protocol. For instance, Fig. 7 shows the histogram of flow accuracy for some selected protocols like SIP (Fig. 7.a), FTP (Fig. 7.b), and HTTPS (Fig. 7.c). It can be noticed that about 90% of flow accuracy is reached by inspecting the first 6 packets.

As a result for per-flow sampling, studied in this section, inspecting the first 4 to 10 packets of a flow (as DPI input for inspection) could maintain the flow classification accuracy at high levels ranging from 90% to 99%.

In choosing the appropriate value of $N_{min}$ for the classifier, two situations should be distinguished according to the classification target:

If the target is to classify only one specific protocol, $N_{min}$ could be easily specified according to Fig. 6 (e.g. for HTTP, $N_{min}$=4). In this case, the classifier would inspect only the minimum number of packets, necessary for flow classification. However, if the target is to classify all protocols, which is the most common situation, $N_{min}$ should be assigned the maximum value of the average packet detection number ($N_{min}$=10) in order to classify most protocols. In this case, and for protocols whose average packet detection number is lower than $N_{min}$, the classifier would inspect more packets than necessary.

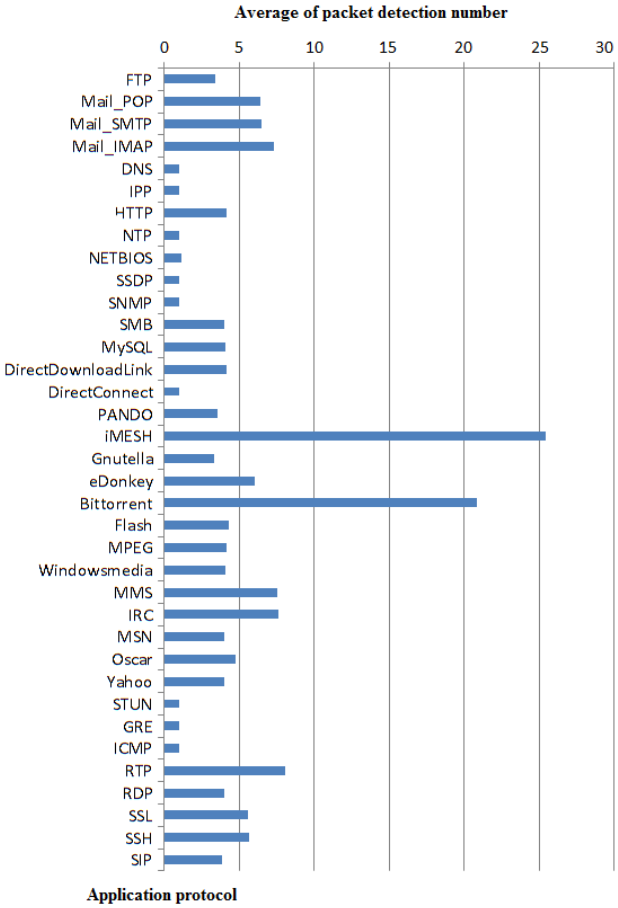



Figure 6. Average detection packet number for each individual protocol.

### E. Computational Measurement

To highlight the optimisation aspect of sampling approaches, we choose to measure the computational gain in processing time for the per-flow sampling technique. Specifically, we measure the processing time consumed by the classification modules inside the classifier's code. As mentioned previously, through the flow sampling process, only the inspection time is optimised and not the packet handling time.

Experiments show that compared to full flow sampling, the per-flow sampling approach can provide 9% of computational time gain and 99% of classification accuracy, when only the first 20 packets ($N_{min}$=20) are inspected.

In comparing to EIM (Equidistant Invariable Mode) [3] having a sampling rate of 7/13, 36% of classification time can be saved due to inspecting less packets with the per-flow sampling approach (for $N_{min}$=20).

## VII. Results Comparison And Analysis

### A. Results Comparison

Table II shows the comparison results as summarized for both sampling techniques, according to the provided flow classification accuracy and the required DPI input. The percentage of input reduction is not shown in this table since it is dataset-dependent and can be simply estimated according to the average number of packets per flow, and the average packet size.

Results shown in Table II indicate that in order to obtain around 90% of flow accuracy, it's mandatory to inspect the first 1280 payload bytes of each packet (as DPI input for inspection), while inspecting the first 4 packets with full payload per flow is sufficient to obtain the same accuracy level. For higher accuracy results at 99%, the first 1408 of payload bytes are required per packet compared to the first 10 packets with full payload per flow.

For the dataset we used, the average number of packets

TABLE II.
Comparison between per-flow and per-packet sampling schemes used for DPI optimisation

| Sampling Scheme used for DPI Optimisation | Required Input | Flow classification accuracy |
|---|---|---|
| Per-packet payload sampling | First 1280 bytes of payload, per packet | 91% |
| | First 1408 bytes of payload, per packet | 99% |
| Per-flow packet sampling | First 4 packets with full payload, per flow | 90 % |
| | First 10 packets with full payload, per flow | 99% |

per flow is 45 packets and the average packet size is 233 Bytes. Thus, in order to obtain at least 90% of flow accuracy, 932 Bytes will be required for inspection with the per-flow sampling compared to 10,485 Bytes with the per-packet sampling.

Thus, compared to per-packet sampling, the per-flow sampling technique will provide higher flow classification accuracy at a cost of less input. Apparently, this applies only to traffic having, on average, more than 4 packets per flow and less than 1400 Bytes per packet, which is common for traffic of most known protocols, as shown in the dataset we used.

Moreover, per-packet sampling leads to many unknown packets, and the packet accuracy drops to 47% when the packet truncation length is 1280 bytes, as depicted in Fig. 2. Thus, optimising DPI/DFI methods through payload truncation could not be considered generally effective (especially when the set includes stateful protocols, which are the more affected).

Obviously, the best trade-off between the required DPI input size and the provided classification accuracy in Table II is with the per-flow sampling when only the first 4 packets with full payload are inspected per flow.

As a result, according to the sampling schemes we defined and regardless of the traffic dataset being tested, the following result can be generalized:

*For DPI optimisation, the per–flow sampling technique is more convenient than the per-packet sampling technique, in terms of the required input and the provided classification accuracy.*

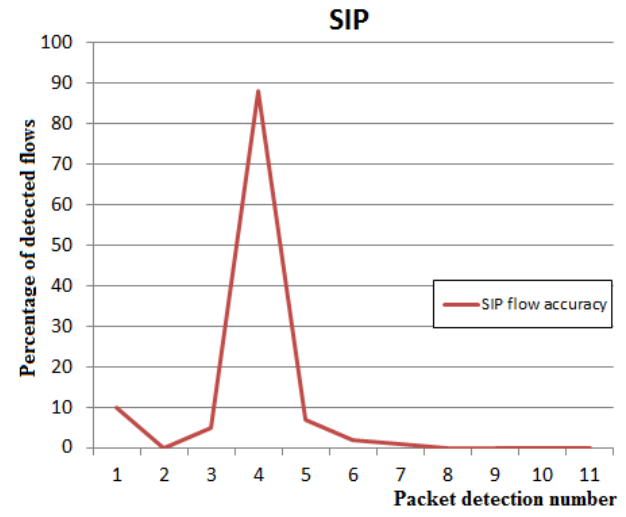
### B. Results Analysis

One interpretation for the obtained result is that by combining DPI with other technologies (such as behavioural and statistical modelling), the task of DPI optimisation through per-packet sampling or payload truncation may render the identification method itself inefficient since the non-parsed part of the data may still be needed for the other added technology. However, to validate this assertion, further experiments are required, in which the pure DPI technique is to be separated from other helper technologies. The per-packet payload truncation can still be useful as an optimisation if, instead of classifying all the traffic, the target is to select some of them based on the application content and depending on the nature of the associated protocol.

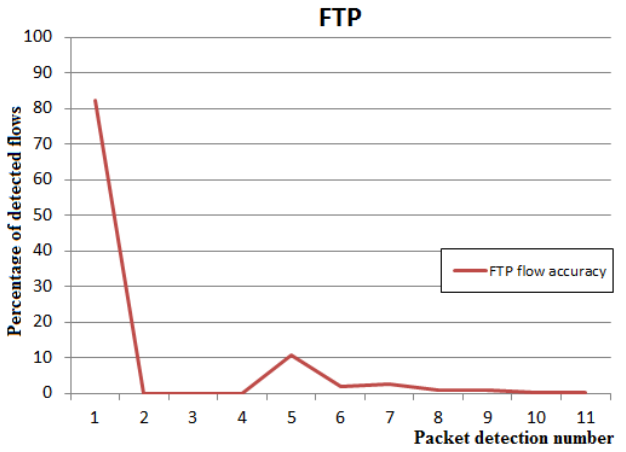
With the per-flow sampling, better results are obtained only if the first packets were sampled. This can be interpreted by the fact that the first packets usually signal the application protocol in use, during the first phase of flow set-up, and are thus of high importance for the classification process.

On the other hand, the default behaviour for PBFS (Packet Based Per Flow State) classifiers [28] is that the entire session must be “marked” as soon as one packet is detected to be holding an application signature.”

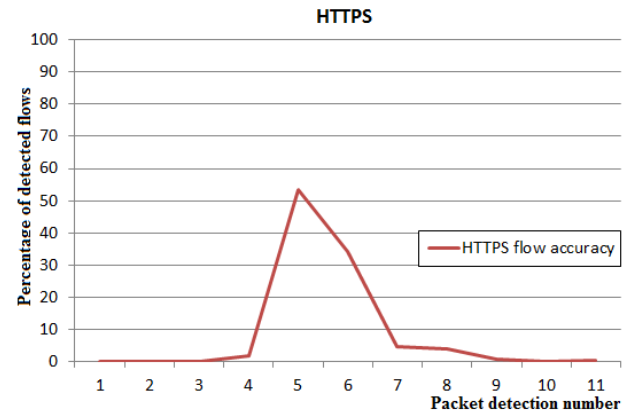
However, not all DPI classifiers are PBFS based, and for some particular cases, not all flows belonging to the same application protocol should be necessarily detected at the same packet detection number. For instance, when




(a)




(b)




(c)

Figure 7. Flow accuracy results for two selected protocols as a function of the packet detection number. a) SIP; b) FTP and c) HTTPS.

the packet holding the signature is delayed or lost, more packets will be inspected until the flow protocol is detected at higher packet numbers or simply marked as unknown. As for the OpenDPI tool, we were able to prove its PBFS like behaviour both theoretically, by interpreting the classifier’s code in Section 6.A, and

practically, by interpreting results for the average packet detection number (being nearly the same per protocol, in Fig. 6) and for the classification time gain (being moderate, in Section 6.E).

## VIII. Conclusions and Future Work

In this paper we explored the effects of sampling on traffic classification accuracy using OpenDPI. Traffic sampling is considered one of the means of DPI optimisation as it reduces the required input size for the classifier. Two sampling techniques were tested and compared: per-packet sampling (through partial payload inspection) and per-flow sampling (through inspecting only a few packets per flow). Comparison is accomplished according to the reduction in the input size and the maintained classification accuracy.

Results show that flow accuracy is less sensitive to flow truncation than to packet payload truncation. With per-packet payload sampling, unless just few bytes (not more than the last 128 Bytes) were omitted during the packet payload inspection, per-packet sampling (or payload truncation) will lead to many unknown packets and flows. With per-flow packet sampling, inspecting the first 4 to 10 packets per flow could maintain flow accuracy at higher levels, ranging from 90% to 99%.

As a result, the per–flow sampling technique is more convenient than the per-packet sampling technique for DPI optimisation.

To provide more richness to this work, future enhancements may include comparing existing sampling methods within the same category (per-flow and per-packet), in which computational gain is evaluated in terms of both processing time and memory usage. Enhancements may involve as well other advanced DPI tools to discard any possible bias to OpenDPI, our best tool of choice.

Finally, in seeking for DPI optimisation through sampling, future enhancements should involve additional experiments which should integrate different techniques in order to define the optimal sampling scheme for the DPI classification process. Once the required input for the DPI classifier is optimally reduced, other DPI optimisation means, as found in the literature, could be used jointly to complete the job of DPI optimisation.

## Acknowledgment

This work has been supported by Spanish MICINN under project TEC2008-06663-C03-02.

## References

[1] T. Nguyen, G. Armitage, “A Survey of Techniques for Internet Traffic Classification using Machine Learning”, IEEE Communications Surveys & Tutorials, 2007, vol. 10, pp. 56-76, doi: 10.1109/SURV.2008.080406.

[2] Allot Communications, 2007. Digging Deeper Into Deep Packet Inspection (DPI). White paper. Available at http://www.dpacket.org 20.01.2012

[3] H. Chen, F. You, X. Zhou, and C. Wang, "The study of DPI identification technology based on sampling", ICIECS 2009, 2009, pp. 1-4, doi: 10.1109/ICIECS.2009.5363202.

[4] Opendpi. http://www.opendpi.org/ 20.01.2012

[5] L7filter. http://l7-filter.clearfoundation.com/ 20.01.2012

[6] Snort. http://www.snort.org 20.01.2012

[7] L. Zhang, D. Li, J. Shi and J. Wang, “P2P-based Weighted Behavioral Characteristics Of Deep Packet Inspection Algorithm, In Proc. of CMCE 2010, 2010, pp. 468-470, doi: 10.1109/CMCE.2010.5610457.

[8] F. Dehghani, N. Movahhedinia, M. Khayyambashi, and S. Kianian, “Real-time Traffic Classification Based on Statistical and Payload Content Features”, In Proc. IWISA 2010, 2010. pp. 1-4, doi: 10.1109/IWISA.2010.5473467.

[9] G. Aceto, A. Dainotti, W. de Donato, and A. Pescapé, "PortLoad: taking the best of two worlds in traffic classification”, In Proc. of INFOCOM 2010, 2010, pp. 1-5, doi: 10.1109/INFCOMW.2010.5466645.

[10] C. Wang, X. Zhou, F. You, and H. Chen, "Design of P2P Traffic Identification Based on DPI and DFI”, In Proc. of CNMT2009, 2009, pp. 1-4, doi: 10.1109/CNMT.2009.5374577.

[11] Y. Yang, H. Le, and V. Prasanna, “High Performance Dictionary-Based String Matching for Deep Packet Inspection”, In Proc. of INFOCOM 2010, 2010, pp. 1-5, doi: 10.1109/INFCOM.2010.5462268.

[12] P. Lin, Y. Lin, T. Lee, and Y. Lai, “Using String Matching for Deep Packet Inspection”, IEEE Computer, 2008, vol. 41, pp. 23-28, doi: 10.1109/MC.2008.138.

[13] G. La Mantia, D. Rossi, A. Finamore, M. Mellia, and M. Meo, “Stochastic Packet Inspection for TCP Traffic”, In Proc. ICC2010, 2010, pp. 1-6, doi: 10.1109/ICC.2010.5502280.

[14] A. Rao and P. Udupa, “A Hardware Accelerated System For Deep Packet Inspection”, In Proc. MEMOCODE’10, 2010, pp. 89-92, doi: 10.1109/MEMCOD.2010.5558646.

[15] R. Jurga and M. Hulbój, "Packet Sampling for Network Monitoring”, Technical Report, CERN | HP Procurve openlab project. Available at http://www.zdnetasia.com 20.1.2012

[16] D. Ficara, G. Antichi, A. Di Pietro, S. Giordano, G. Procissi, and F. Vitucci "Sampling Techniques to Accelerate Pattern Matching in Network Intrusion Detection Systems”, In Proc. ICC2010, 2010, pp. 1-5, doi: 10.1109/ICC.2010.5501751.

[17] S. Fernandes, R. Antonello, T. Lacerda, A. Santos, D. Sadok, and T. Westholm, "Slimming Down Deep Packet Inspection Systems”, In Proc. INFOCOM Workshops 2009, 2009, pp. 1-6, doi: 10.1109/INFCOMW.2009.5072188 .

[18] Z. Guo and Z. Qiu, "Identification Peer-to-Peer Traffic for High Speed Networks Using Packet Sampling and Application Signatures", In Proc. ICSP2008, pp. 2013-2019, doi: 10.1109/ICOSP.2008.4697540.

[19] M. Canini, D. Fay, D. Miller, A. Moore, and R. Bolla, "Per Flow Packet Sampling for High-Speed Network Monitoring”, In Proc. COMSNETS'09, 2009, pp. 1-10, doi: 10.1109/COMSNETS.2009.4808888.

[20] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on NetFlow traffic classification”, Computer Networks, Volume 55, Issue 5, 1 April 2011, pp. 1083-1099.

[21] M. Lee, M. Hajjat, R. Kompella, and S. Rao, "RelSamp: Preserving Application Structure in Sampled Flow Measurements”, In Proc. INFOCOM 2011, 2011, pp. 2354-2362, doi: 10.1109/INFCOM.2011.5935054.

[22] S. Dharmapurikar, P. Krishnamurthy, T. Sproull, and J. Lockwood, "Deep Packet Inspection using Parallel Bloom Filters", In Proc. High Performance Interconnects 2003, 2003, pp. 44-51, doi: 10.1109/CONECT.2003.1231477.

[23] Y. Li "Memory Efficient Parallel Bloom Filters for String Matching", In Proc. NSWCTC 2009, 2009, pp.485-488, doi: 10.1109/NSWCTC.2009.280.
[24] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, “Sketch-based change detection: methods, evaluation, and applications”, In Proc. of ACM SIGCOMM Internet Measurement Conference IMC’03, October 2003, doi: .
[25] R. Cong, J. Yang and G. Cheng, "Research of Sampling Method Applied To Traffic Classification", In Proc. ICCT 2010, 2010, pp. 112-115, doi: 10.1109/ICCT.2010.5689208.
[26] Ipoque. http://www.ipoque.com/ 20.01.2012
[27] J. Khalife, J. Verdejo, and A. Hajjar, “On the Performance of OpenDPI in Identifying P2P Truncated Flows”, AP2PS 2011, Lisbon, Portugal, 2011.
[28] F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus "Lightweight, Payload-Based Traffic Classification: An Experimental Evaluation", In Proc. ICC 2008, 2008, pp. 5869-5875.

**Jawad khalife** born in Lebanon, 1980, holds a master’s degree in telecommunications networks, University of Saint-Joseph, Beirut, Lebanon, 2003, and an engineering degree in computer and communication, the Lebanese University, Beirut, Lebanon, 2002. Since 2004, his teaching activities covered mainly network-related topics including network administration under Linux, Cisco and security courses in well-known faculties and institutions in Lebanon. Since 2002, he works as a Network Engineer in the central administration of the Lebanese university, Beirut, Lebanon. His current research interests covers enhancing traffic classification methods and their use in the security field, especially for intrusion detection systems. Eng. Khalife is a student member of IEEE.

**Amjad S. Hajjar** was born on May 12, 1964 in Chehim, Lebanon. He obtained his engineering diploma in electricity and electronics from the Lebanese university in 1986, and his Ph.D in computer-aided design (CAD) from the university of Paris-VI in 1992. He is currently assistant professor at the faculty of engineering of the Lebanese University, where he teaches computer networks, operations research and operating systems. His fields of interest in research are peer-to-peer (P2P) networks, traffic analysis and P2P activity detection.

**Jesús Díaz Verdejo** is a professor in the Department of Signal Theory, Telematics and Communications of the University of Granada. He received his B.Sc. in physics in 1989 and a Ph.D. degree in physics in 1995 from the University of Granada.

His initial research interest was related to speech technologies, especially automatic speech recognition. He is currently working on computer and network security, mainly focused in intrusion detection systems and traffic engineering from the point of view of security. He has also developed some work in telematics applications and e-learning systems.

# Anomaly Detection Using Metaheuristic Firefly Harmonic Clustering

Mario H. A. C. Adaniya, Taufik Abrão, Mario Lemes Proença Jr.
Department of Computer Science, State University of Londrina, Londrina, BR
Email: {mhadaniya}@gmail.com, {taufik, proenca}@uel.br

***Abstract*— The performance of communication networks can be affected by a number of factors including misconfiguration, equipments outages, attacks originated from legitimate behavior or not, software errors, among many other causes. These factors may cause an unexpected change in the traffic behavior and create what we call anomalies that may represent a loss of performance or breach of network security. Knowing the behavior pattern of the network is essential to detect and characterize an anomaly. Therefore, this paper presents an algorithm based on the use of Digital Signature of Network Segment (DSNS), used to model the traffic behavior pattern. We propose a clustering algorithm, K-Harmonic means (KHM), combined with a new heuristic approach, named Firefly Algorithm (FA), for network volume anomaly detection. The KHM calculate the weighting function of each point to calculate new centroids and circumventing the initialization problem present in most center based clustering algorithm and exploits the search capability of FA from escaping local optima. Processing the DSNS data and real traffic data is possible to detect and classify intervals considered anomalous with a trade-off between the 80% true-positive rate and 20% false-positive rate.**

***Index Terms*— anomaly detection, data clustering, firefly algorithm, k-harmonic means.**

## I. Introduction

Nowadays the network is a vital part of any company and even became an important part of our daily life. Technology innovations brought us a facility to gather and share information, to communicate our ideas with others, the opportunity to work from home and other small gestures that have became part of everyday life and it would be impossible to survive without the Internet. The infrastructure behind the convenience, in most cases, are monitored to prevent possible failures and loses of performance. The causes can be a simple misconfiguration to attacks that can harm the system, among many other. One of the causes that may affect the operation of the network is an anomaly behavior, which has a focus on areas such as Network Traffic, Data Mining, Image Processing, Credit Card Transactions and other pointed in [1].

In [1] and [2], the authors provide a structured and comprehensive overview of the research on anomaly detection summarizing other survey articles and discussing the importance and the applications in anomaly detection. The basis of the techniques and the manner how they are applied in some domains is described and the techniques are classified according to: Classification, Clustering, Nearest Neighboor, Statistical, Information Theoritic and Spectral. Still a challenge in anomaly detection, specially in the Network Traffic, is the identification of what could be considered an anomaly or not. In [3], the authors present a benchmark suite for volume anomalies based on the parameters: shape, duration, intensity and target. Throught the four parameters a set of sixteen scenarios have been presented.

In our work is considered an anomaly anything that is outside a threshold range created using the Digital Signature of Network Segment (DSNS) generated through GBA tool (Automatic Backbone Management) presented in [4] and briefly described in section III. The threshold range adopted is described in section IV.

Clustering is a technique where it is possible to find hidden patterns that may exist in datasets and it is possible to infer better conclusions. Clustering techniques are applied in Data Mining and known as "vector quantization" when dealing with Speech and Image data [5]. The most popular Clustering algorithm is K-means (KM) because it can deal with a large amount of data [6], is fast in most cases and it is simple to implement. The main basic idea is to partition the dataset into K clusters. Two weak aspects of KM are the sensitivity to initialization and the convergence to local optima [7]. To solve the initialization sensitivity Zhang proposed the K-Harmonic means (KHM) [8], minimizing the harmonic mean average of all points of N in all centers of K. In section V the KHM is discussed in detail.

In the literature heuristic there are methods where the main advantage pointed out by the authors is the characteristic of not converging rapidly to local optima. Tabu Search, Simulated Annealing, Particle Swarm Optimization, Ant Colony Optimization are examples of such methods. Firefly Algorithm (FA) is a relatively new method developed by Yang [9] in 2008. FA is inspired by the behavior of fireflies, the intensity of the lights and the attraction are the keys to the proper functioning of the algorithm. In section VI the algorithm is described in more detail.

This paper is based on "Anomaly Detection Using Firefly Harmonic Clustering Algorithm" by M. H. A. C. Adaniya, M. F. Lima, L. H. D. Sampaio, T. Abrão, and M. L. Proença Jr., which appeared in the Proceedings of the 8th International Joint Conference on e-Business and Telecommunications (ICETE), Seville, Spain, July 2011. ©2011 SciTePress.


doi: 10.4304/jnw.8.1.82-91

In this paper we proposed a hybrid data clustering algorithm based on KHM and FA, called Firefly Harmonic Clustering Algorithm (FHCA) described in section VII. Exploring the advantages of both algorithms to apply them to detect anomalies in real network traffic is possible to achieve a trade-off between the 90% true-positive rate and 30% false-positive rate.

In Section II some related works are discussed in the literature using heuristic, clustering and both techniques applied for detecting anomalies. Section III describes the GBA tool. Section IV describes the context anomaly adopted. Section V introduces KHM clustering. Section VI is relative to the Firefly Algorithm. Section VII is about the proposed algorithm. Section VIII presents the results achieved by the proposed algorithm. Section IX presents the conclusion and future improvements.

## II. Related Work

The anomaly detection receives special attention in Network Traffic, because it concern directly to quality and security of service provide to end-users, companies and other services are directly affected. In [10] is presented a survey on anomaly detection methods: statistical, based on classifier, machine learning and use of finite state machines. According to this classification, our model is based on a classifier, where the anomaly detection depends on the idea that normal characteristics behavior can be distinguished from abnormal behavior. Digital Signature of Network Segment (DSNS) generated by GBA (Automatic Backbone Management) tool is assumed as a normal traffic.

Techniques of clustering have the characteristic of grouping data objects into clusters, where the objects in each cluster are similar and different from others clusters. Through clustering it is possible to find new patterns in datasets that may need a new way to observe to make new . In [2], the authors pointed out the ability to learn from the data set, without the need to describe the various anomalies types, resulting in a reduction of time spent in training. Two main approaches are training using raw data containing both regular and anomaly samples and training only by using regular data.

Traditional signature based on automatic detections methods compare the network data values with a set of attack signatures provided by experts. One weakness is the dependence of data being provided by human expert leading the system to detect previously known attacks. Hereupon, in [11], the authors present an *unsupervised anomaly detection*, which trains on unlabeled data. Since the algorithm is designed to be general, the normalization is an important step of preparation of the training dataset. For each new data instance it computes the distance between it and each centroid of the clusters. The cluster presenting the shortest distance and if the distance is less than a adopted constant W (cluster width) then the instance is assigned to that cluster. The metric of distances is the Euclidean distance.

In [12], the authors are preoccupied with the notion of distance-based outliers. The approach presented is based on the Euclidean distance of the $k^{th}$ nearest neighbor from a point $O$. Herewith, it is created a list of "degree of outliersness" of point $O$ denoting the distances from $k^{th}$ neighbors. Ranking the list and assuming the top points as outliers they developed a highly efficient partition-based algorithm for mining outliers. Partitioning the input data set into disjoint subsets, and then cutting off entire partitions as soon as it is determined that they cannot contain outliers. This greedy approach results in substantial savings in computation.

In [13], the authors use the concept of outliers to differentiate between normal and abnormal data. Random forest is a algorithm of classification and classifies the data set in trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Random forests algorithm uses proximities to find outliers whose proximities to all other cases in the entire data are generally small.

In [14], the method concentrates on user command-level data and the authors proposed a system with host-based data collection and processing. The reasons the author point to adopt clustering are: a cluster presenting low variance is efficiently represented by its center, with a constraint on the cluster support is possible to reduce noise and retain more relevant clusters and if the intra-cluster similarity threshold (i.e., the minimum acceptable similarity between a cluster's center and any other sequence assigned to it) is set high enough, some test sequences may remain unassigned. "ADMIT" as the authors called the system works basic in three steps: 1) data pre-processing, 2) clustering user sequences, and 3) cluster refinement. The authors method achieve detection rate as high as 80% and a false positive rate as low as 15.3%.

Clustering and heuristics to form a hybrid solution with better results is found in [15] where the authors make use of the Bee Algorithm to overtake the K-means (KM) local optima problem, but there is still the initialization problem. In [16], the authors make use of Particle Swarm Optimization (PSO) and in [5] the authors use Simulated Annealing (SA), but the adopted clustering algorithm is the K-Hamornic Means (KHM).

The use of clusters and heuristics combined in the area of anomaly detection is also found in the literature. Lima [17] uses the PSO algorithm with the K-means clustering for detecting anomalies using as regular data the DSNS generated by BLGBA [4]. The system seeks to find the distances between points and their centroid, and for a given threshold value the system triggers alarms for the network administrator.

## III. Traffic Characterization

The first step to detect anomalies is to adopt a model that characterizes the network traffic efficiently, which

represents a significant challenge due to the non-stationary nature of network traffic. Large networks traffic behavior is composed by daily cycles, where traffic levels are usually higher in working hours and are also distinct for workdays and weekends. So an efficient traffic characterization model should be able to trustworthily represent these characteristics. Thus, the GBA tool is used to generate different profiles of normal behavior for each day of the week, meeting this requirement. These behavior profiles are named Digital Signature of Network Segment (DSNS), proposed by Proença in [4] and applied to anomaly detection with great results in [18].

Hence, the BLGBA algorithm was developed based on a variation in the calculation of statistical mode. In order to determine an expected value to a given second of the day, the model analyzes the values for the same second in previous weeks. These values are distributed in frequencies, based on the difference between the greatest $G_{aj}$ and the smallest $S_{aj}$ element of the sample, using 5 classes. This difference, divided by five, forms the amplitude $h$ between the classes, $h = (G_{aj} - S_{aj})/5$. Then, the limits of each $L_{Ck}$ class are obtained. They are calculated by $L_{Ck} = S_{aj} + h * k$, where $C_k$ represents the $k$ class $(k = 1 \ldots 5)$. The value that is the greatest element inserted in the class with accumulated frequency equal or greater than $80\%$ is included in DSNS.

The samples for the generation of DSNS are collected second by second along the day, by the GBA tool. The DSNS is generated for each day of the week. Figure 1 shows charts containing workdays and the weekend of monitoring of UEL network for the first week from August 2011. Data were collected from SNMP object *tcpInSegs*, at the University's Proxy server. The respective DSNS values are represented by the blue line, the collected data are represented in green and the red line is the collected data that surpasses the blue line.

In figure 1, the network behavior is well captured by the DSNS and can be observed that the blue line follows close along all the day the green area representing the monitored network traffic. The work hours cycles from the workdays is captured and seen in with the increase of traffic volume at 7 a.m. in the beginning of the university (office hours and classes) and decreasing around 12 a.m. for the lunch break, increasing again near 2 p.m. At 6 p.m. the traffic volume decreases but as the university offers courses at night, the traffic maintain a volume until around 11 p.m. For the weekends, Saturday and Sunday present different behaviors from each other. On Saturday some courses or events may occur generating a traffic volume but less than the traffic volume during the workdays. On Sunday the most important thing is a peak occurred between 7 a.m. and 8 a.m. caused by the backup system.

Because the DSNS can highly represent the daily cycles for workdays and weekends it is an efficient traffic characterization model and adopted as regular behavior of the network for our proposed anomaly detection system.

## IV. Anomaly Description

In this section, we introduce a definition adopted in our context to create a template in order to compare the classified intervals by the algorithm, Firefly Harmonic Clustering Algorithm (FHCA). This template is generated using the Digital Signature of Network Segment (DSNS) described in section III.

The anomalies causes can arise from various situations as flash crowds, network elements failures, misconfigurations, outages and malicious attacks such as worms, DoS (Denial of Service) and DDoS (Distributed Denial of Service). In our scenario, the anomaly description is preocupied in describe flash crowds (involuntarily there is an increase in network traffic for a short period of time saturating the server-side resources [19]), network elements failures, misconfigurations and outages which are situations that typically occur in the university campus.

Given $\mathbf{d}$, which represents the Digital Signature of Network Segment (DSNS) data, it can be described as a vector with $N$ positions, the position index is related to the timestamp of collect value and $\mathbf{d}$(index) = the collected value. The $J$ represents the value of the total lenght of intervals described as follows: $J = N/\Delta$, where $\Delta$ is the hysteresis interval and $N$ is the positions. For example, one day have $86400$ seconds. As explained in section III, we collect data from every second throughout the day, N = 86400, and assuming intervals duration for anomaly analysis of $\Delta = 300$ seconds, we have $J = 288$ intervals in a single day. Lets $\mathbf{d}_j$ be the vector of data values on the $j$-th interval. We can extend the analysis defining the concatenated vector $\mathbf{d}$ as:

$$[\mathbf{d}_1, \ldots, \mathbf{d}_j, \ldots, \mathbf{d}_J] = \mathbf{d}, \tag{1}$$

with dimension of $N \times 1$.

The parameter $\lambda$ is a measure of the deviation level occurred in the DSNS. We can adopt a constant value based on prior knowledge of the network or using a statistical measure. In our work, $\lambda$ is statistically characterized by:

$$\lambda = \frac{\frac{1}{J}\sum_{i=1}^{J}\sigma(\mathbf{d}_j)}{max[\sigma(\mathbf{d}_j)]}; \tag{2}$$

where $\sigma$ is the standard deviation of $\mathbf{d}_j$ and $max[]$ return the highest $\sigma$ of all the DSNS intervals.

The real traffic must go through the DSNS on a different scale and certain deviation is tolerable. In figure 2, the lines drawn represents the acceptable range created from DSNS. It is possible to observe that the traffic (red line) follows, in most of the time, the DSNS (blue line) and is inside the threshold range most of it. Depending on the network segment and the MIB objects collected data, the parameters may vary because of the volume. For example, on a HTTP server the traffic is measure by the IP address of destination and origin, and it is different from a Firewall, where all traffic passes before entering and/or leaving a network segment. As the volume passing

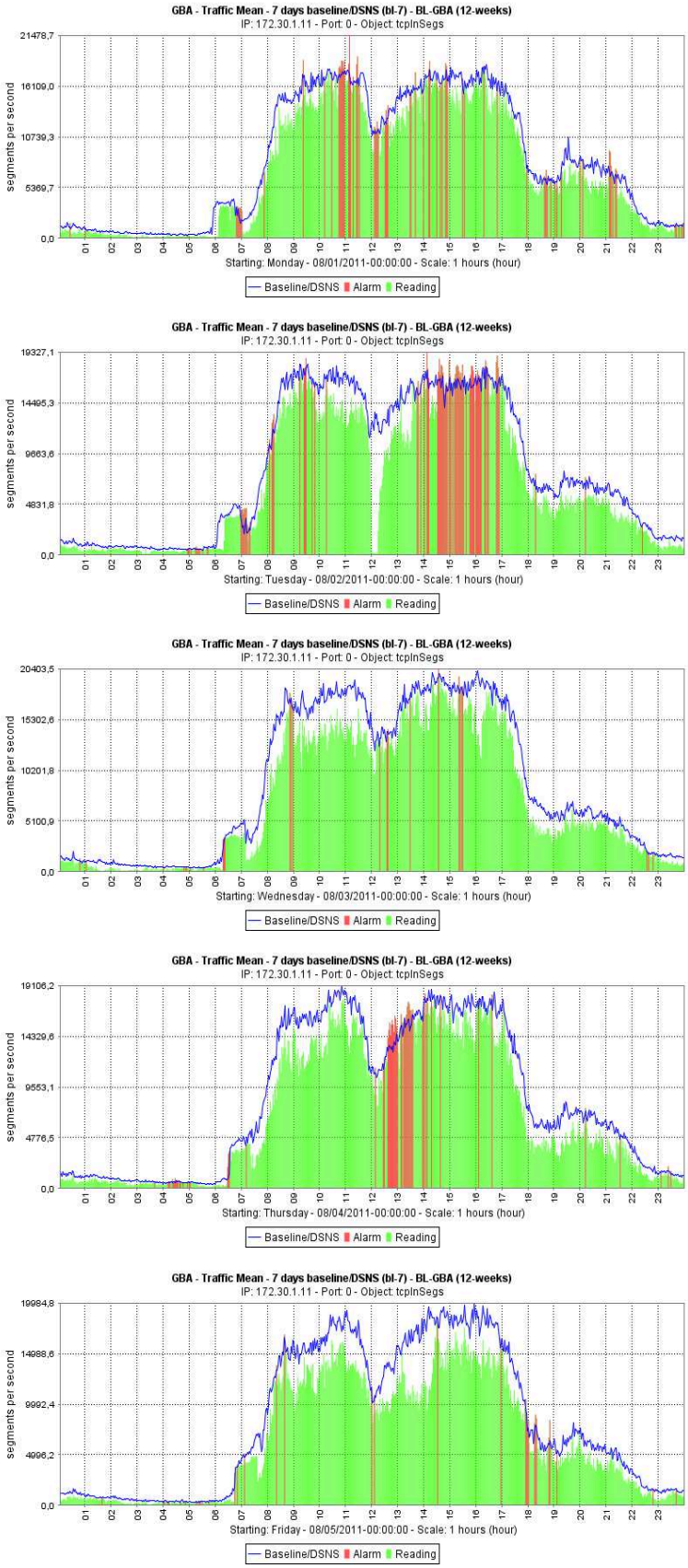



Figure 1. DSNS and real traffic collected from GBA from 08/01/2011 to 08/05/2011.

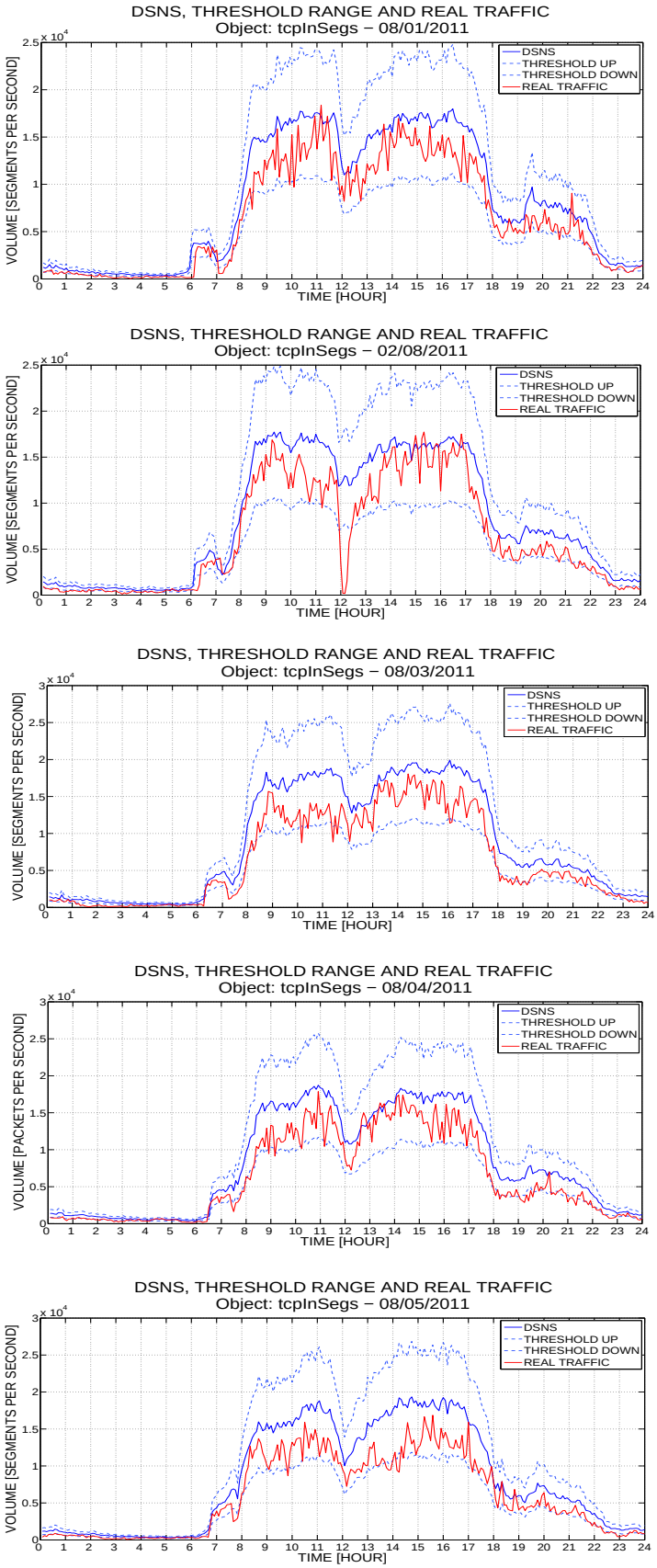



Figure 2. DSNS, the threshold range and real traffic of *tcpInSegs* SNMP object, on Web-Server of State University of Londrina.

by through the Firewall is larger than a HTTP server, the $\lambda$ is different.

To determine if a $\mathbf{d}_J$ is an anomaly or not, the equation 3 describes:

$$\mathbf{a}(j) = \begin{cases} 0, \lambda_{DOWN} < \mathbf{d}_j < \lambda_{UP} \\ 1, c.c. \end{cases} \quad (3)$$

where $J = N/\Delta$. $\mathbf{a}(j)$ is a vector contain boolean values for each $j^{th}$ interval, resulting into the template used to compare against the results from the tested algorithms in section VIII.

The $\lambda_{DOWN}$ and $\lambda_{UP}$ are calculated by the equations 4:

$$\begin{aligned} \lambda_{DOWN} &= \mathbf{d}_j - (\mathbf{d}_j * \lambda) \\ \lambda_{UP} &= \mathbf{d}_j + (\mathbf{d}_j * \lambda) \end{aligned} \quad (4)$$

We can observe that the presented threshold range can capture the normal variations of the traffic, classifying the intervals in regular and anomaly in a automatic and more precise manner. This generated template is used to compare to the intervals classified from the proposed anomaly detection system.

## V. K-Harmonic means clustering

A method of unsupervised classification of patterns into groups is called Clustering. In analyzing the data, the clustering problem has combinatorially characteristic. The existing clustering techniques are classified according to some features: agglomerative vs. divisive, monothetic vs. polythetic, hard vs. fuzzy, deterministic vs. stochastic, incremental vs. non-incremental [20]. Another important aspect in clustering is the similarity measure that define how similar a given data $x$ is to a cluster.

The K-means (KM) algorithm is a partitional center-based clustering method and the popularity is due to simplicity of implementation and competence to handle large volumes of data. The similarity function adopted is the Euclidean distance described in equation 5.

$$d(x, y) = \sqrt{\sum_{i=1}^{m} |x_i - y_i|^2} = \|\mathbf{x}_i - \mathbf{y}_i\| \quad (5)$$

The problem to find the $K$ center locations can be defined as an optimization problem to minimize the sum of the Euclidean distances between data points and their closest centers, described in equation 6. To state the problem the following notations are used [5]:

$$KM(\mathbf{x}, \mathbf{c}) = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - c_j\| \quad (6)$$

where $\mathbf{x}$ is the data set to be clustered and $\mathbf{c}$ the centers.

The KM randomly select $k$ points and make them the initial centres of $k$ clusters, then assigns each data point to the cluster with centre closest to it. In the second step, the centres are recomputed, and the data points are redistributed according to the new centres. The algorithm stop when the number of iterations is achieved or there is no change in the membership of the clusters over successive iterations [15]. One issue founded in KM is the initialization due to partitioning strategy, when in local density data results in a strong association between data points and centers [16].

In [8], Zhang proposed the K-Harmonic means (KHM), where the main idea is throught of the harmonic mean distance between a data point to all the centers. The author demonstrate that KHM is insensitive to the initialization of the centers due to the membership function (8) and weight function (9). The KHM optimization function is presented in equation 7:

$$KHM(\mathbf{x}, \mathbf{c}) = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \quad (7)$$

where $p$ is an input parameter of KHM and assume $p \geq 2$. In the harmonic mean, increasing the parameter $p$ the value of $\|x_i - c_j\|^p$, decreasing the value of $\frac{1}{\|x_i - c_j\|^p}$. The value of KHM converge to some value as $p \to \infty$.

The KHM calculate the membership function (8) describing the proportion of data point $x_i$ that belongs to center $c_j$:

$$m(c_j|x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}, \quad (8)$$

increasing the parameter $p$ in the membership function it give more weight to the points close to the center.

The weight function (9) defining how much influence data point $x_i$ has in re-computing the center parameters in the next iteration:

$$w(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{(\sum_{j=1}^{k} \|x_i - c_j\|^{-p})^2} \quad (9)$$

After calculating the membership function (8) and weight function (9), the algorithm calculate the new center location described by:

$$c_j = \frac{\sum_{i=1}^{n} m(c_j|x_i) w(x_i) x_i}{\sum_{i=1}^{n} m(c_j|x_i) w(x_i)} \quad (10)$$

The new center will be calculated depending on the $x_i$ and its $m(c_j|x_i)$ and $w(x_i)$, if $x_1$ is closer to $c_1$ and far from $c_2$, it will present a $m(c_1|x_1) > m(c_2|x_1)$ and calculating the new center $c_1$, $x_1$ will be more representative into the final answer for $c_1$ and less representative for $c_2$.

The pseudocode of KHM is presented in Algorithm 1 [5]:

## VI. Firefly Algorithm

Firefly Algorithm (FA) was designed by Yang [9] in 2008. FA was developed based on the behavior of fireflies and the behavior of light emitted. Many biologists still debate the importance and usage of the flashes used by fireflies, but it is known that is used to attract partners for mating, some cases to attract future prey, often as a security mechanism. Some important features are the length of the

**Algorithm 1** K-Harmonic means

**1**. Initialize the algorithm with randomly choose the initial centers;
**2**. Calculate the objective function value according to equation (7);
**3**. For each data point $x_i$ compute the membership value according to equation (8);
**4**. For each data point $x_i$, calculate the weight function according to equation (9);
**5**. For each center $c_j$, recompute its location based on the equation (10);
**6**. Repeat steps 2-5 until KHM(**x**,**c**) does not change or predefined number of iterations;
**7**. Assign data point $x_i$ to cluster $j$ with the biggest $m(c_j|x_i)$.

brightness, the brightness level and rhythm. It is known that the brightness level of $I$ is inversely proportional to the distance $r$, and $I \propto 1/r^2$, the brightness decreases with distance from the observer [21].

The proposed algorithm follows three rules: 1) all fireflies are unisex and can attract and be attracted, 2) The attractiveness is proportional to the brightness by moving the firefly fainter toward the brighter, 3) The brightness is directly linked to the function of the problem treated.

Two important issues must be addressed: the variation of light intensity and the formulation of attractiveness. The author suggests a simplifying assumption that the attractiveness of a firefly is determined by its brightness, which in turn is associated with the objective function encoded. The pseudocode is presented in Algorithm 2 [21]:

**Algorithm 2** Firefly Algorithm

Objective function $f(\mathbf{x}), \mathbf{x} = (x_1, ..., x_d)^T$
Initialize a population of fireflies $\mathbf{x_i}(i = 1, 2, ..., n)$
Define light absorption coefficient $\gamma$
**while** (t < MaxGeneration)
  **for** $i = 1 : n$ all $n$ fireflies
    **for** $j = 1 : i$ all $n$ fireflies
      Light intensity $I_i$ at $\mathbf{x_i}$ is determined by $f(\mathbf{x_i})$
      **if** ($I_j > I_i$) Move firefly $i$ towards $j$ in all $d$ dimensions **end if**
      Attractiveness varies with distance $r$ via exp$[-\gamma r]$
      Evaluate new solutions and update light intensity
    **end for** $j$
  **end for** $i$
Rank the fireflies and find the current best
**end while**

Yang [9] based the movement and the distance, given two fireflies, $i$ e $j$, the distance is given by:

$$r_{ij} = \|\mathbf{x_i} - \mathbf{x_j}\| = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2}, \qquad (11)$$

and the attraction movement performed by the firefly $i$ in relation to the brighter firefly $j$ is describe by:

$$\mathbf{x_i} = \mathbf{x_i} + \beta_0 e^{-\gamma r_{ij}^2}(\mathbf{x_j} - \mathbf{x_i}) + \alpha(\mathbf{rand} - \frac{1}{2}), \qquad (12)$$

where the second term is the attraction and the third term is randomization with $\alpha$ being the randomization parameter. **rand** is a uniform distribution in $[0, 1]$. The author adopt values of $\beta_0 = 1$ and $\alpha \in [0, 1]$. According to [9], the emission intensity of light from a firefly is proportional to the objective function, i.e., $I(x) \propto f(x)$, but the intensity with which light is perceived by the firefly decreases with the distance between the fireflies. The agents in FA have adjustable visibility and more versatile in attractiveness variations, which usually leads to higher mobility and thus the search space is explored more efficiently, getting a better ability to escape local minimum or maximum [9].

According to [9], the emission intensity of light from a firefly is proportional to the objective function, i.e., $I(x) \propto f(x)$, but the intensity with which light is perceived by the firefly decreases with the distance between the fireflies. Thus, the perceived intensity of a firefly is given by: $I(r) = I_0 e^{\gamma r^2}$, where $I_0$ is the intensity of light emitted, $r$ is the Euclidean distance between $i$ and $j$ firefly, $i$ being the bright and $j$ the less bright; and $\gamma$ the absorption coefficient. The attractiveness $\beta$ of a firefly is defined by:

$$\beta(r) = \beta_0 \exp^{-\gamma r^2}, \qquad (13)$$

The agents in FA have adjustable visibility and more versatile in attractiveness variations, which usually leads to higher mobility and thus the search space is explored more efficiently, getting a better ability to escape local minimum or maximum [9].

## VII. Proposed Anomaly Detection System

This section present the proposed anomaly detection system based on a clustering optimized algorithm, Firefly Harmonic Clustering Algorithm (FHCA), which cluster the Digital Signature of Network Segment (DSNS) data and the network traffic samples. The intervals are classified according to the Alarm System Generator.

### *A. Firefly Harmonic Clustering Algorithm*

Presented the K-Harmonic means (KHM) algorithm to clustering data in section V and the heuristic Firefly Algorithm (FA) in section VI, into this section will be discussed the implementation of Firefly Harmonic Clustering Algorithm (FHCA).

Merging and using the benefits of the two algorithms we propose the FHCA and applied it to the volume anomaly detection of network traffics. The first step is almost the same as presented in V, adding a step after (2), where we use the FA to optimize the equation 7.

The FHCA have a complexity of $O(NKDM^2)$, where N = data points, K = number of centers, D = dimension and M = the population of fireflies, resulting in a quickly convergence due to the fact of the Firefly Algorithm to perform local and global search simultaneously.

**Algorithm 3** Firefly Harmonic Clustering Algorithm

**Input**: Real traffic samples, DSNS;
**Output**: Clustered traffic and DSNS, traffic centroids ($\mathbf{ct}_j^k$), DSNS centroids ($\mathbf{cd}_j^k$);
1. Initialize a population of fireflies in random positions;
2. Define light absorption coefficient $\gamma$;
3. **While** (i $<$ Iteration)||(error $<$ errorAcepted(KHM(**x**,**c**)))
   Calculate the objective function according to equation (7);
   **For** $i = 1$ to $M$
      **For** $j = 1$ to $M$
         Light intensity $I_i$ at $x_i$ is determined by f($x_i$)
         **if** $(I_j > I_i)$
            Move firefly $i$ towards $j$ in all d dimensions
         **end if**
         Attractiveness varies with distance
         Evaluate new solutions and update light intensity
      **endFor** $j$
      Compute the membership function (Equation (8))
      Compute the weight function (Equation (9))
      Recompute $c_j$ location based on the Equation (10);
   **endFor** $i$
   Rank the fireflies and find the current best
**end while**

### B. Alarm System Generator

Once the centroids are defined the classification part classify the intervals in anomalous or normal. The following steps described by the Algorithm 4.

**Algorithm 4** Alarm System Algorithm

**Input**: Real traffic samples, DSNS, traffic centroids ($\mathbf{ct}_j^k$), DSNS centroids ($\mathbf{cd}_j^k$);
**Output**: Labeled intervals;
1. Calculate the distance ($\mathbf{D}_d$) between the DSNS ($\mathbf{d}_j$) and the centroids ($\mathbf{cd}_j^k$);
2. $M = max[\mathbf{D}_d]$;
3. Calculate the distance ($\mathbf{D}_t$) between the traffic samples ($\mathbf{t}_j$) and the centroids ($\mathbf{ct}_j^k$);
4. $MT = max[\mathbf{D}_t]$;
5. Define $\Lambda$;
6. **if** $MT > M$
      cont+1
   **endif**
7. **if** cont $> (\Lambda *$ size of MT)
      Classify the interval as ANOMALY
   **else**
      Classify the interval as NORMAL
   **endif**
8. Post process results and visualization

The parameters $\mathbf{D}_d$ and $\mathbf{D}_t$ are the distance matrix calculated by the distance between the samples and the centroids generated by FHCA. The operator $max[]$ returns the highest value.

The parameter $\Lambda$ describes a porcentage of acceptable points to exceed the distances diferences inside the analyzed interval. As we increase $\Lambda$ the number of anomaly points increase, $\Lambda \in [0, 1]$. For example, if we assume $\Delta = 300$ and $\Lambda = 0.2$, it means that 60 points inside the interval can exceed the distance diferences between DSNS and traffic samples from its centroids and be considered a normal interval, but if 61 points exceed, the interval will be consider anomalous.

The labeled intervals generated by the Alarm System are compared to the template generated by the Anomaly Description described in section IV.

## VIII. RESULTS

To validate the proposed algorithm were real data collected from the Proxy server of the network environment from State University of Londrina (UEL) which receive traffic from 5,000 computers connected to its network. One week starting from 08/01/2011 (Monday) until 08/07/2011 (Sunday) and the MIB object *tcpInSegs* which represent the total number of segments received.

To measure if the proposed approach is feasible or not, the metrics adopted are classical and discussed in [22]. Changing the nomenclature to our context, the metric is composed of several variables:

- **True Positive**: If the instance is anomaly and is classified as an anomaly;
- **False Positive**: If the instance is normal and is classified as an anomaly;

Through the declared variables can be calculated:

$$\textbf{False-positive rate (FPR)} = \frac{\textbf{False Positive}}{\textbf{Number of Normal Data}} \quad (14)$$

$$\textbf{True-positive rate (TPR)} = \frac{\textbf{True Positive}}{\textbf{Number of Anomaly Data}} \quad (15)$$

$$\textbf{Precision rate (PRE)} = \frac{\textbf{True Positive}}{\textbf{True Positive + False Positive}} \quad (16)$$

Equation (14) describes how much of the interval pointed by the algorithm was classified wrongly. Equation (15) describes the successes of the proposed algorithm classifying the intervals. Equation (16) is the percentage of corrected data classified throughout all the data classified. The Receiver Operating Characteristics (ROC) graph is a technique to visualize the performance based on the parameters and demonstrated the better trade-off between false-positive rate and true-positive-rate.

For the KHM parameters, $p = 2$ and FA parameters, $\gamma = \alpha = 0, 2$ and $\beta_0 = 1$ and the population, $N = \Delta/2$. These parameters did not influence directly in the results for the cluster formation because the data dimension treated in our scenario is the volume (one SNMP object per scenario), $D = 1$. The number of cluster formed into the dataset is an important characteristic, aiming betters clusters and proposing an center based algorithm method we tested $K = 2$, 3 and 4. In figure 3 is presented the true-positive rate varying in $\Delta$ intervals. $K = 2$ present the highest true-positive rate and low changing among the intervals.

The best result is achieved when $K = 2$, it can be explained by the average of the cluster might be low, resulting in a smaller number of objects associated and the clusters might be located far apart. An average result

is achieved when $K = 3$. $K = 1$ was tested with reservations, our group believe that one group does not represent all the network traffic data comprehended in that timespace. $K = 4$ present the lower rates, meaning a poorly classification. For the following graphs and results, we assume $K = 2$.

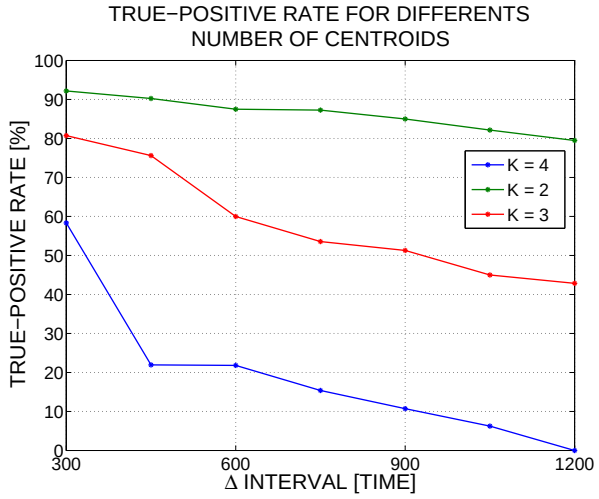



Figure 3. Comparison of true-positive rates for differents values of K.

The figure 4 present the highest and lowest perfomance of the algorithm respectively. The parameter $\Lambda$ of the Alarm System Generator is tested for [0,1]. As we increase $\Lambda$, the TPR also increase but the FPR increase as well. Higher the value of $\Lambda$ more greedy the algorithm behaves accepting more points inside the interval to exceed a maximum amount resulting in intervals wrongly classified.

The first conclusion about the proposed method is the better perfomance in workdays. The real traffic does not follow a general rule and for different days we have different results, the 08/01 is a Monday and the 08/06 is a Saturday. As discussed in section III, it was expected a different behavior from workdays and weekend, therefore, the Alarm System Generator need to be refined for weekends behavior. In figure 5 is presented the overall ROC curve compose of the average of the results for the workdays only.

From figure 5 we can conclude that the presented algorithm achives a trade-off 80% TPR and around 20% FPR. From all the intervals anomalous the proposed algorithm classify 80% intervals right. Unfortunately, the algorithm classify 20% of the regular intervals as anomalous. Naturally the challenge in the anomaly detection is to increase the TPR and decrease the FPR, and our results prove to be promising.

As stated before, the overall ROC takes the workdays and weekends to calculate, where the workdays pull up the TPR and the weekends pull down the FPR. Our research show promises results, because our classification model present a 80% TPR, meaning that our model can classify the rights interval, and with a little bit more of refinement it is possible to form clusters more accurate,

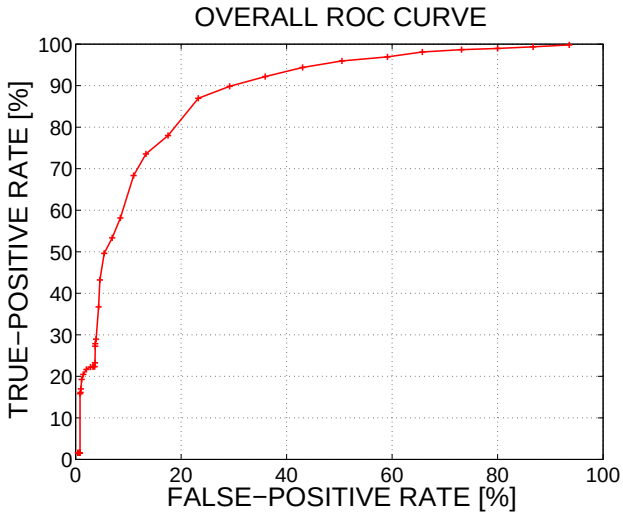



Figure 5. Overall ROC curve for the workdays.

and increase our classification and in exchange, our FPR will decrease.

## IX. CONCLUSIONS

In our work we proposed a new algorithm based on the merge of two algorithms: K-Hamornic means (KHM) and Firefly Algorithm (FA), named Firefly Harmonic Clustering Algorithm (FHCA). The FHCA utilizes the strength of KHM giving weight to members in calculating the centroids, circumventing the initialization problem present in center based clustering algorithm and exploits the search capability of FA in escaping local optima, resulting in better clusters.

Applying the FHCA to detect abnormalities in volume, the results achieved by the algorithm are satisfactory presenting high true-positive rates and low false-positive rates. The results present a true-positive rate above 80% and false-positive rates of nearly 20%. For workdays the algorithm results in better results than for the weekend, thus, the conditions to be consider an anomaly for the weekends and the parameters for the Alarm System Generator will be redesigned.

The next step is to combine the power of FHCA with another technique, i.e., Principal Component Analysis (PCA) or Support Vector Machine (SVM) to use other objects collected from the same segment network to group the results adding more information to increase the precision to classify correctly the intervals.

## ACKNOWLEDGEMENTS

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through a post-graduate master's degree level and SETI/Fundação Araucária and MCT/CNPq by the financial support for the Rigel Project.

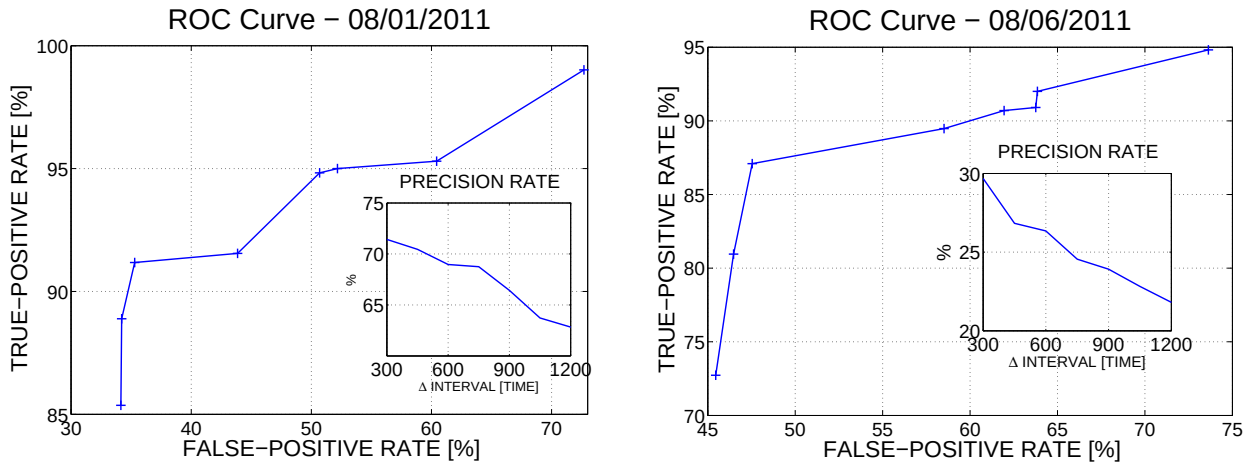



Figure 4. Highest and lowest perfomance of the algorithm.

## REFERENCES


[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey." *ACM Computing Surveys.*, vol. 41, no. 3, 2009.

[2] A. Patcha and J. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks: The International Journal of Computer and Telecommunications Networking*, vol. 51, pp. 3448–3470, August 2007.

[3] S. Shanbhag and T. Wolf, "Anombench: A benchmark for volume-based internet anomaly detection," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, December 2009, pp. 1–6.

[4] M. L. Proença Jr., B. B. Zarpelão, and L. S. Mendes, "Anomaly detection for network servers using digital signature of network segment," in *Advanced Industrial Conference on Telecommunications, 2005, Advanced ICT 2005. IEEE*, 2005, pp. 290–295.

[5] Z. Güngör and A. Ünler, "K-harmonic means data clustering with simulated annealing heuristic." *Applied Mathematics and Computation*, vol. 184, no. 2, pp. 199–209, 2007.

[6] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. USA: University of California Press, 1967, pp. 281–297.

[7] S. Z. Selim and M. A. Ismail, "K–means type algorithms: A generalized convergence theorem and characterization of local optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 81–86, 1984.

[8] B. Zhang, M. Hsu, and U. Dayal, "K-harmonic means - a data clustering algorithm," Hewlett-Packard Laboratories, Palo Alto, Tech. Rep. HPL-1999-124, Outubro 1999.

[9] X. Yang, *Nature-Inspired Metaheuristic Algorithms*. UK: Luniver Press, 2008.

[10] W. Zhang, Q. Yang, and Y. Geng, "A survey of anomaly detection methods in networks," in *International Symposium on Computer Network and Multimedia Technology.*, ser. CNMT 2009. USA: IEEE, January 2009, pp. 1–3.

[11] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," in *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001)*, 2001, pp. 5–8.

[12] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2000, pp. 427–438.

[13] J. Zhang and M. Zulkernine, "Anomaly based network intrusion detection with unsupervised outlier detection," in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 5, June 2006, pp. 2388–2393.

[14] K. Sequeira and M. Zaki, "ADMIT: anomaly-based data mining for intrusions," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 386–395.

[15] D. T. Pham, S. Otri, A. A. Afify, M. Mahmuddin, and H. Al-Jabbouli, "Data clustering using the bees algorithm," in *Proc. 40th CIRP Int. Manufacturing Systems Seminar*, Liverpool, 2007.

[16] F. Yang, T. Sun, and C. Zhang, "An efficient hybrid data clustering method based on k-harmonic means and particle swarm optimization." *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9847–9852, 2009.

[17] M. F. Lima, B. B. Zarpelão, L. D. H. Sampaio, J. J. Rodrigues, T. Abrão, and M. L. P. Jr., "Anomaly detection using baseline and k-means clustering," in *International Conference on Software, Telecommunications and Computer Networks*, ser. SoftCOM. USA: IEEE, September 2010, pp. 305–309.

[18] B. B. Zarpelão, L. de Souza Mendes, M. L. P. Jr., and J. J. P. C. Rodrigues, "Parameterized anomaly detection system with automatic configuration." in *GLOBECOM*. IEEE, 2009, pp. 1–6.

[19] L. Xie, P. Smith, D. Hutchison, M. Banfield, H. Leopold, A. Jabbar, and J. Sterbenz, "From detection to remediation: A self-organized system for addressing flash crowd problems," in *Communications, 2008. ICC '08. IEEE International Conference on*, may 2008, pp. 5809–5814.

[20] A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," *ACM Computing Survey*, vol. 31, no. 3, pp. 264–323, 1999.

[21] X. Yang, "Firefly algorithms for multimodal optimization." in *SAGA*, ser. Lecture Notes in Computer Science, O. Watanabe and T. Zeugmann, Eds., vol. 5792. Springer, 2009, pp. 169–178.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2005.

# Efficient DoS-limiting Support by Indirect Mapping in Networks with Locator/Identifier Separation

Daochao Huang, Dong Yang, Hongke Zhang
School of Electronic and Engineering, Beijing Jiaotong University, Beijing, China
Email: {08111043, dyang, hkzhang}@bjtu.edu.cn

Fuhong Lin
School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China
Email: dr.lin.bjtu@139.com

***Abstract*—Recent research in the designing of an elegant mapping service to map identifiers onto locators in networks with locator/identifier separation, focuses on solving practical issues related to mapping system. However, how to provide entire secure support in separation networks is still an open issue. In this paper, we present the design and evaluation of a hierarchical indirect mapping system (HIMS). It provides indirect mapping from connection identifier (CID), a novel flat identifier space introduced to stamp packets, to endpoint identifier (EID), can limit the impact of full range of destination attacks such as Denial of Service (DoS) floods from the outset by EID hidden, and fulfills the requirements such as low latency, efficient network utilization and scalability. Based on an efficient merging rule, HIMS build a hierarchical Chord architecture which can scale to Internet level by preserving the locality and convergence of the inter-domain path. We present scalability assessment and numerical results to demonstrate the performance gains of the proposed approach.**

***Index Terms*—Locator/identifier separation, connection identifier mapping, indirect mapping, Denial-of-Service, hierarchical DHT**

## I. INTRODUCTION

It has been widely recognized that today's Internet architecture is facing serious security problems. Two reasons lead to these security problems. First, the original Internet protocol stack design deliberately did not include solutions for security [1][2]. The original Internet architecture was designed to provide unicast point-to-point communication between fixed locations. This makes traditional Internet vulnerable to destination attacks, such as denial of service (DoS) and distributed denial of service (DDoS) [6][7]. Because packets transferring across the Internet encapsulate the source and destination IP addresses in their headers, the malicious nodes can easily capture data packets and launch attacks. Second, the trend toward network-based Cloud computing is driving security concern more seriously [3][4][5]. With the success of cloud computing, the Internet is increasingly a platform for online services—such as Web search, social networks, and video streaming—this makes people fetch services more convenient while leading to more security issues.

During the last few years, there has been a considerable number of proposals to address security issues, for a partial list of proposals, consider [8][9][10]. However, most existing security approaches rely on adding overlays to current Internet architecture, while these proposals achieve the desired functionality; they do so in a very disjointed fashion in that solutions for one service are not solutions for other services. So far, none of the proposals are centered on the idea of providing a novel entire secure Internet architecture which includes inherent security mechanism.

In this paper, we base on two recent proposals, the locator/identifier (Loc/ID) split routing architecture and identifier mapping system. The Loc/ID split, by [11], provides a clean-slate redesign of Internet architecture which is introduced to resolve traditional Internet architecture design problems, such as semantic overloading of IP address, lack of mobility support and poor robustness etc. The most related representative approach is LISP [12]. Both of these studies provide a Loc/ID split mechanism to solve the scalability issue of current Internet routing and addressing system.

The service identifier mapping system, SIDMAP [13], is a proposal to effectively resolve a service identifier (SID) to end-point identifier (EID) in Loc/ID separation networks when users want to apply to services identified by the SIDs. The SIDMAP consists of resolvers and a Chord-based mapping system. The Chord-based mapping system is an overlay network and comprises a collection of mapping servers that are used to store and retrieve SID-to-resolver mapping entries. A resolver stores and retrieves SID-to-EID mappings for end hosts and servers in its corresponding zone. With SIDMAP, both mapping resolution delay and maintain overhead are significantly reduced by multilayer cache mechanism and local entries update. Unfortunately, the basic SIDMAP is only designed to directly resolve a SID to one or more EIDs and unable to deal with the type of flooding denial-of-service attacks.


doi: 10.4304/jnw.8.1.92-99

In order to fully support security isolation through mapping system, we introduces a new identifier space, connection identifier (CID), into Loc/ID separation networks and design a hierarchical Distributed hash table (DHT)-based CID-to-EID pair indirect mapping system (HIMS). We argue that HIMS, with some modifications to SIDMAP, could provide a DoS-limiting infrastructure for Loc/ID separation networks. To be effective, the HIMS system must satisfy several important goals for its users. It must be:

**Secure**: The HIMS should address destination attacks such as DoS, DDoS, making service providing and communication among users more secure.

**Effective**: The HIMS should enable a receiver to detect attack traffic without inflicting damage to other legitimate hosts and with litter overhead burden to receiver or routers.

**Scalable**: Based on Chord or any other DHT algorithms, HIMS should resolve mapping requests quickly without introducing unacceptable resolution delay even widely deployed in large-scale Internet.

**Reliable**: If one indirect mapping server fails to work, the HIMS should limit the impact of the failure, and provide stable mapping service to relevant legitimate communications.

The rest of the paper is organized as follows. Section II describes the relevant background and related work; While Section III presents a concrete design and implementation of an indirect mapping system. Section IV analyzes the scalability and performance of HIMS. Section V evaluates our approach through numerical results and Section VI concludes the paper.

## II. Background And Related Work

Our work encompasses Loc/ID separation architecture and many aspects of DoS limiting mechanisms. In what follows we first present the main ideas of Loc/ID split architecture, then summary some representative pieces of research closely related to our work.

### *A. The Main Ideas of Loc/ID Split*

The namespace of current Internet, IP address, is used for two separate functions: 1) as an end-point identifier to uniquely identify “who” a device is; 2) as a locator for routing purposes, which describes “where” a device is attached to the network. This "overloading" of functions makes it virtually impossible to build an efficient routing system without forcing unacceptable constraints on end-system use of addresses. The Loc/ID split is introduced to split these functions apart by using different numbering spaces for EIDs and route locators (RLOCs). This decoupling yields several advantages, including improved scalability of the routing system through greater aggregation of RLOCs, persistent identity in the EID space and, in some cases, the efficiency of network mobility support. But while its benefits have been widely discussed, there has been less attention paid to the entire security approach that is key to Loc/ID split networks.

In the Locator/Identifier Separation mechanism Internet end-hosts depend on the network elements (routers) to look up the mapping between EID and RLOC. Typically, the mappings are stored in a distributed database called the mapping system, which responds to the lookup queries. Some prototype implementations such as, LISP-DHT [14] and DHT-MAP[15] have already been implemented.

### *B. Existing DoS Defence Mechanisms*

In the area of DoS, many approaches have been proposed in the past, such as Ingress filtering[16], Overlay based filtering (SOS)[17], network filtering[18], Capability based approach[19], SIFF (Stateless Internet Flow Filter)[20]. Some of the proposals mainly focus on making all sources identifiable. For examples, Ingress filtering [16] discards packets with widely spoofed addresses at the edge of the network, and traceback uses routers to create state so that receivers can reconstruct the path of unwanted traffic. However, attackers may still launch packet floods with unspoofed packets under these solutions.

By giving legitimate hosts an authenticator off-line that permits them to send to specific destinations, SOS can block the attacking traffic from malicious attacker effectively. Besides, typically filtering mechanism is the most commonly used method. Unfortunately, these filters will block some legitimate traffic from the receiver because there is no clean way to discriminate attack traffic from other traffic, for example, ingress filtering suffered from the shortcoming that a single unprotected ingress allows remote spoofing. In summary, these approaches only address an aspect of the problem but not the entire problem, and they do not provide a complete solution by themselves. A robust and efficient systematic approach that overcomes the shortcomings of current packet filtering techniques by allowing destination to control what it receives and automatically validating senders without prior arrangement is needed.

Packet filtering is a traditional tool for migrating DoS flooding attacks: when a receiver does not want to receive traffic from a sender, it can request to install filters to block the traffic. However, existing packet filtering systems are rendered ineffective by source address spoofing because it is easy for attackers to spoof source IP addresses to evade attack detection and packet filtering in the current Internet. Source address spoofing also enables attackers to launch reflector attacks, in which the attackers can hide behind innocent sources and the attack traffic can be significantly magnified. Our solution to authenticate source addresses is using self-certified flat CIDs instead of IP addresses during communication process.

In [20], by enabling routers stamp packets with a key that reaches the receiver and is returned to authorize the sender, SIFF built a stateless Internet flow filter that can eliminate the separate overlay channel for request packets and per-flow state. However, the SIFF proposal suffered from the following weaknesses: 1) for efficiency, only short stamps (2 bits) embedded in normal IP packets were adopted, and thus potentially discoverable by brute-force attack. 2) Initial request packets are forwarded with low

priority. This allows malicious hosts to take over the connected links.

Capabilities are short-term authorizations that senders obtain from receivers and stamp on their packets. This provides the permission in the form of capabilities to those senders whose traffic it agrees to accept. Both capability-based and filter-based approaches are promising building blocks for DoS flooding defense systems. They both enable a receiver to control the traffic it receives, but differ dramatically in methodology. Under the DoS-limiting traffic validation architecture (TVA)[21], it is required that each packet carries unique capacities that are not easily forgeable or usable if stolen by other party. Routers on the path validate these "stamps" but are not required to trust the hosts. Through this way, capacities expire to control the flow to destination while causing little overhead both in computation and bandwidth. When the attack power is very low, filters are more effective than capacities, while combining with per-source-AS fairness, capacities might be more cost-effective than filters. However, traditional capability-based approaches leave many questions unanswered, such as how capabilities are granted without being vulnerable to attack.

## III. Design Overview

In this section, we present the main ideas of our design. The overall goal is to strictly hide endpoint identifiers so that two hosts can communicate despite attacks by other hosts using an indirect mapping. To this end, we start with the definition of connection identifier. We then introduce standard packets forwarding and routing with CIDs in Loc/ID split based network. Detailers of indirect mapping scheme are described finally.

### *A. The Definiton of CID*

We design connection identifier mapping servers as entities that offer CID-to-EID pair resolution services to Internet end-hosts. The corresponding identifier space needs to uniquely identify connection between the two sides of communication by using a unique, temporary, and global scope CID. The lifetime of a CID depends on the session length, that is, it starts when a connection is established and ends when the session finished.

CIDs have significant potential benefits compared to other schemes such as filters and capabilities. They do not require routers to participate in filtering unwanted packets using implicit features. However, to be viable as a DoS-limitting solution, CIDs must meet several implied requirements. First, they must be generated by destination and authenticated by indirect mapping system, so that they can be distributed to the sender and stamped on packets. Second, routers in Loc/ID separation networks must be able to verify CIDs through communicating with indirect mapping servers. Third, CIDs must expire so that a destination can cut off a sender from whom it no longer wants to receive packets, in other words, the lifetime of a CID is strictly limited in the session period. Fourth, CIDs is designed as flat, self-certifying identifier excluding any semantic information of the session. Finally, CIDs must add little overhead in the common case. The following CIDs' format design is geared towards meeting these requirements.

CIDs is the cryptographic hash of the information such as sender's EID, destination's EID, timestamp and a small random number. We can summarize the format as follows:

**CID=hash(src EID, dest EID, timestamp, random number)(160 bits)**

Figure 1. Format of CIDs

### *B. Packets with CIDs*

With CID indirect mapping system, a typical session between a sender $H_s$ with $EID_s$ and a receiver $H_d$ with $EID_d$ is:

- $H_s$ sends a service request to Internet, asking for service from $H_d$. The Loc/ID split based network with CID indirect mapping system answers with a CID.
- $H_s$ sends traffic to $H_d$, with CID embedded in each packet.

All CID-stamped packets are piggy-backed into normal traffic between senders and receivers. We illustrate this through an example shown in Fig. 2, assuming that each AS has one or several HIMS servers that handle CID requests from its ITR/ETR and other ASes.

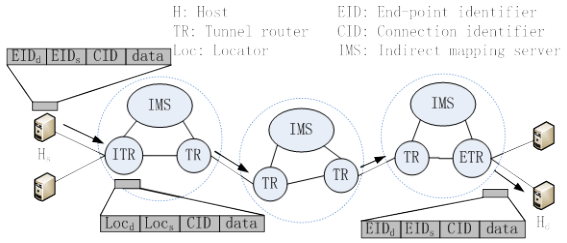



Figure 2 Illustration for the communication process.

- Step1 A source host $H_s$ that wishes to require a service from a destination host $H_d$ sends a service request to its tunnel router $ITR$. This request includes the communication's source and destination endpoint identifiers: $(EID_d, EID_s)$, the connection identifier CID and the payload data.
- Step2 The tunnel router $ITR$ resolves the identifiers $EID_d$ and $EID_s$ to corresponding locator $Loc_d$ and $Loc_s$ respectively using existing methods such as LISP-DHT, DHT-MAP etc., and then forwards the request including source and destination locators: $(Loc_d, Loc_s)$, the connection identifier CID and the payload data to the destination tunnel router $ETR$.
- Step3 The CID indirect mapping server in the destination $H_d$ 's AS forwards an inter-domain HIMS request to the destination indirect mapping server of CID to resolve CID to corresponding

$(EID_d, EID_s)$ pair. Finally, *ETR* forwards the request to the destination host.

The impact of destination attacks should be limited. In such attack, attackers send mass attack traffic to the target, congest the link to the target, and exhaust the target's resources. We use CIDs instead of EIDs in core network to achieve EIDs hidden and limit the impact of destination attacks.

### *C. Basic Design of Indirect Mapping*

There are several challenges in designing this indirect mapping system:

There must be a way for senders and receiver to request CID from mapping system, and such mapping system should ensures that the increased communication delays after CID notification are not significantly higher than the delays prior to existing approaches without mapping system. In our design, mapping system servers are organized as distributed system based on hierarchical DHT algorithm.

CIDs should be unforgeable, and ITR/ETR should be able to efficiently verify CIDs. We use consistent hashing function to generate a unique, temporary, and global scope Connection Identifier to ensure Unforgeability of CIDs as well as efficiency in CID generation and verification. The temporary characteristic indicates that the lifecycle of each CID is limited and only be used to identify one connection within a limited time period. Even if an attacker can obtain CIDs by pretending to be a good sender, it cannot abuse it later.

The CIDs generation and distribution should be bound in order to make mapping system scalable. In our design, each mapping server does not have to keep the global knowledge to locate the sender/receiver. We also designed an algorithm that uses distributed hash table to forward the request/response between multiple mapping servers to obtain the destination server of a CID.

Since CIDs have to be embedded into each packet, the header overhead should be minimized. We allow CIDs caching on mapping servers to significantly reduce the header overhead and mapping resolution delay.

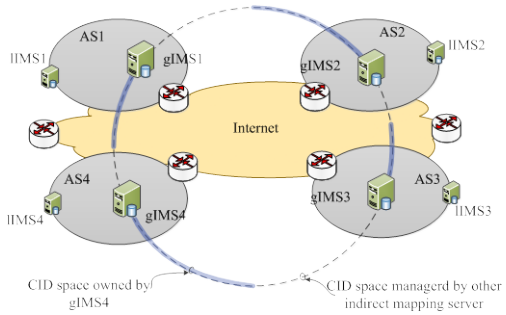



Figure 3. Example of HIMS infrastructure

To meet the requirements described above, we propose to use a hierarchical Chord ring in which CIDs are used as ChordIDs, and have domains create a local Chord ring for each of ASes (see Fig.3). As shown in Fih.3, The proposed two levels hierarchical indirect mapping system (HIMS) consists of low level components, local indirect mapping server (lIMS), and top level components, global indirect mapping server (gIMS). Such a deterministic approach allows the HIMS to benefit from following advantages: 1) Typically, inter-domain traffic accounts for a large proportion; hierarchical design enables mapping resolution delay decreases significantly with efficient caching and bandwidth usage. 2) The proposed approach can achieve good performance with limited costs. Due to large quantity of CIDs will exist simultaneously; the tradeoff between performance (e.g. resolution delay) and cost (maintenance overhead) needs to be considered. By reasonable design of merging multiple Chord rings, HIMS can use less resource usage to achieve the same performance. 3)The HIMS supports adaptation to the underlying physical network. One of the main challenges in Chord is that end-hosts have little control over the location of their connection identifiers. This is because mapping server identifiers are randomly chosen, and, therefore, mapping servers close together in the identifier space can be far apart in the underlying network. To solve this problem, locality and convergence of inter-domain paths are considered to adapt to underlying physical networks in HIMS. When the sender which starts the CID lookup and the destination indirect mapping server is in the same domain, then the lookup never leaves this domain. This is called locality of inter-domain paths. When different nodes from one domain A route to the same node in another domain, all the different routes exit the domain A through the same node. This node is the closest successor of the target node's identifier in the domain A. This is called convergence of inter-domain paths. In view of above considerations, in HIMS, each lIMS x in one sub-Chord ring creates a link to a lIMS y in another sub-Chord ring if and only if:

- y is the closest lIMS that is at least distance $2^k$ away for some $0 \le k < m$, where $m$ is the length of a CID.
- y is closer to x than any node in x' sub-Chord ring.

Fig.4 depicts the merging process for lIMS 0 (in sub-Chord ring A) and lIMS 7 (in sub-Chord ring B). In view of merging rule described above, lIMS 0 links to lIMS 2 and 4, lIMS 7 links to lIMS 9 and 11.

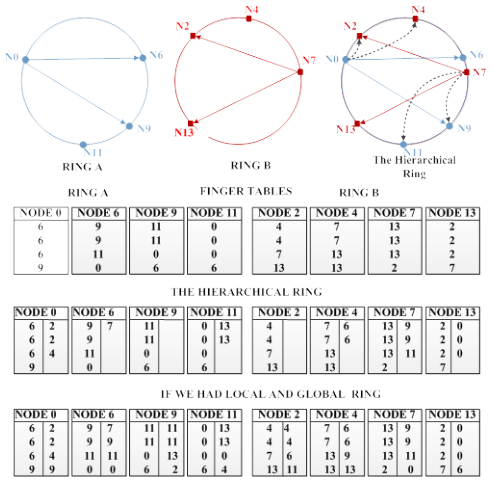



Figure 4. Illustration of merging multiple Chord rings

Compared with two Chord rings, one local and one global, the proposed HIMS only needs 14 extra links while two Chord rings 32 extra links.

### C. Registration of CID-to-EID pair Mappings

Whenever an end host wants to communicate with other hosts, it should obtain a CID from the proposed indirect mapping system. We take a classical client-server process for example to illustrate the steps of registration process of a CID in HIMS. Assuming that every service is accompanied by a metadata file that includes the server's public key and as well as its digital signature over the publication data.

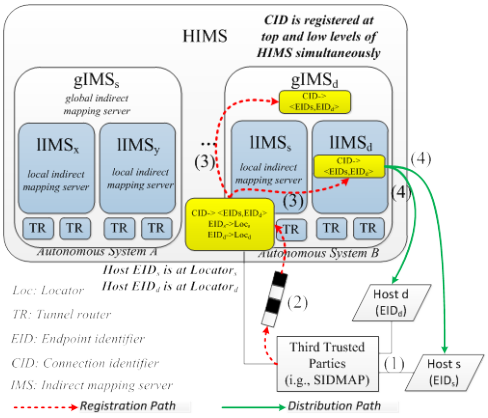



Figure. 5 Register a CID into HIMS

In HIMS, a registration process comprises the following steps.

- Step 1. When a sender (or client) wants to obtain a service, it first depends on third trusted parties such as Domain Name System (DNS) or SIDMAP for service retrieval and revocation, then resolves this service's name to its EID using external, reliable mapping mechanisms (e.g. SIDMAP) in Loc/ID split based network (see Fig.5 (1)).
- Step 2. Instead of returning the resolution results to the client, the third trusted parties create a CID for the request which is used to stamp on subsequent packets of this session later and forward the CID with relevant information including source EID, destination EID etc. to the nearest lIMS server through the tunnel router (TR) which further resolves the endpoint identifier to its corresponding locators using mapping system such as LISP-DHT. The details of this process can be illustrated in Fig.5 (2).
- Step 3. When the lIMS server receives the message from the third trusted parties, it first check whether or not the CID belongs to its range according to Chord protocol. If it is the case, the lIMS is the destination lIMS of the CID which is responsible for creating a new CID-to-< $EID_s$, $EID_d$ > and then storing in its local mapping table. After that, the lIMS needs to forward the mapping to the gIMS in the same AS. If not, the lIMS forwards the CID registration message to the CID's destination lIMS and gIMS (see Fig.5 (3)). Both the destination lIMS and gIMS add a new CID-to-<$EID_s$, $EID_d$> mapping entry in their local database. Then, the CID is registered at top and low levels of HIMS simultaneously.
- Step4 The destination lIMS needs to answer the sender with the CID by distributing the CID to the sender and the required server simultaneously. After receiving the CID, the sender sends service request with the CID to tunnel router which is responsible for verifying and forwarding the packets to the server.

### D. Resolving EID Pair for a CID

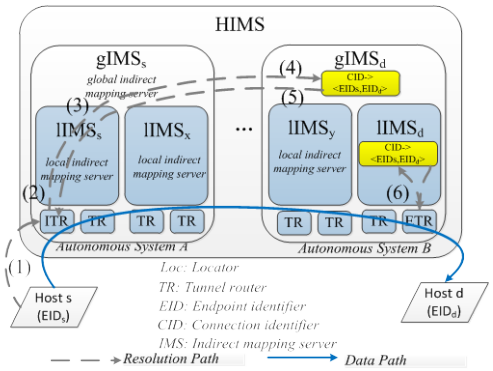



Figure. 6 Resolving a CID to EID pair

The steps of resolution process of a CID in HIMS are as follows:

- Step 1. The end host sends a message which includes the CID and payload data to its ITR to indicate that it will communicate with the destination network through that TR (see Fig.6 (1)).
- Step 2. When the ITR receives the message, it sends a mapping request to the lIMS in its domain. The mapping request should contain the CID information and some signature used for security (see Fig.6 (2)).
- Step 3. When the lIMS receives such a mapping request, it first finds in its local CID-to-EID pair mapping table whether there is a CID-to-EID pair for the required CID or not. If so, then directly return back resolution results to the ITR. Otherwise, the lIMS should forward the mapping request to its nearest gIMS (see Fig.6 (3)).
- Step 4. When the gIMS receives the mapping request, it first finds whether the requested CID is in its local domain. If so, returns the CID-to-EID pair mapping to the ITR. If not, the gIMS should forward the mapping request to its nearest gIMS that is closest to the destination gIMS (see Fig.6 (4)).
- Step 5. When the destination gIMS receives the mapping request, it sends the resolution result back to the ITR. Then, the ITR further resolves the EIDs to their locator and forwards the packet to

the destination. In this transforming process, EIDs are replaced by the CID (see Fig.6 (5)).

- Step 6. When the packet is arriving at ETR, the ETR first sends a resolution request to the HIMS to resolve the CID, then forwards the packets to the destination (see Fig.6 (6)).

### *E. Security Primitive Provided by HIMS*

DoS attack occurs when an attacker uses a thousand systems to simultaneously launch attacks against a remote host, and then floods the bandwidth or resources of a targeted host. These collections of systems are known as botnets. The major advantages to an attacker of using a DoS attack are that multiple machines can generate more attack traffic than one machine, multiple attack machines are harder to turn off than one attack machine, and that the behavior of each attack machine can be stealthier, making it harder to track down and shut down. These attacker advantages cause challenges for defense mechanisms in traditional Internet.

The root cause of the problem lies in the IP address leakage. The original Internet architecture was designed to provide unicast point-to-point communication between fixed locations. In this basic service, the sending host knows the IP address of the receiver and the malicious node can easily obtain the IP addresses of sender and receiver, and then attaches a destination attack to the sender or the receiver. With HIMS, communication partners, neither clients nor servers, never know the exact access identifiers of each other, in other words, HIMS hides the identifiers of communication hosts. At the client, when a packet is sent out to local ITR, the source EID and the destination EID are $EID_s$ and $EID_d$ respectively. After local ITR receives the packet, it first extracts the CID and matches it with its cache entries of CID mappings, then forwards the packet to the destination indirect mapping server of the CID. The same process proceeds at the server. By EID hidden, communication security is provided as built-in function of HIMS in universal network. Instead of explicitly sending a packet to a destination using IP addresses or EID pairs, each packet is associated with a connection identifier; this identifier is then used by the sender and the receiver to obtain delivery of the packet. As a result, HIMS provides an efficient method to alleviate or, to some extent, eliminate DoS attack.

## IV. SCALABILITY AND PERFORMANCE ANALYSIS

We give a preliminary and simplified assessment of the system scalability in terms of required number of world-wide HIMS servers. For this assessment, we assume a world-wide HIMS with two levels (gIMS, lIMS). In order to store a binding record in a valid CID mapping table, it details a storage space of 4KB (>160 bits $EID_{src}$+160 bits $EID_{dst}$+160 bits CID+8 bits timestamp+ 16 bits random number etc.). Solid State Disk (SSD) memory instead of traditional hard drive is used to store these binding records to offer sufficiently fast access (15μs, let $\bar{W}$ denote the average sojourn time that equals to waiting time plus service time, then $\bar{W} = 15$ μs). Current state-of-the-art SSD storage servers have 4TB of memory [22]. Assuming that $10^{15}$ (>8.4 × $10^9$ indexed Web pages [23], about $10^5$ connections per page) CIDs globally with mapping records of 4KB coexist simultaneously, and each gIMS can store $10^9$ mapping records. Therefore, $10^6$ indirect mapping servers are required for a world-wide HIMS with $10^{15}$ CIDs.

Typically, regular Chord is not equal to the number of routers since each hop on the logical, overlaying network connecting the Chord nodes may comprise of a number of physical communication links and their routers. HIMS efficiently solve this problem by implementing intra-domain paths locality and inter-domain paths convergence. To illustrate the efficiency of HIMS and analyze the performance, we present a modeling as follows: Given a random graph, the average number of routers between two peers in the network is given by[23]

$$\langle d \rangle = \frac{\ln\left[\left(N_R - 1\right)\left(\hat{z}_2 \hat{\hat{z}}_1 - z_1\right) + z_1^2\right] - \ln\left(z_1^2\right)}{\ln\left(\hat{z}_2 / z_1\right)}$$

where $\hat{z}_i$ is the average number of $i$ hop neighbors and $N_R$ is the total number of nodes in the router graph. Since the HIMS is implemented as a hierarchical Chord ring, the expected average number of routing hops between two nodes is $\frac{1}{2}\log_2\left(N-1\right)+\frac{1}{2}$ with $N > 1$, where $N = 2^m$ is the total number of HIMS servers. Therefore, each query is forwarded for $\left\lceil \frac{1}{2}\log_2\left(N-1\right)+\frac{1}{2} \right\rceil$ hops with an average of $\langle d \rangle$ routers per hop and assuming that the return path has an equal number of routers. The expected network delay of the resolution process is then

$$E\left[T_{QS}\right] = \left[\left(\log_2\left(N-1\right)+1\right)\langle d \rangle \sum_{I=1}^{N_R}\left(E\left[W_{Q_i}\right]+\tau_i\right)\right] \Big/ N_R$$

where $\left[\sum_{I=1}^{N_R}\left(E\left[W_{Q_i}\right]+\tau_i\right)\right] \Big/ N_R$ is the average queuing delay at a router, and $E\left[W_{Q_i}\right]$ is the expected waiting time in the $i$ th router which is given by [23] as follows:

$$E\left[W_{Q_i}\right] = \tau_i \rho_i \left(c_{ai}^2 + c_{si}^2\right) g_i \Big/ 2\left(1-\rho\right)$$

In which $g_i = g_i\left(\rho_i, c_{ai}^2, c_{si}^2\right)$ is defined as

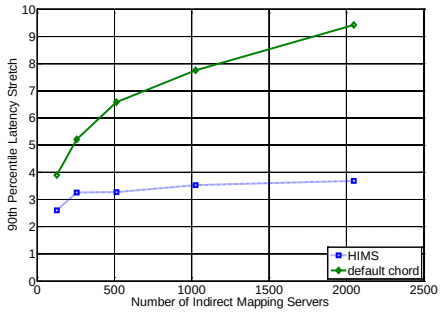
$$g_i\left(\rho_i, c_{ai}^2, c_{si}^2\right) = \begin{cases} \exp\left[-\dfrac{2\left(1-\rho_i\right)\left(1-c_{ai}^2\right)^2}{3\rho_i\left(c_{ai}^2+c_{si}^2\right)}\right] & , c_{ai}^2 < 1 \\ 1 & , c_{ai}^2 \geq 1 \end{cases}$$

The total average lookup latency of a request being resolved by HIMS is denoted as $\bar{L}$, which is the sum of the average sojourn time and network delay. Thus,

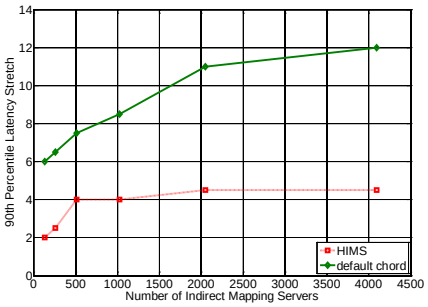
$$\bar{L} = \bar{W} \cdot \left\lceil \frac{1}{2}\log_2\left(N-1\right)+\frac{1}{2} \right\rceil + E\left[T_{QS}\right]$$

## V. Simulation Results

We evaluate the efficacy and scalability of the HIMS using numerical simulations. These simulations are based on the Chord protocol [25] and uses recursive style routing. We consider the following two network topologies in our simulations: *Topology 1:* A real network topology generated with the kingdata [26] with 2501 nodes. In this topology, the distance between two DNS servers is used to simulate the distance of two indirect mapping servers. Therefore, the simulation results greatly reduce the deviation between simulation and real network because it includes a connection of real network RTT (Round-trip Time) values. *Topology 2:* A power-law random graph topology generated with INET topology generator [27] with 16, 384 nodes, where the delay of each link is uniformly distributed in the interval (1, 80) ms. The HIMS servers are randomly assigned to the network nodes.




(a)




(b)

Figure 6. The 90th percentile latency stretch in the case of (a) A real network topology generated with the kingdata with 2501 nodes, and (b) a power-law random network topology with 16384 nodes.

Consider a sender $EID_s$ communicates with a receiver $EID_d$ via $CID_{sd}$. As discussed in Section 3.3, packets are routed in locator/identifier separation network with CID instead of EIDs, and then EID hidden is achieved to limit destination attacks such as DoS. During the communication process, mapping requests are sent to HIMS to resolve CID to its corresponding EID pairs. We use the ratio of the inter-node latency on the HIMS network to the inter-node latency on the underlying network to evaluate the routing efficiency of HIMS.

As shown in Fig.6, both in *Topology 1* and *Topology 2*, the 90th percentile latency stretch can be reduced up to 1.5-3 times as compared to the default Chord protocol since intra-domain paths locality and inter-domain paths convergence are considered in HIMS.

To evaluate how efficiently our design choices use HIMS for better lookup performance, we introduce the average number of bytes sent per node per unit time as the cost metric. This cost accounts for all messages sent by a node, including periodic routing table refresh traffic, lookup traffic, and join traffic. The performance vs. cost simulation results are as follows:

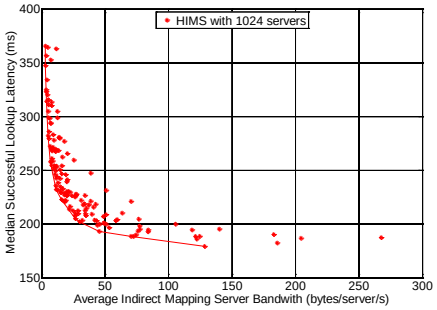



Figure 7. Performance vs. cost tradeoff in HIMS with 1024 servers

Fig.7 presents the performance/cost tradeoff in HIMS. Each point represents the median lookup latency of successful lookups vs. the communication cost achieved for a unique set of parameter values. The convex hull (solid line) represents the best achievable performance/cost combinations. The lookup latency and the maintenance overhead are contradictive, which cannot achieve the best at the same time.

## VI. Conclusions

With the disadvantages of traditional Internet architecture becoming more and more obvious, a number of novel identifier split mechanisms are proposed so as to solve the issues mentioned above. However, most of these proposals cannot provide entire security support. In light of this issue, a novel secure hierarchical DHT-based connection identifier mapping system called HIMS is proposed in this paper, the basic idea of which is to set a connection of mapping servers that store CID-to-EID pair mapping entries for the corresponding CIDs. Using HIMS, terminals communicate with each other without knowledge of the correspondence node's EID, so as to limit destination attacks such as DoS attack, providing security support in locator/identifier separation Network.

To demonstrate the feasibility of our approach, we have built a connection identifier indirect mapping system based on the hierarchical Chord lookup system. Preliminary experience with suggests that the system is highly flexible and provides secure, scalable and good worst-case lookup performance.

## Acknowledgment

This work was supported in part by a grant from the National 863 Key Project (2011AA010701), and Basic Research Supporting of Beijing Jiaotong University (2009JBZ004-2).

## REFERENCES


[1] A. Feldmann, “Internet clean-slate design: what and why?”, SIGCOMM Comput. Commun. Rev., vol. 37, no. 3, pp. 59–64, 2007.

[2] J. Roberts, The clean-slate approach to future internet design: a survey of research initiatives, Annals of Telecommunications, vol. 64, no. 5, pp. 271–276, Jun. 2009.

[3] L. M. Kaufman, “Data security in the world of cloud computing, ” IEEE Security &Privacy Magazine, vol. 7, no. 4, pp. 61–64, Jul. 2009.

[4] H. Takabi, J. B. D. Joshi, and G.-J. Ahn, “Security and privacy challenges in cloud computing environments”, IEEE Security & Privacy Magazine, vol. 8, no. 6, pp. 24–31, Nov, 2010.

[5] LM Vaquero, L Rodero-Merino, D Morán, “Locking the sky: a survey on IaaS cloud security”, Computing, vol. 91, no. 1, pp. 93-118, 2011.

[6] R. K. C. Chang, “Defending against flooding-based distributed denial-of-service attacks: a tutorial”, Communications Magazine, IEEE, vol. 40, no. 10, pp. 42–51, 2002.

[7] V. Thing, M. Sloman, and N. Dulay, “A survey of bots used for distributed denial of service attacks”, New Approaches for Security, Privacy and Trust in Complex Environments, vol. 232, ch. 20, pp. 229–240, 2007.

[8] E. Gelenbe and G. Loukas, “A self-aware approach to denial of service defence”, Computer Networks, vol. 51, no. 5, pp. 1299–1314, Apr. 2007.

[9] D. K. Y. Yau, J. C. S. Lui, F. Liang, and Y. Yam, “Defending against distributed denial-of-service attacks with max-min fair server-centric router throttles”, IEEE/ACM Transactions on Networking, vol. 13, no. 1, pp. 29–42, 2005.

[10] Y. Kim, W. C. Lau, M. C. Chuah, and J. H. Chao, “PacketScore: A Statistics-Based packet filtering scheme against distributed Denial-of-Service attacks”, IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 2, pp. 141–155, 2006.

[11] D. Yang, H.-c. Zhou, H.-k. Zhang, “Research on Pervasive Services Based on Universal Network”, ACTA ELECTRONICA SINICA, vol. 35; no. 4, pp. 607-613, 2007.

[12] D. Farinacci, V. Fuller, D. Oran, and D. Meyer, “Locator/ID Separation Protocol (LISP)”, IETF Internet draft, draft-farinaccilisp-07.txt, work in progress, Apr. 2008.

[13] D.-c. Huang, D.Yang, F. Song, H.-k. Zhang, “SIDMAP: A Service-oriented Mapping System for Loc/ID split Internet Naming”, Journal of Communications, vol. 6, no. 8, pp. 601-609, 2011.

[14] L. Mathy, L. Iannone, and O. Bonaventure, “LISP-DHT: Towards a DHT to Map Identifiers onto Locators”, IETF Internet draft, draftmathy-lisp-dht-00.txt, work in progress, Feb. 2008.

[15] H. Luo, Y. Qin, H.-k. Zhang, “A DHT-Based Identifier-to-Locator Mapping Approach for a Scalable Internet”, IEEE Trans. Parallel Distrib. Syst. (TPDS), pp. 1790-1802, 2009.

[16] P. Ferguson and D. Senie. Network Ingress Filtering: Defeating Denial of Service Attacks that Employ IP Source Address Spoofing. Internet RFC 2827, 2000.

[17] A. Keromytis, V. Misra, and D. Rubenstein, SOS: Secure Overlay Services, ACM SIGCOMM, 2002.

[18] F. Huici, and M. Handley, “An edge-to-edge filtering architecture against DoS”, ACM SIGCOMM CCR, vol. 37, no. 2, pp. 41 - 50, April 2007.

[19] T. Anderson, T. Roscoe, and D. Wetherall, “Preventing Internet denial-of-service with capabilities”, ACM SIGCOMM CCCR, vol. 34, no. 1, pp. 39-44, Jan. 2004.

[20] A. Yaar, A. perrig, and D. Song, “SIFF: a stateless Internet flow filter to mitigate DDoS flooding attacks”, in Proc. IEEE Symp. Security and Privacy, 2004, pp. 130 - 143.

[21] X. Yang, D. Wetherall, and T. Anderson, “TVA: a DoS-limiting network architecture”, IEEE/ACM Transactions on Networking, vol. 16, no. 6, Dec. 2008, pp. 1267 - 1280.

[22] Texas Memory Systems, www.ramsan.com. Last checked: June 2011.

[23] http: //www.worldwidewebsize.com/

[24] K.K. Ramachandran and B. Sikdar, A Queuing Model for Evaluating the Transfer Latency of Peerto-Peer Systems, IEEE Transactions on Parallel and Distributed Systems, vol. 21, 2010, pp. 367-378.

[25] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-To-Peer Lookup Service for Internet Applications. SIGCOMM 2001, pp. 149-160.

[26] http: //pdos.csail.mit.edu/p2psim/kingdata/

[27] http: //topology.eecs.umich.edu/inet/



**Dochao Huang** received the M.S. degree in communication systems from the Beijing Jiaotong University of electronic and engineering of China, Beijing, China, in 2007. Now he is a Ph.D. candidate of Beijing Jiaotong University. His research interests are in the areas of communication networks including cloud computing, data center and Next Generation Internet technologies.

**Dong Yang** received the M.S. and Ph.D. degrees in communication and information systems from the Beijing Jiaotong University of electronic and engineering of China, Beijing, China, in 2003 and 2008, respectively. His research interests are in the areas of communication networks including wireless networks, and Internet technologies.

**Fuhong Lin** received the M.S. and Ph.D. degrees in communication and information systems from the Beijing Jiaotong University of electronic and engineering of China, in 2006 and 2010, respectively. He is a post doctor at University of Science and Technology Beijing. His research interests include social network, P2P network, and Future Internet technologies.

**Hongke Zhang** received M.S. and Ph.D. degrees in electrical and communication systems from the University of Electronic Science and Technology of China, Chengdu, China, in 1988 and 1992, respectively.

From September 1992 to June 1994, he was a Post-Doctoral Research Associate with Beijing Jiaotong University (BJTU), Beijing, China. In July 1994, he joined BJTU, where he is currently a Professor in School of Electronic and Information Engineering. He is also the Chief Scientist of the National Basic Research Program of China.

# Comparison and Handover Performance Evaluation of the Macro-mobility Protocol

NIE Gang
School of Mathematics & Computer Science, Wuhan Textile University, Wuhan, China
nie.gang@qq.com

QING XiuHua
School of Electronic & Electrical Engineering, Wuhan Textile University, Wuhan, China
niegang@gmail.com

***Abstract*—Future internet works will include large numbers of portable devices moving among small, wireless cells. In order to support real-time applications, users demand seamless mobility and Quality-of-Service (QoS) provisioning. The handover technology is one of the most important technologies for the quality of real-time operation which guarantee the QoS of mobile Internet. Mobility management issues are solved by separating macro-mobility (inter-domain) generally using Mobile IP and micro-mobility (intra-domain). In recent years, many enhancements for macro-mobility protocols have been proposed, designed and implemented in Mobile IPv6. Hierarchical Mobile IPv6 (HMIPv6) is one of them that is designed to reduce the amount of signaling required and to improve handover speed for mobile connections. This paper presents a comparative study of the Hierarchical Mobility IPv6 and Mobile IPv6 protocols. The architecture and operation of each protocol is studied and are evaluated based on the QoS parameter: handover latency. The simulation was carried out by using the Network Simulator-2. The first simulation results show that, HMIPv6 performs best under macro-domain mobility compared to MIPv6. The MIPv6 suffers large handover latency. The second simulation results show that the proposed scheme can get better performance in terms of packet loss and handoff delay**

***Index Terms*—Macro Mobility protocol, Quality of Service, Handover, Hierarchical Mobile IPv6, Performance Evaluation**

## I. INTRODUCTION

Next generation wireless networks offer the promise of high speed access as well as IP-based data services to the mobile hosts. Protocols would be required to maintain the same level of performance in the wireless networking environment with frequent handoffs, as in the wire-lined environment.

As a mobile node (MN) travels between wireless cells, data transfer between the MN and the correspondent node (CN) will be typically changed from an old to a new access router (AR). In most mobility solutions, this process involves changes of routing entries in the MN and the CN, in addition to some designated mobility agents (home agent and/or foreign agent), and is called a handoff. It must ensure that end-to-end connectivity is maintained in a seamless way despite the changed path. So, the most famous protocol that supports mobility in IP networks, Mobile IP, does not handle these requirements of future wireless IP networks efficiently. It produces a lot of control traffic inside the local domain that increases the handoff delay and the risk of packet losses. Therefore, mobility management issues are solved by separating macro-mobility generally using Mobile IP and micro-mobility. Macro-mobility concerns the management of mobile movements on a large scale between different wide wireless access networks connected to the Internet. Micro-mobility, on the other hand, covers the management of local movement inside a particular wireless network, or domain. In order to achieve a fast handoff mechanism with minimum packet loss and a signaling system with minimum registration, a number of IP macro-mobility protocols that answer these performance and scalability issues have been developed.

Mobile IPv6 (MIPv6) [12] describes how mobile node can change its point of attachment from one access router to another. As a demand for wireless mobile devices increases, many enhancements for macro-mobility (inter-domain) protocols have been proposed, designed and implemented in Mobile IPv6. Hierarchical Mobile IPv6 (HMIPv6) [11] is one of them that is designed to reduce the amount of signaling required and to improve handover speed for mobile connections. This is achieved by introducing a new network entity called Mobility Anchor Point (MAP).

In this paper, we present a comparative study of the HMIPv6 and MIPv6 protocols. The architecture and operation of each protocol is studied and are evaluated based on the Quality of Service (QoS) parameter: handover latency. The simulation was carried out by using the Network Simulator-2.

The rest of this paper is organized as follows. Section II discusses some related work. Section III discusses Mobile IPv6 Handover Process. Section Ⅳ presents our performance analysis. Finally, in Section Ⅴ, we conclude this paper and give the future work.

## II. RELATED WORK


doi: 10.4304/jnw.8.1.100-107

### *A. Mobile IPv6*

The mobile IPv6 protocol enables a mobile node to communicate with other nodes after changing its point of attachment from one IP subnet to another without changing its IP address. Without this feature the Mobile Node (MN) would not be able to maintain transport and higher layer protocol sessions, which depend on a static IP address. The problem is solved by assigning two IP addresses to a mobile host: a permanent IP address, which is called IP home address, and a temporary IP address, named care-of-address (CoA), which is assigned each time the mobile node is visiting a new foreign subnet. Home address is used for transport and higher layer sessions whereas CoA is needed to route packets correctly to the actual point of attachment. In this way, the impact of host mobility is reflected only in the routing layer.

A mobile node (MN) is a node that can change its point of attachment to the Internet, whereas a correspondent node (CN) is a host communicating with the MN. A home agent is a router in the MN´s home network, handling the mobility of the MN. Every time that the MN gets a new CoA, it has to register it to the home agent and the CN, by sending binding update (BU) messages. Packets of new calls are intercepted by the home agent and then tunneled to the MN´s CoA by using IPv6 encapsulation, whereas packets belonging to active sessions are sent directly to CoA. Packets from the MN are sent to the CN through the Internet as usual. In basic Mobile IP, each handover implies end-to-end signaling. This can take a considerable time before all the CNs and the home agent have received the BU messages. Packets sent to the old CoA during this time period are potentially lost. For local mobility, several protocols have been proposed as extensions to Mobile IP to enable seamless handover.

### *B. Hierarchical Mobile IPv6*

Hierarchical schemes separate mobility management into micro mobility and macro mobility or otherwise known as intra-domain mobility and inter-domain mobility respectively [1]. The central element of this framework is the inclusion of a special conceptual entity called Mobility Anchor Point (MAP). It is a router or a set of routers that maintains a binding between itself and mobile nodes currently visiting its domain. It is normally placed at the edges of a network, above the access routers, to receive packets on behalf of the mobile nodes attached to that network. When a mobile node attaches itself to a new network, it registers with the MAP serving that network domain (MAP domain).

The MAP acts as the local home agent for the mobile node. It intercepts all the packets addressed to the mobile node it serves and tunnels them to the corresponding on-link Care of Address (CoA) of the mobile node. If the mobile node changes its current address within a local MAP domain, it only needs to register the new on-link address with the MAP since that the global CoA does not change. If a mobile node moves into a new MAP domain, it needs to acquire a regional address (RCoA) and an on-link address (LCoA). The mobile node then uses the new MAP's address as the RCoA, while the LCoA address can be formed as stated in [11]. After forming these addresses, the mobile node sends a regular MIPv6 BU to the MAP, which will bind the mobile node's RCoA to its LCoA. If successful, the MAP will return a binding acknowledgment (BAck) to the mobile node indicating a successful registration. In addition to the binding at the MAP, the mobile node must also register its new RCoA with its home agent by sending another BU that specifies the binding between its home address and the RCoA. Finally, it may send similar BU to its current corresponding nodes, specifying the binding between its home address and the RCoA. The HMIPv6 Architecture is depicting in Fig. 1.

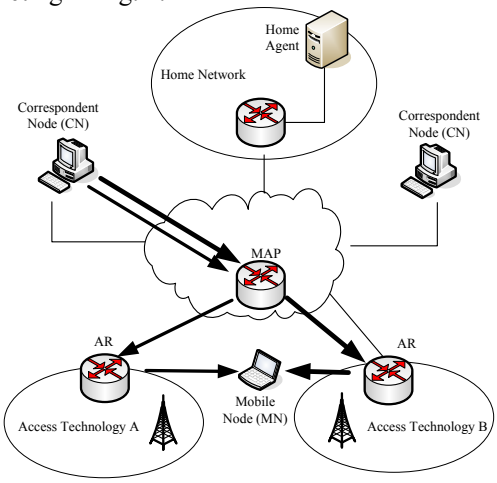



Figure 1. HMIPv6 Architecture.

### *C. Types of Handovers*

The term handoff or handover1 refers to the process of a mobile node (MN) moving from one point of attachment to the Internet to a different point of attachment. There are different types of handover according to which layers of the communication stack are affected. In general, handovers that only affect the link layer (i.e. L2) without resulting in a change of IP (i.e. L3) state are known as horizontal handovers. An example of this is when a MN moves between different Wireless LAN Access Points that are served by the same IP Access Router. In 802.11 terminologies, both Access Points belong to the same Extended Service Set (ESS). Handovers that affect both L2 and L3 (i.e. a new IP address is obtained by the MN) are known as vertical handovers.

Some literature makes a distinction between hard and soft handovers. A hard handover is when all the links (usually radio) in the MN are disconnected before the new link(s) are established. Conversely, a soft handover refers to the case where the MN is always connected to the network via at least one link. In this way, there is an overlap of different link usage during the handover process. Of course, this implies either multiple interfaces or multiple radio modules on a single interface are available on the MN.

All the above types of handover may be either inter-technology or intra-technology handovers. In inter-technology handovers the handover is between different network technologies, which would usually mean separate interfaces on the MN. Intra-technology handovers are handovers of the same network technologies. Horizontal handovers would usually be of the intra-technology type, although, technically, different network technologies could be used provided the IP layer sees no change its connectivity and associated state. Vertical handovers can just as easily be inter-technology as intra-technology.

To be somewhat pedantic, one could also categories L1 handovers such as when a Wireless LAN station switches between different frequencies and/or coding schemes of its current link. However, these are not considered to be of much relevance in the scope of this research.

### *D. Intra-domain Mobility in HMIPv6 And MIPv6*

Intra-domain mobility means a Mobile Node (MN) moves within the domain. The MN gets a CoA on its new point-of attachment. HMIPv6 is considered as a proposed protocol for intra-domain mobility because it differentiates between intra-domain mobility management scheme and inter-domain mobility management scheme. The MN will have 2 CoAs which is Local Care of Address (LCoA) and Regional Care of Address (RCoA). The RCoA remains constant as long as the MN is roaming locally. Thus, MN mobility is completely hidden in intra-domain mobility. This is achieved by introducing the new entity, Mobility Anchor Point (MAP), which will improve the handover latency as well [3].

Fig. 2 briefly explains the Macro-mobility in HMIPv6:

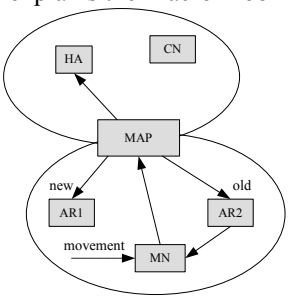



Figure 2. Macro-Mobility in HMIPv6.

- MAP sends the advertisement (contains RCoA) to all Access Router (AR) within its region.
- A MN entering a MAP domain receives Router Advertisement (RA) from nearest AR that contains information (RCoA) on one or more local MAPs.
- Then, MN sends Local Binding Update (LBU) only which contains LCoA to MAP.

While in Fig. 3, it explains the intra-domain mobility in MIPv6:

- AR sends the advertisement (contains CoA) to MN.
- Then, MN will choose the nearest AR.
- Finally, it will send a Binding Update (BU) to CN and HA.

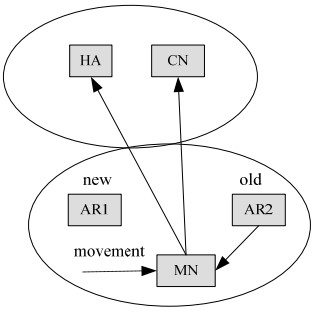



Figure 3. Macro-Mobility in MIPv6.

## III. MOBILE IPV6 HANDOVER PROCESS

### *A. Handoff in Mobile IPv6*

The Mobile IPv6 (MIPv6) specification is a proposed standard by the IETF to provide transparent host mobility within IPv6. The protocol enables a Mobile Node to move from one network to another without the need to change its IPv6 address. A Mobile Node is always addressable by its home address, which is the IPv6 address that is assigned to the node within its home network. When a MN is away from its home network, packets can still be routed to it using the MN's home address. In this way, the movement of a node between networks is completely invisible to transport and other higher-layer protocols.

When a MN changes its point of attachment to the Internet from one IPv6 network to another IPv6 network (also referred to as roaming), it will perform the MIPv6 handover procedure. The MIPv6 handover procedure is similar to the auto-configuration procedure of an IPv6 node booting up onto a network, but has some important differences:

- MN must somehow detect that it has moved onto a new network.
- Once configured, the MN must inform its home agent (HA) and each correspondent node (CN) of its new location.
- During the handover procedure, upper layer connections will still be active so the handover procedure should be performed as quickly as possible to minimize disruption from lost and severely delayed packets.

The MIPv6 handover procedure is illustrated in Fig. 4. In current IPv6 specifications, each stage of the procedure is mandatory with the exception of authentication and authorization, although this stage will be present, at least in some form, in most employed networks [5].

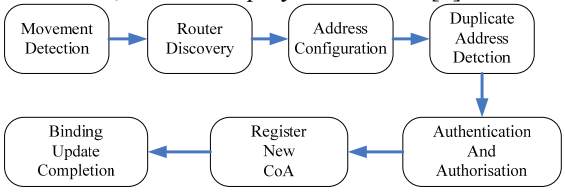



Figure 4. MIPV6 handover procedure.

Over Wireless LAN, we can identify the latency that can significantly affect handover delay during an IPv6

handover. The handoff mainly comprises the following components [6]:

- Movement detection time ($t_d$): this is the time required by the MN to detect and establish that it has moved to a new point of attachment (i.e. the discovery of a new on-link router).
- IP CoA configuration time ($t_a$): this is the time between the establishment of having moved and the time that a globally rout-able IPv6 address has been configured (this includes duplicate address detection (DAD)).
- Context establishment time ($t_c$): this is the time between the establishment of a globally rout-able CoA and the establishment of the appropriate context state.
- Binding registration time ($t_r$): past the establishment of context-specific state of the MN, this is the time between the dispatch of a binding updated (BU) signal to the HA to the receipt of an acknowledged BU piggybacked on the first packet from its communication peer.
- Route optimization time ($t_o$): this is the time from registering the new CoA with the HA to completing route optimization with the current list of CNs. This includes the return route-ability procedure which, if used, must occur before a BU is sent by the MN to a CN.

The total IP handover delay defined as the sum of these components: $t_h = t_d + t_a + t_c + t_r + t_o$. This is depicted in Fig. 5.

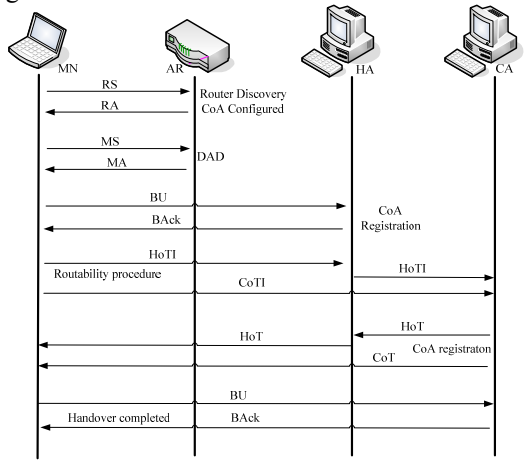



Figure 5. Components of IPv6 handover.

### B. Handoff in HIMPv6

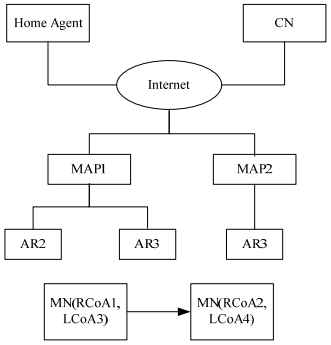



Figure 6. Inter-domain Handoff in HMIPv6.

The system for Inter-domain handoff is based on the following scenario: the MN has a regional Care-of Address RCoA1 and an on-link Care-of Address LCoA3, which is depicted in Fig.6. When the CN sends packets to the MN, the packets will be sent through MAP1 to the MN's LCoA3.

*1)* When the MN is about to move from the MAP1 domain to the MAP2 domain:

*a)* The MN sends a request control message to MAP1 to construct a multicast group for the MN.

*b)* MAP1 forms a multicast group for the MN and sends a multicast group join request to all other neighboring ARs. The neighboring ARs send response messages after receiving these multicast group requests toward MAP1 to show their availability to receive multicast packets from MAP1.

*c)* The packets encapsulated by MAP1 are tunneled from the CN to the multicast group members. These ARs buffer the packets. As a final point, these neighboring ARs forward the packets.

*2)* When the MN travels from MAP1 domain to MAP2 domain:

*a)* The MN initially acquires a new address from the MAP2 network (RCoA2, LCoA4). The MN sends a Binding Update to MAP2 through AR4 and sends a message requesting AR4 to forward a multicast message. AR4 receives the request message, and subsequently forwards the buffered packets to the MN.

*b)* Whereas AR4 constantly sends multicast packets to MN, MAP2 receives the Binding Update and checks for DAD. MAP2 sends a Binding Update to the MN's Home Agent after receiving the DAD. After that MAP2 waits for a binding acknowledgment from the Home Agent. MAP2 followed by sends a Binding Acknowledgment to the MN.

*c)* The MN receives the Binding Acknowledgment and sends a Binding Update to the CN via MAP2.

*d)* After receiving the Binding Update, CN changes the destination address RCoA1 to new RCoA2 and consequently directs the packets to MN in the new network via MAP2 and AR4.

*e)* AR4 stops sending multicast packets from MAP1 as soon as it receives new packets intended to the MN. MN at this moment receives packets directly from the CN as with Hierarchical Mobile IPv6.

## IV. Simulation And Discussion

### A. Analysis and Evaluation of MIPv6 Handovers

Fig. 7 shows the small MIPv6 testbed used for performing the handover tests. A MN running MIPL v1.1 is away from home (the HA also running MIPL v1.1) and can attach to one of two networks represented by the SSIDs 'roam1' and 'roam2'. We decided to conduct handover tests from one foreign network to another (e.g. rather than from home network to foreign network) as the nature of mobility implies that when one is mobile, one is very rarely located at the home network.

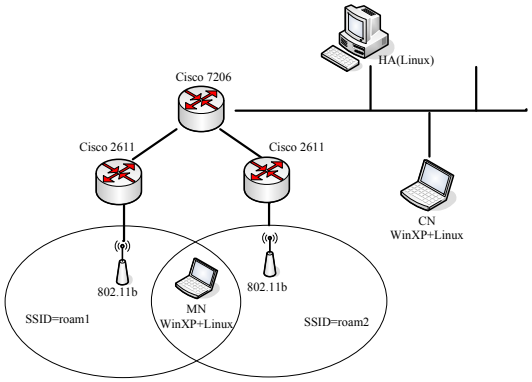



Figure 7. Simple MIPv6 handover testbed.

Handovers from one network to another were forced by turning off one of the APs so that the MN would immediately associate with the other AP and thus receive different RAs than on the previously connected network.

The handover times were measured from the point at which the AP is switched off, when the Binding Acknowledgement is received from the Correspondent Node with each event being time stamped in the relative logs. Note that the Return Rout-ability protocol was enabled for Route Optimization.

In order to demonstrate the effects of reducing the RA interval we performed handover tests with various configurations of RA intervals on the Cisco 2611 access routers. Upon detecting movement, the MN will issue a RS assuming it hasn't received a new RA already. As can be seen from Fig. 8, the time it takes for the MN to receive a solicited RA is fairly random within a given (configurable) time window.

The test experiment was shown in Table I. The IPv6 capable VoD server was located at the Correspondent Node and streamed 1.5Mbps MPEG1 video clip of 30 seconds duration was streamed to the Mobile Node. At 10 seconds into the video clip the AP to which the MN was associated with was switched off, forcing a handover to the other network. At the end of the clip the handover latency and packet loss (reported by the VoD client) were noted. This was repeated 10 times for each value of RA interval configured on the Cisco routers. These RA intervals were 300ms (the RFC 3775 minimum), 1000ms and 3000ms.

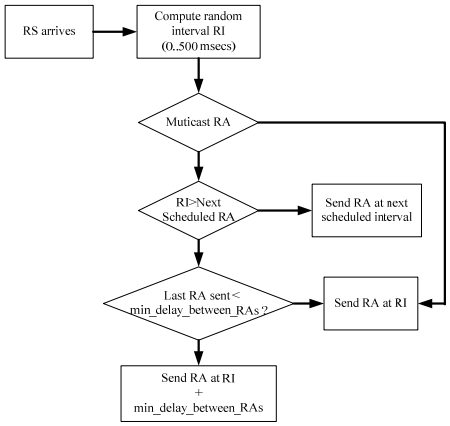



Figure 8. Simple MIPv6 handover testbed.

TABLE I.
RESULTS OF RA INTERVAL TESTS

| | *Avg Latency (seconds)* | *Avg Packets Lost* |
|---|---|---|
| 300 ms | 1.917 | 245.376 |
| 1000 ms | 2.448 | 313.344 |
| 3000 ms | 3.013 | 385.664 |

It can be seen that changing the RA interval does not have as much effect on reducing the overall latency as we would like. This can be explained in that the rest of the handover procedure after receiving a RA, i.e. CoA configuration, DAD and CoA registration with the HA and CN is completely unaffected by reducing the RA interval. One can also see the number of packets lost in the video stream. On the client playback the stream would recover itself after handover but the break in the video and audio seemed about 1 or 2 seconds longer than the handover latency reported in the logs. It is easy to conclude that even tuning the RA interval to the lowest possible value will not suffice for real-time voice and video applications in a mobile environment.

### *B. Analysis of FMIPv6 Handovers*

The aim of the Fast Handovers for Mobile IPv6 (FMIPv6) protocol is to allow a MN to configure a new CoA, before it moves and connects to a new network. It also allows the MN to use the new CoA immediately upon connecting to the new network. Furthermore, the FMIPv6 protocol seeks to eliminate the latency involved during the MN's BU procedure by providing a bi-directional tunnel between the old and new networks while the BU procedures are being performed. Thus, compared to normal MIPv6 operation, the FMIPv6 protocol claims to be more efficient in two respects:

- It eliminates IPv6 configuration delay introduced by Router Discovery, Address Configuration and DAD.
- It removes the delay introduced by the MN performing BU procedures with it's HA and CNs.

Technically speaking, the handover latency using FMIPv6 will not be any quicker than for MIPv6. That is, if we define handover latency to be the time between losing connectivity from one network to resuming existing communications on the new network using the New Care of Address (NCoA). This definition makes sense in MIPv6 as there is no way for the MN to resume communications until the NCoA has been registered with its HA. However, we have seen that FMIPv6 allows existing communications to continue throughout the entire handover process (assuming the New Access Router (NAR) has sufficient resources to buffer packets until the MN attaches to the new link). Theoretically, the only effect on existing traffic flows for a predicted handover will be the latency involved when packets are buffered at the NAR. Thus, if we define this as the real, effective handover delay it can be expressed as:

$$t_h = t_{connect} - t_{disconnect}$$

with the latency experienced by packets being:

$$t_h = t_{FNA} + t_{deliver}$$

where $t_{FNA}$ is the time it takes for Fast Neighbor Advertisement (FNA) to complete and $t_{deliver}$ is the time it takes for the NAR to deliver the packets to the MN.

Thus for movement that can be predicted, a FMIPv6 handover should not result in any lost packets and any jitter for real-time streams will be minimized to the time it takes to perform the movement at L2 (a good rule of thumb being ~50ms for re-association in 802.11 networks, although OS and driver specific timers can push this time up considerably).

Of course this only holds true for predicted handovers. For reactive handovers there remains the possibility of packet loss since from the moment the MN disconnects from the old network, the Previous Access Router (PAR) has nowhere to send packets destined for Previous Care of Address (PCoA) until it receives the Fast Binding Update (FBU) from the MN after it has arrived on the new network.

One of the open issues with the FMIPv6 is choosing when to tear down the bi-directional tunnel between the MN's PCoA and NCoA. Intuitively, this should be done once the MN has completed the MIPv6 BU procedure with all of its CNs. However, the current FMIPv6 specification does not provide any signaling exchange for the MN to inform the PAR that it can stop forwarding packets.

Thus, a soft state timer in the PAR set to a 'reasonable' value is the most likely solution that will be implemented.

### C. First Simulation Scenario

The Network Simulator, NS-2 [7] that supports for HMIPv6 which is ns-2.1b7a, was used for the evaluation of the protocols. The goal of this simulation is to examine and compare between HMIPv6 and MIPv6 in terms of handover latency. Handover Latency is defined for a receiving MN as the time that elapses between the last packets received via the old access router (oldAR) and the arrival of the first packet along the new access router (newAR) after a handover. Thus, this is the time during which the Mobile Node can neither receive, nor send IP traffic. Fig. 9 shows the network topology used for simulation experiment handover.

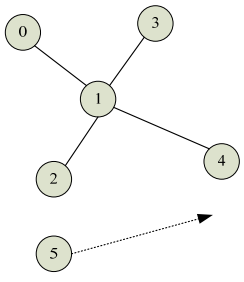



Figure 9. Simulation Network Topology.

The link characteristics namely the bandwidth (megabits/s) and the delay (milliseconds), are shown beside the link. The access routers are set to be 70 meters apart with free space in between. The wireless coverage area of the access router is approximately 40 meters. Finally, our model assumes a well-behaved mobile node movement pattern where the mobile node moves linearly from one access router to another at a constant speed of 1m/s. Table II explains the nodes topology for both protocols.

TABLE II.
MODE DESCRIPTION

| Node | Macro-mobility Protocol | |
|---|---|---|
| | ***HMIPv6*** | ***MIPv6*** |
| Node 1 | Mobility Anchor Point | -- |
| Node 2 | Old Access Router | Old Access Router |
| Node 3 | Home Agent | Home Agent |
| Node 4 | New Access Router | New Access Router |
| Node 5 | Mobile Node | Mobile Node |

### D. Results and Discusses of First Simulation

Fig. 10 shows the handover effect where it is evaluated based on the graph of cumulative sum of the packets sent from CN to MN versus time in seconds.

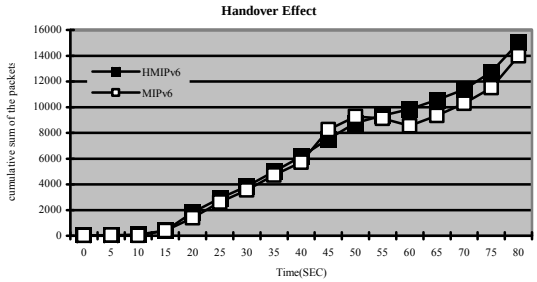



Figure 10. Handover latency in Mobile IPv6.

The following observations can be made about Fig. 10:

1) CN starts to communicate with MN by sending packets at 11s after it finishes its registration and all the setup links with HA and MN.

2) Then, at around 40s, packets lost/reordering begin to occur where at this moment, MN has moved to a new AR. However, this situation only happened in HMIPv6. In MIPv6, even though it supposes to start at the same time but due to the location of HA is quite further away from MN, thus, the delay will be increased. Since MN is always contact to HA in order to tunnel the packet from/to CN, then, it affects the movement of MN to the new AR. As a result of the packets lost/reordering, slow start activity can be observed thereafter.

3) After around t =45s for HMIPv6 and 49s for MIPv6, eventually, the transmission returns to normal. The overall handover latency, defined as the time when the MN detaches from the network at layer-2 till the disrupted communication session is returned to full operational state, is approximately 4500ms for HMIPv6 and 9000ms for MIPv6.

From the figure, we found that the time in HMIPv6 protocols between the last moment where the MN can receive and send packets through the old Access Router and the first moment where it can receive and send packets through the new Access Router is shorter compared to MIPv6.

### E. Second Simulation Setup

We use NS2 to simulate our proposed protocol. In our second simulation, the channel capacity of mobile hosts is set to the same value: 2 Mbps. We use the distributed

coordination function (DCF) of IEEE 802.11 for wireless LANs as the MAC layer protocol. It has the functionality to notify the network layer about link breakage. Table III summarizes the simulation settings.

For the simulation, we make use of Hierarchical Mobile IP (HMIP) implementation, which has implemented in Columbia IP Micro-mobility Software (CIMS) [7]. It supports micro mobility protocols for instance Hawaii, Cellular IP, and HMIP extension meant for the ns-2 network simulator based on version 2.1b6. We have additionally included MAP functionality to provide regional registration with the existing CIMS implementations.

TABLE III.
SIMULATION SETTINGS

| | |
|---|---|
| No. of Nodes | 15 |
| Area Size | 1000 X 1000 |
| Mac | 802.11 |
| Simulation Time | 50 sec |
| Traffic Source | CBR |
| Packet Size | 512 |
| Speed | 5,10,15,20 |
| Transmission range | 75m |
| Routing Protocol | AODV |

The simulation has carried out using the network topology shown in Fig. 11.

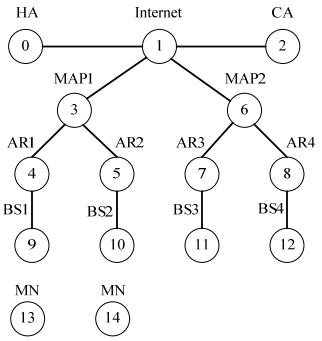



Figure 11. Network Topology.

Initially the mobile node MN13 was in MAP1 in the domain AR1. During the simulation we perform intra and inter domain handoff on MN13.

Initially, at time t1, the mobile node performs intra domain handoff by moving from AR1 to AR2 within MAP1. Next at time t2, it start moving towards AR3 from AR2, thus by performing inter domain handoff. At time t3, it moves from AR3 to AR4, within MAP2. Finally at time t4, it moves back to AR1, once again performing inter-domain handoff. In all the movements, PDAD is performed before getting the new CoA.

We evaluate the performance of our scheme based on the following parameters.

- *Handoff latency:* The handoff latency is defined as the time interval from last packet received form serving BS to and new packet received from target BS.
- *Packet loss:* The packet loss counts from the MS disconnecting to serving BS to receiving new packets from the target BS.

### *F. Second Simulation Results*

We move the mobile node to AR1, AR2, AR3 and AR4. Fig.12 shows the packet loss for HMIP handoff with PDAD and DAD schemes. We can see that the packets loss is less in PDAD based handoff.

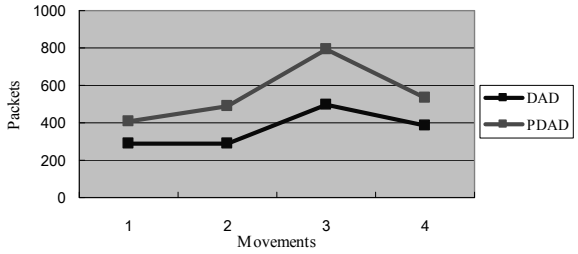



Figure 12. Movements vs. packet loss.

Fig.13 shows the handoff delay for PDAD and DAD based schemes. Clearly the handoff delay for PDAD is significantly less when compared with DAD.

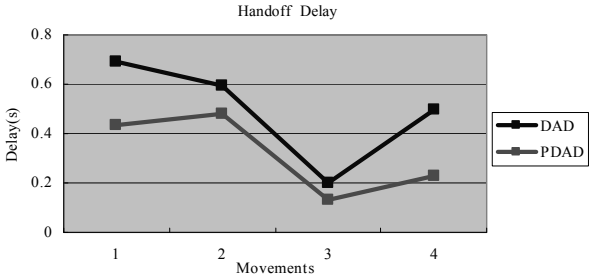



Figure 13. Movements Vs Handoff Delay.

Next we vary the speed of the MN as 5, 10, 15 and 20 m/s. From Fig.14, we can see that the packet loss is once again less in the case of PDAD scheme when compared with DAD.

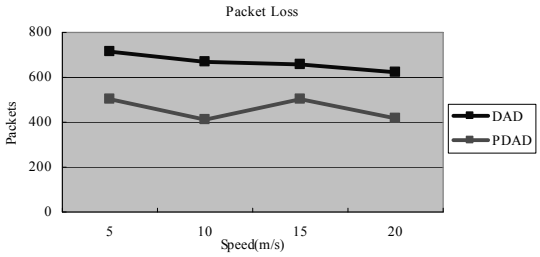



Figure 14. Speed vs. packet loss.

## V. CONCLUSION AND FUTURE WORKS

Mobile IPv6 is a key element of the future of All-IP wireless network to allow users to traverse freely between domains and still be connected to a service network. This paper presents a comparative study of intra-domain mobility between HMIPv6 and MIPv6. We have shown through the simulation that HMIPv6 perform best in terms handover latency compared to MIPv6. MIPv6 suffers longer handover latency because the time to send back the BU at new AR to the CN takes longer time. However, the handover latency for HMIPv6 is still quite large. Future work will have to design some scheme for

reducing handover delay and packet loss in hierarchical Mobile IPv6 Networks.

## ACKNOWLEDGMENT

The authors wish to thank Dr. Ming.Hu and Dr. Wu Xinyi from Wuhan Textile University, China. This work was supported in part by a grant from Ministry of Education, Hubei province, China.

## REFERENCES

[1] Joe. Inwhee, Lee. Hyojin, "An efficient inter-domain handover scheme with minimized latency for PMIPv6," International Conference on Computing, Networking and Communications (ICNC), pp. 332-336, March 2012.

[2] Sivakami. T., Shanmugavel. S., "An overview of mobility management and integration methods for heterogeneous networks," Third International Conference on Advanced Computing (ICoAC), pp.41-45, Dec. 2011.

[3] Bo Hu, Shanzhi Chen, Xiaoyan Jiang, "A Performance Evaluation of IP Mobility Support Protocols," Second International Conference on Multimedia and Information Technology (MMIT), pp. 17-20, April 2010.

[4] Xinyi Wu, Gang Nie, "Comparative Study and Performance Analysis of the Macro-mobility Protocol," Asia-Pacific Conference on Information Processing (APCIP), pp. 497-500, July 2009.

[5] Gang Nie, Xiuhua Qing, "Analysis and evaluation of an enhanced handover scheme in hierarchical mobile IPv6 networks," International Colloquium on Computing, Communication, Control, and Management (CCCM), pp.329-332, Aug. 2009.

[6] Xinyi Wu, Gang Nie, "Comparison of different mobility management schemes for reducing handover latency in Mobile IPv6," International Conference on Industrial Mechatronics and Automation (ICIMA), pp. 256-259, May 2009.

[7] L. Chen, X. Cai, R. Sofia, and Z. Huang, "A cross-layer fast handover scheme for mobile WiMAX," Proc. of Vehicular Technology Conference, Sep. 2007.

[8] H.-J. Jang, Y. H. Han, and S.-H. Hwang, "A cross-layering handover scheme for IPv6 mobile station over WiBro networks," Journal of KIISE, vol. 34, no. 1, Feb. 2007.

[9] T. Mahmoodi, V. Friderikos, O. Holland, and A. H. Aghvami, "Cross-layer design to improve wireless TCP performance with link-layer adaptation," In Proc. of Vehicular Technology Conference, Sep. 2007.

[10] H.S. Flarion, C.Castellucia, K. El-Malki, and L. Bellver, "Hierarchical Mobile IPv6 mobility management (HMIPv6)," IETF RFC 4140, August 2005.

[11] R. Koodly, "Fast Handovers for Mobile IPv6 (FMIPv6)," IETF RFC 4068, July 2005.

[12] D. Johnson, C. Perkins and J. Arkko, "Mobility Support in IPv6," RFC 3775, June 2004.

[13] Andrew T. Campbell et al., "Comparison of IP Micro-mobility Protocols," IEEE Wireless Communications Magazine, vol. 9, February 2002.

[14] Xavier P´erez-Costa and Hannes Hartenstein, "A Simulation Study on the Performance of Mobile IPv6 in a WLAN-Based Cellular Network," Computer Networks, special issue on", Towards a New Internet Architecture", September 2002.

[15] Xavier P´erez-Costa and Marc Torrent-Moreno, "A Performance Study of Hierarchical Mobile IPv6 from a System Perspective," Proceedings of IEEE International Conference on Communications (ICC), May 2003.

**NIE Gang** received his B.S. and M.S. degrees in computer engineering from South West University, China, in 1997 and 2002 respectively. Currently, he is an assistant professor of School of Mathematic & Computer Science at Wuhan Textile University, China, and leading the modern Network technology group. His research interests include mobile networking & computing, embedded systems & software.

# Radar Emitter Signal Analysis with Estimation of Distribution Algorithms

Haina Rong
School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China
Email: ronghaina@126.com

Jixiang Cheng and Yuquan Li
School of Electrical Engineering, Southwest Jiaotong University, Chengdu, China
Email: chengjixiang0106@126.com, wolf-quan@qq.com

***Abstract*—This paper proposes a novel approach (short for iEDA/TFAD) based on estimation of distribution algorithms and time-frequency atom decomposition for analyzing radar emitter signals. In iEDA/TFAD, an improved estimation of distribution algorithm combining Gaussian and Cauchy probability models is presented to implement time-frequency atom decomposition to analyze several typical radar emitter signals by extracting their features and recognizing them. The introduction of iEDA can greatly reduce the computational complexity of TFAD. Experimental results show that EDA/TFAD can efficiently recognize several radar emitter signals at a high correct rate.**

***Index Terms*—radar emitter signal, estimation of distribution algorithm, time-frequency atom decomposition, feature extraction, recognition**

## I. INTRODUCTION

Radar emitter signal recognition is to determine the identity of a radar emitter by using the signals intercepted in an electromagnetic environment. It is a very important step in an electronic reconnaissance system composed of electronic support measures, electronic intelligence and radar warning receiver sub-systems [1-3]. This step always provides a critical support to an electronic jamming system and its subsequent military operations. In conventional methods, five inter-pulse parameters, including pulse amplitude, pulse width, carrier frequency, direction of arrival and time of arrival, were used to recognize different radar emitter signals. As more and more advanced radars with various modulations are utilized in the modern electronic warfare, radar emitter signal recognition becomes a very challenging issue and attracts much attention of many countries. In recent years, some high-level recognition approaches, such as Wigner-Ville distribution-Radon transform [4], resemblance coefficient [2], Wigner and Choi–Williams time-frequency distributions [5], have been presented to recognize certain advanced radar emitter signals.

Time-frequency atom decomposition (TFAD) is the process that a subset of elementary components with good time and frequency resolution is selected from a redundant collection of waveforms to represent a signal in a linear superposition manner. As usual, the elementary component is called time-frequency atom and correspondingly the collection is called time- frequency atom dictionary. In the TFAD, the time- frequency atom dictionary is always redundant or over-complete and consequently the time-frequency atoms can be flexibly chosen [3, 6]. Thus, the TFAD becomes a good technique to analyze a non-stationary signal because it uses a certain number of time- frequency atoms in a flexible way to match the local structures of the signal. In recent years, the TFAD are receiving much attention, for example, it was successfully applied to analyze electric disturbance signals [7] and radar complex signals [8]. However, the computational complexity of TFAD is very high. How to reduce the computing time of TFAD is an ongoing issue.

In this paper, a fast algorithm based on an improved estimation of distribution algorithm (iEDA) combining Gaussian and Cauchy probability models is introduced to implement the TFAD. Through analyzing and processing parameters of the time-frequency atoms decomposed, three distinctive features reflecting the intra-pulse modulations of radar emitter signals are extracted to form a feature vector as inputs of support vector machine classifiers to distinguish the radar emitter signals with various signal-to-noise ratios and a wide range of modulation parameters. As compared with conventional time-frequency methods such as Wigner-Ville and Choi–Williams, the introduced approach has smaller computational complexity because it needs only a small number of atoms and avoids the time-consuming process of signal reconstruction. Extensive experiments verify the effectiveness of the proposed technique.

The organization of this paper is as follows. Section 2 describes iEDA as a fast implementation approach of TFAD. Section 3 presents feature extraction of radar emitter signals by using EDA/TFAD. Experimental results are provided in Section 4. Concluding remarks follow in Section 5.

## II. iEDA FOR TFAD

Corresponding author: Haina Rong.


doi: 10.4304/jnw.8.1.108-115

In this section, we will start from a brief introduction of the TFAD and a description of a pseudocode algorithm for implementing the TFAD by using a generic evolutionary algorithm. Then we will turn to iEDA.

*A. TFAD*

The TFAD is to select an appropriate countable subset of time-frequency atoms $g_\gamma(t)_{\gamma\in\Gamma}$ from a redundant time-frequency atom dictionary $\mathcal{D}=(g_\gamma(t))$ to expand a signal $f(t)$ into a linear sum of $g_\gamma(t)$ [6], i.e.,

$$f(t)=\sum_{h=1}^{+\infty}a_h g_{\gamma_h}(t), \tag{1}$$

where $a_h$ is the expansion coefficient of the atom $g_{\gamma_h}(t)$. The problem that the best atoms and their corresponding best expansion coefficients in (1) are found to optimally approximate a signal in a redundant time-frequency dictionary is NP-hard [9]. Matching pursuit, introduced in [6], could be one of the most successful methods to solve this problem.

Matching pursuit is an iterative algorithm and starts by projecting $f(t)$ on an atom $g_{\gamma_0}\in\mathcal{D}$ and computing the residual $Rf$:

$$f=\langle f,g_{\gamma_0}\rangle g_{\gamma_0}+Rf, \tag{2}$$

where $Rf$ is the residual vector after approximating $f$ in the direction of $g_{\gamma_0}$. By projecting the residual $Rf$ on an atom of $\mathcal{D}$, matching pursuit can subdecompose the residual $Rf$ sequentially. At the $h$th iteration, a best time-frequency atom $g_{\gamma_h}$ is searched from a time-frequency atom dictionary $\mathcal{D}$ to maximize the module $\left|\langle R^h f,g_{\gamma_h}\rangle\right|$, where $R^h f$ is the $h$th order residual of the signal $f(t)$ and $R^0 f=f$. Finally, the signal $f$ can be represented as

$$f=\sum_{h=0}^{H-1}\langle R^h f,g_{\gamma_h}\rangle g_{\gamma_h}+R^H f, \tag{3}$$

where $H$ is the maximal number of iteration. Since $R^H f$ is orthogonal to $g_{\gamma_h}$, the module of $f$ will be

$$\|f\|^2=\sum_{h=0}^{H-1}\left|\langle R^h f,g_{\gamma_h}\rangle g_{\gamma_h}\right|^2+\left\|R^H f\right\|^2, \tag{4}$$

where $\left\|R^H f\right\|$ converges exponentially to 0 when $H$ tends to infinity [6].

In [6], a greedy strategy was used to search the time-frequency atoms in a redundant dictionary. The greedy strategy tries to find the best time-frequency atom in an enumerative way at each iteration and hence is a local and time-consuming method. To reduce the computational load, we use an evolutionary algorithm to search the suboptimal time-frequency atom at each iteration from a redundant time-frequency atom dictionary, instead of the optimal one.

**Begin**

$R^0f=f$; $h=0$;

**While** (not termination condition) **do**

Set parameters of time-frequency atoms;

(*) Search the suboptimal time-frequency atom $g_{\gamma_h}$ in $\mathcal{D}$ using an evolutionary algorithm;

Compute $\left|\langle R^h f,g_{\gamma_h}\rangle g_{\gamma_h}\right|$;

$R^h f \leftarrow (R^h f-\langle R^h f,g_{\gamma_h}\rangle g_{\gamma_h})$;

$h\leftarrow h+1$;

**End**

**End**

Figure 1. Pseudocode algorithm for the TFAD based on an evolutionary algorithm [3].

The TFAD based on an evolutionary algorithm [3] is summarized as a pseudocode algorithm shown in Figure 1, where step (*) is the most important step. This paper uses iEDA, which will be described in the next subsection, as the approximate optimization algorithm.

*B. iEDA*

Since Estimation of Distribution Algorithms (EDAs) were proposed by Baluja in 1994 [10], EDAs quickly become an important branch of evolutionary algorithms because they have better mathematical foundation than other evolutionary algorithms. On the basis of statistical learning theory, EDAs use some individuals selected from the population at the current evolutionary generation to build a probability model and then produces offspring for the next generation by sampling the probability model in a probabilistic way. Many investigations in [10-13] show that EDAs have good optimization performance in both combinatorial problems and numeric optimization problems. As usual EDAs built one probability model to produce offspring.

This approach combines the advantages of a Gaussian probability model and a Cauchy probability model to produce a better search capability. As usual a Gaussian probability model is helpful for the local search capability and a Cauchy probability model is helpful for the global search capability. The flowchart of iEDA is shown in Figure 2, where each step is described in detail as follows.

Step 1 *Population initialization*: a population $P$ with $M$ individuals is uniformly generated in the feasible domain of a problem, i.e., $P_0=(\mathbf{x}_1,\cdots,\mathbf{x}_i,\cdots\mathbf{x}_M)$, where $\mathbf{x}_i=(x_{i1},\cdots x_{ik},\cdots x_{in})$, where $n$ is the dimension of a variable.

Step 2 *Fitness evaluation*: an evaluation function related to an optimization problem is applied to evaluate each individual and a fitness value is assigned to it.

Step 3 *Selection operation*: all individuals are sorted in a descending order by using their fitness values. The first $R$ individuals $\mathbf{x}^{se\_1}, \mathbf{x}^{se\_2}, \cdots, \mathbf{x}^{se\_R}$ with the best fitness values are selected.

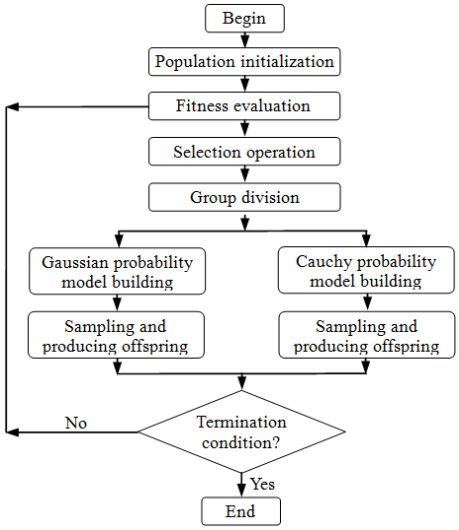



Figure 2. Flowchart of iEDA.

Step 4 *Group division*: the $R$ individuals $\mathbf{x}^{se\_1}$, $\mathbf{x}^{se\_2}, \cdots, \mathbf{x}^{se\_R}$ are divided into two groups. One group has $R \cdot \gamma$ individuals and the other has $R \cdot (1-\gamma)$ individuals.

Step 5 *Probability model building*: the individuals in the first and second groups are used to build a Gaussian probability model and a Cauchy probability model, respectively, by applying (5) and (6).

$$f_j(x) = N(\mu_j, \sigma_j), \tag{5}$$

where $\mu_j$ and $\sigma_j$ are the mean value and standard deviation of the *j*th dimension of individuals, respectively.

$$f(x_j) = \frac{1}{\pi}\left[\frac{\alpha}{(x-\theta)^2 + \alpha^2}\right], \tag{6}$$

where $\theta$ and $\alpha$ are two parameters representing the peak position and the scale that is a half of the peak value, respectively. As usual $\theta = 0, \alpha = 1$.

Step 6 *Sampling and producing offspring*: the Gaussian probability model built is sampled to produce $M \cdot \gamma$ individuals by using (7). The Cauchy probability model built is sampled to produce $M \cdot (1-\gamma)$ individuals by using (8).

$$\mathbf{x}^{new} = \mathbf{x}^{s} + E \cdot N(\mu, \sigma), \tag{7}$$

$$\mathbf{x}^{new} = \mathbf{x}^{s} + F \cdot C(0,1), \tag{8}$$

where $\mathbf{x}^{new}$ and $\mathbf{x}^{s}$ are the new and original individuals, respectively; $E$ and $F$ are extension rates of sampling parameters.

Step 7 *Termination condition*: a prescribed number of evolutionary generations are used as the termination condition.

## III. Feature Extraction Using iEDA/TFAD

This section will start from the analysis of time-frequency atoms, which are obtained from radar emitter signals by using the iEDA/TFAD, and then turn to the presentation of feature extraction. Three time-frequency atom features reflecting the intra-pulse modulations of radar emitter signals are extracted and they are a correlation ratio of the first Gabor atom (CrFG) to the original signal, a variance of center frequencies of Gabor atoms (VCFG) and a correlation ratio of the first Chirplet atom (CrFC) to the original signal, respectively.

First of all, we use Gabor time-frequency atoms to analyze radar emitter signals. The Gabor atom [6, 13], frequently applied in the literature, is defined as

$$g_{\gamma}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-u}{s}\right)\cos(vt + w), \tag{9}$$

where the index $\gamma = (s, u, v, w)$ is a set of parameters and they represent scale, translation, frequency and phase, respectively. $g(\cdot)$ is a Gauss-modulated window function described as

$$g(t) = e^{-\pi t^2}, \tag{10}$$

We employ the iEDA/TFAD to decompose five radar emitter signals into Gabor time-frequency atoms. The five signals are conventional radar emitter signals (CON), linear-frequency modulated radar emitter signals (LFM), non-linear-frequency modulated radar emitter signals (NLFM), binary phase shift-key radar emitter signals (BPSK) and binary frequency shift-key radar emitter signals (BFSK), respectively. The energy decaying rates (EDRs) of 20 time-frequency atoms of five radar emitter signals with 3 dB signal-to-noise rate (SNR) are shown in Figure 3. The EDR of a time- frequency atom $g_r$ is defined as

$$EDR(h) = \frac{\left| < R^h f, g_r > \right|}{\|f\|}, \tag{11}$$

where $f$ and $R^h f$ are the same as those in Eq. (3).

Figure 3 shows that the EDRs of the five radar emitter signals generally decrease as the number of time-frequency atoms increases and different signals almost have different values of EDRs. The several first time-frequency atoms have large EDRs, especially, CON and BPSK have very large EDRs at the first one time-frequency atom. This results from that the Gabor time-frequency atom is a signal with fixed frequency and hence easily matches such kind of signals with fixed carrier frequency as CON and BPSK. This observation inspires us to give the first feature, a correlation ratio of

the first Gabor atom $g_{\gamma_0}$ to the original signal *f*, which is defined as

$$P_1 = \frac{\left| < f, g_{\gamma_0} > \right|}{\| f \|}, \quad (12)$$

where $0 \le P_1 \le 1$.

As compared to the Gabor atom, a Chirplet time-frequency atom [10, 13] has one more parameter and described as

$$g'_{\gamma}(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) \cos(vt + \frac{c}{2}t^2 + w), \quad (13)$$

where the index $\gamma = (s, u, v, c, w)$ is a set of parameters representing scale, translation, frequency, modulation slope and phase, respectively. $g(\cdot)$ is a Gauss-modulated window function described in (10). We illustrate EDRs of 20 Chirplet time-frequency atoms of the five radar emitter signals with 3 dB SNR in Figure 4.

Figure 4 shows the similar trends to Figure 3, that is, the EDRs are decaying curves in general. It is noting that CON, LFM and BFSK have very large EDRs at the first one time-frequency atom. So we present the second feature, a correlation ratio of the first Chirplet atom $g'_{\gamma_0}$ to the original signal *f*, which is defined as

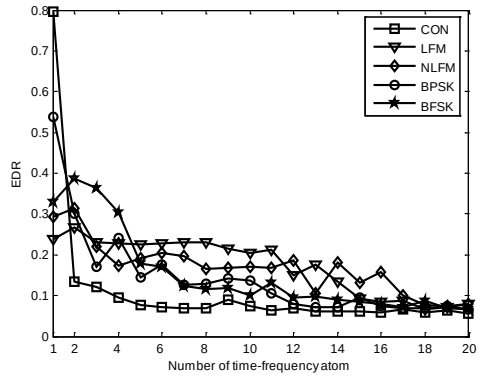



Figure 3. EDR as the number of Gabor time-frequency atoms

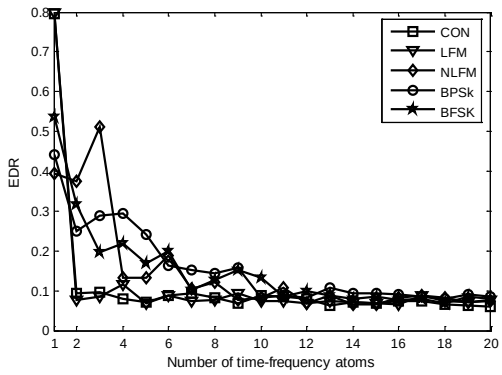



Figure 4. EDR as the number of Chirplet time-frequency atoms

$$P_2 = \frac{\left| < f, g'_{\gamma_0} > \right|}{\| f \|}, \quad (14)$$

where $0 \le P_2 \le 1$.

We go further to analyze the parameters of Gabor time-frequency atoms and observe that the normalized frequency parameter *v* and the translation parameter *u* can illustrate the modulation types of different radar emitter signals with a small number of atoms. Figure 5 shows the *u*-*v* graph of the five radar emitter signals.

We observe from Figure 5 that different radar emitter signals have distinctive frequency characteristics. Thus we obtain the last one feature, a variance of center frequencies of Gabor time-frequency atoms, which is described as

$$P_3 = \frac{1}{n-1} \sum_{k=1}^{n} (v'_k - \bar{v}')^2, \quad (15)$$

$$v'_k = |v_k| / 2\pi, \quad (16)$$

$$\bar{v}' = \frac{1}{n} \sum_{k=1}^{n} v'_k, \quad (17)$$

where $v'_k$ is the normalized frequency of the *k*th time-frequency atom $v_k$; $\bar{v}'$ is the center of frequencies $v'_1, v'_2, \cdots, v'_n$; *n* is the maximal number of Gabor time-frequency atoms which can reflect the modulations of radar emitter signals. In what follows we will discuss how to determine the value of *n* and consequently a new termination criterion for the TFAD algorithm shown in Figure 1 can be derived.

We also observe from Figure 5 that the presentation of frequency characteristics of different radar emitter signals needs different number of Gabor time- frequency atoms, for instance, CON uses only three atoms, while NLFM needs much more atoms. This observation exactly corresponds to the phenomena in Figure 3 that the EDRs of different radar emitter signals have different descending speeds as the number of Gabor time-frequency atoms increases. Thus, if we use a certain value $T_{EDR}$ of the EDR in (11) as a threshold to be a termination criterion of the

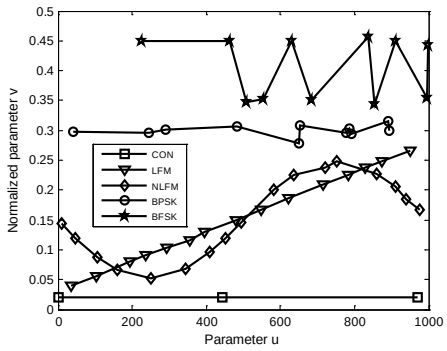



Figure 5. The *u*-*v* graph of five radar emitter signals

TFAD algorithm shown in Figure 1, we can obtain the appropriate number of Gabor atoms for each type of radar emitter signals, for example, three atoms can sketch the

frequency characteristic of CON in Figure 5, while in Figure 3, the EDR of CON rapidly decreases to a certain value. Actually the different numbers of Gabor atoms of the five radar emitter signals in Figure 5 are obtained by using the threshold $T_{EDR}$=0.1.

Finally, the three features above are utilized to construct a feature vector $P$=[$P_1$ $P_2$ $P_3$].

## IV. Experiments and Results

In this section, the parameter setting of iEDA is first discussed and then the performance of iEDA/TFAD is analyzed. Finally the effectiveness of iEDA/TFAD is tested by using five typical radar emitter signals.

### A. Parameter Setting

To investigate the setting of the parameter $\gamma$ in iEDA, we use the following three benchmark functions [14, 15]

$$F_{elp} = \sum_{i=1}^{n} i x_i^2 , \tag{18}$$

$$F_{sch} = \sum_{i=1}^{n} \left( \sum_{j=1}^{i} x_j \right)^2 , \tag{19}$$

$$F_{ros} = \sum_{i=1}^{n-1} (100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2) , \tag{20}$$

to conduct experiments. The three functions include single- and multi-modal test problems, which are widely used to check the performance of an optimization algorithm in the literature.

Let the parameter $\gamma$ varies from 0 to 1 with an interval 0.1, where the value 0 means that only the Cauchy probability model is applied in iEDA, and the value 1 means that only the Gaussian probability model is employed in iEDA. The parameters $M$ and $R$ are assigned as 200 and 100, respectively. The number $5\times10^5$ of function evaluations is used as the termination condition of the first problem. The 2nd and 3rd problems use the number of function evaluations as the termination conditions. The number of dimensions is 20. We record the mean of best solutions (short for Mean) and their standard deviation (short for Std) over 30 independent runs, which are listed in Table I.

Table I shows that as the parameter increases, the solutions of the three problems generally become better. When the values of the parameter are 0.9, 0.8 and 0.7, the quality of solutions is best. Thus the parameter could be assigned as 0.7-0.9.

### B. Performance Analysis of iEDA/TFAD

This section discusses the computational complexity of the iEDA/TFAD. Due to the introduction of an approximate optimization algorithm in the TFAD, we investigate the computational complexity in an experimental manner. A LFM radar emitter signal with various lengths is taken as an example to conduct the experiment. Two algorithms, iEDA and the greedy algorithm (GrA) [16] are employed to implement the TFAD. Figure 6 illustrates the average elapsed time of each time-frequency atom as the length of the signal increases. As shown in Figure 6, the greedy algorithm consumes much more time than the iEDA, which verifies the necessity of introducing approximate optimization algorithms into the TFAD. Actually the greedy algorithm has an exponential increase of time for solving the TFAD. We go further to provide the elapsed time of extracting three features, $P_1$, $P_2$, $P_3$, as the length of the signal goes up. The experimental results are shown in Figure 7. We can conjecture from these results that the computational complexity of the feature extraction may be . In [2], the time complexity of extracting resemblance coefficient features is not less than $O(n \log n)$. It is obvious that the time complexity of the presented approach is much lower than other time-frequency analysis techniques such as Wigner-Ville and Radon [4], and Wigner and Choi–Williams transforms [5].

TABLE I.
Experimental Results with Different $\gamma$

| $\gamma$ | $F_{elp}$ | $F_{sch}$ | $F_{ros}$ |
|---|---|---|---|
| | Mean ± Std | Mean ± Std | Mean ± Std |
| 0 | 4.88E+1 ± 1.31E+1 | 2.41E+3 ± 5.40E+2 | 5.26E+2 ± 7.13E+2 |
| 0.1 | 3.87E-1 ± 1.56E-1 | 7.23E+2 ± 2.57E+2 | 1.69E+1 ± 9.00E-1 |
| 0.2 | 1.56E-3 ± 1.29E-3 | 1.77E+2 ± 7.89E+1 | 1.38E+1 ± 4.89E-1 |
| 0.3 | 3.01E-6 ± 1.68E-6 | 3.86E+1 ± 2.17E+1 | 1.15E+1 ± 5.46E-1 |
| 0.4 | 1.47E-9 ± 1.31E-9 | 1.02E+1 ± 9.92E+0 | 9.48E+0 ± 6.09E-1 |
| 0.5 | 1.8E-13 ± 1.5E-13 | 8.66E-1 ± 1.18E+0 | 7.27E+0 ± 6.52E-1 |
| 0.6 | 4.3E-18 ± 4.5E-18 | 4.04E-2 ± 5.10E-2 | 5.81E+0 ± 6.76E-1 |
| 0.7 | 1.2E-23 ± 1.4E-23 | 6.56E-4 ± 1.08E-3 | 4.95E+0 ± 6.61E-1 |
| 0.8 | 3.4E-30 ± 4.5E-30 | 5.1E-13 ± 1.0E-12 | 5.71E+0 ± 9.97E-1 |
| 0.9 | 3.0E-45 ± 8.2E-45 | 5.74E-9 ± 2.19E-8 | 7.03E+0 ± 5.45E-1 |
| 1.0 | 1.9E-37 ± 5.1E-37 | 2.50E-6 ± 3.62E-6 | 9.44E+0 ± 8.37E-1 |

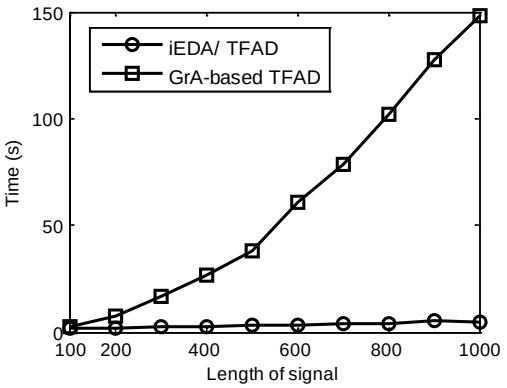
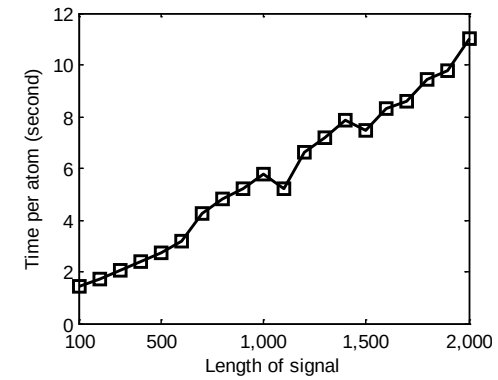
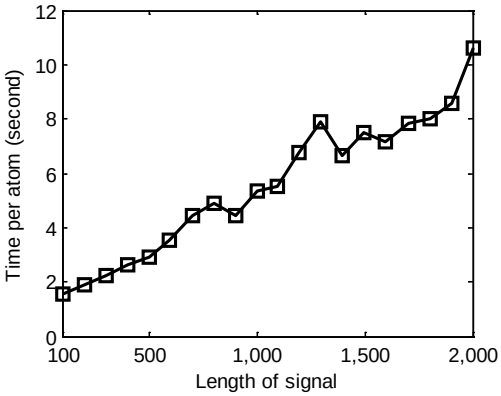


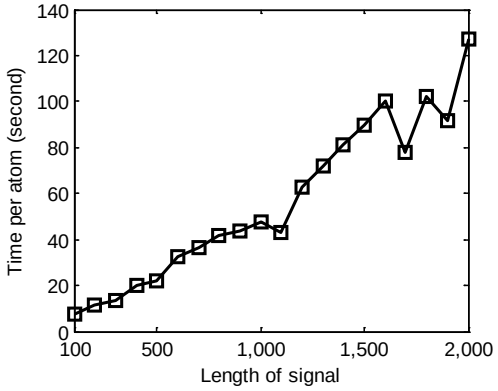

Figure 6. Comparisons of elapsed time




(a) Elapsed time of extracting feature CrFG




(b) Elapsed time of extracting feature CrFC

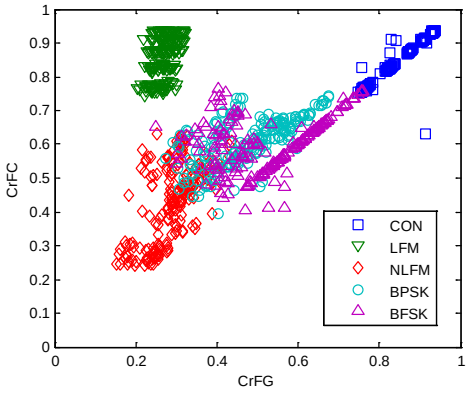



(c) Elapsed time of extracting feature VCFG

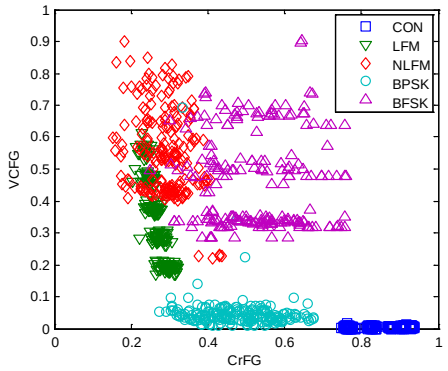
Figure 7. Elapsed time of feature extraction

### *C. Radar Emitter Signal Recognition*

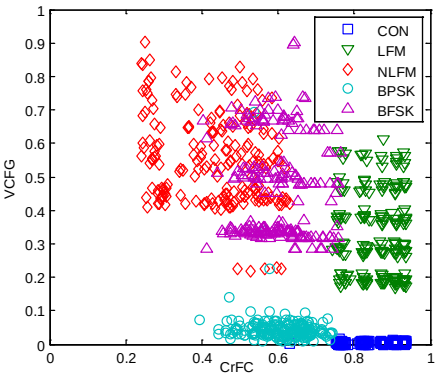
In this section, experiments conducted on the five types of radar emitter signals with various SNRs and a wide range of modulation parameters are used to test the effectiveness of the proposed features. The five types of radar emitter signals include CON, LFM, NLFM, BPSK and BFSK. We choose 40 radar emitter signals for each modulation type, i.e., 200 signals in total, to carry out the experiment. The pulse width of signals is 10 us and the sampling frequency is 100 MHz. The carrier frequencies of 40 CON signals vary from 6 MHz to 45 MHz. The starting frequencies and frequency slopes of LFM signals are chosen in the ranges [5 MHz, 35 MHz] and [5 MHz, 45 MHz], respectively. NLFM uses sinusoidal frequency modulated signals with 0.5, 1.0, 1.5 or 2 periods and the peak frequencies vary from 10 MHz to 45 MHz.



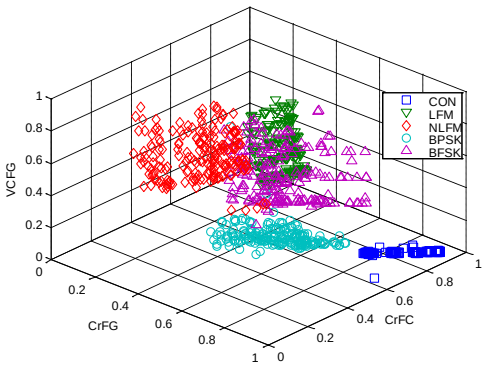

(a) Feature distribution of CrFG and CrFC




(b) Feature distribution of CrFG and VCFG




(c) Feature distribution of CrFC and VCFG




(d) Feature distribution of CrFG, CrFC and VCFG

Figure 8. Feature distributions of radar emitter signals

BPSK signals employ the frequencies ranged from 10 MHz to 45 MHz as carrier frequencies and 4, 5, 7, 11 and 13 Barker codes as code sequences. As for BFSK signals, we also use 4, 5, 7, 11 and 13 Barker codes and frequency differences varying from 10 MHz to 20 MHz.

As for each signal, SNR is chosen as 2 dB, 4 dB, …, 10 dB. Thus we generate 5 samples for each signal and therefore 200 samples for each modulation type of radar emitter signals. In this way 1000 samples for the five types of radar emitter signals are obtained to be training set of classifiers. The testing samples are gained by using 9 SNRs varying from 2 dB to 10 dB, that is, each signal has 9 testing samples and hence 200 signals have 1800 testing samples in total. We use the proposed approach to extract three features from each radar emitter signal. Figure 8 illustrates the feature distributions in two-dimensional and three- dimensional spaces. In this figure, all the features are from the training samples and the VCFG is normalized.

The feature distribution graphs shown in Figure 8 provide several indications. As shown in Figure 8 (b) and (c), the VCFG values of CON signals are close to zero, and the CrFG and CrFC of them mainly depend on the SNR and are not relative to the carrier frequencies. It can be seen from Figure 8 (a) and (b) that the CrFG and VCFG of LFM signals principally depend on the values of frequency slopes and that the CrFC is directly relative to the SNR, being independent of parameters of LFM signals. NLFM signals have smaller values of CrFG and CrFC, while their larger VCFG is dependent on non-linearity of frequency modulation. BPSK signals have fixed carrier frequency and therefore have larger CrFC and CrFG and their changing phases resulting in corresponding frequency fluctuations cause a relatively small values of VCFG, which is implied in Figure 8 (b) and (c). BFSK signals may be viewed as a combination of two CON signals and have larger CrFG and CrFC, which is similar to CON signals. The VCFG of BFSK signals depends on the difference between two frequencies and hence has a strip of features.

To quantitatively evaluate the performances of the three features, we use support vector machines (SVMs) to design classifiers. SVM, developed principally by Vapnik [17], provides a novel means of classification using the principles of structure risk minimization. The subject of SVM covers emerging techniques that have been proven to be successful in many traditional neural network-dominated applications [17-19]. A SVM is primarily designed for binary classification problems. Several binary SVMs are usually combined to solve the multiclass classification problem. Among several approaches for combining binary-class SVMs, the directed acyclic graph (DAG) [18, 19] is verified to be a valid and practical way for solving muticlass classification problems. So we use DAG to classify radar emitter signals.

In the DAG, for an *M*-class classification problem, *M* (*M*-1) /2 hyperplanes need be constructed to separate each class from each other and some decision-making scheme is used. In the training phase, DAG need construct *M* (*M*-1) /2 binary-class SVMs, in which each one for each pair of classes. In the testing phase, the architecture shown in Figure 9 is used to combine *M* (*M*-1) /2 binary-class SVMs. In this figure, five classes are taken for an example.

In the experiments, SVMs use Gaussian kernel function. The 1000 training samples (mentioned above) of radar emitter signals are applied to train the classifiers designed by using DAG to combine 10 binary SVMs. The 1800 testing samples are utilized to test the SVM classifiers trained. Experimental results are listed in Table 3.

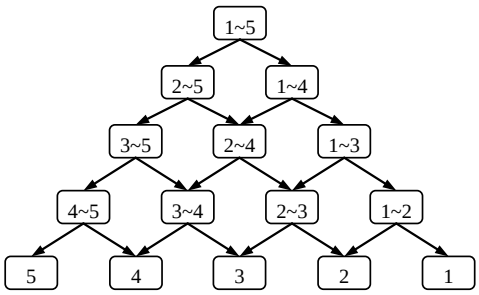



Figure 9. Structure of DAG

TABLE II.
RECOGNITION ERROR RATES (RER) OF RADAR EMITTER SIGNALS

| Types | CW | LFM | NLFM | BPSK | BFSK |
|---|---|---|---|---|---|
| RER | 0.0 | 0.0 | 2.3 % | 1.3 % | 3.5 % |

Table II shows that the proposed method can obtain a low recognition error rate for each of the five typical radar emitter signals and the average of recognition error rates is only 1.72%, which verifies the effectiveness of the features again and further the introduced feature extraction classification algorithms.

## V. CONCLUSIONS

This paper analyzes radar emitter signals by using an improved estimation of distribution algorithm and time-frequency atom decomposition. iEDA has the advantages of both Gaussian and Cauchy probability models and can reduce the complexity of time-frequency atom decomposition of radar emitter signals. On the basis of analyzing time-frequency atoms obtained from radar emitter signals, three features, CrFG, CrFG and VCFG, are presented to construct an input vector of SVM classifiers to implement automatic recognition of signals. Extensive experiments verify the success of the application. Our further work will focus on how to use the TFAD to recognize the unintentional modulations underlying the radar emitter signals.

## ACKNOWLEDGMENT

The authors wish to thank G. Zhang for his valuable comments for improving this paper. This work was supported in part by the National Natural Science Foundation of China (61170016) and the Fundamental Research Funds for the Central Universities (SWJTU11BR173).

## REFERENCES


[1] E. Granger, M. A. Rubin, S. Grossberg, and P. Lavoie, “A what-and-where fusion neural network for recognition and tracking of multiple radar emitters, “ Neural Networks, vol. 14, no. 3, pp. 325-344, 2001.

[2] G. X. Zhang, W. D. Jin, and L. Z. Hu, “Resemblance coefficient based intrapulse feature extraction approach for radar emitter signals, " Chinese Journal of Electronics, vol. 14, no. 2, pp. 337-341, Apr 2005.

[3] G. X. Zhang, "Time-frequency atom decomposition with quantum-inspired evolutionary algorithms, “ Circuits, Systems and Signal Processing, vol. 29, no. 2, pp. 209-233, Apr. 2010.

[4] T. O. Gulum, P. E. Pace, and R. Cristi, "Extraction of polyphase radar modulation parameters using a Wigner-Ville distribution-Radon transform, " in Proceedings of IEEE International Conference on Acoustics, Speech and Signal, pp. 1505-1508, 2008.

[5] J. Lunden and V. Koivunen, "Automatic radar waveform recognition, " IEEE Journal of Selected Topics in Signal Processing, vol. 1, no. 1, pp. 124-136, 2007.

[6] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries, " IEEE Transactions on Signal Processing, vol. 41, no. 12, pp. 3397-3415, 1993.

[7] M. P. Tcheou, L. Lovisolo, E. A. B. da Silva, M. A. M. Rodrigues, and P. S. R. Diniz, "Optimum rate-distortion dictionary selection for compression of atomic decompositions of electric disturbance signals, " IEEE Transactions on Signal Processing Letters, vol. 14, no. 2, pp. 81-84, 2007.

[8] G. Lopez-Risueno, J. Grajal, and O. Yeste-Ojeda, "Atomic decomposition-based radar complex signal interception, " IEE Proceedings-Radar Sonar Navigation, vol. 150, no. 4, pp. 323-331, 2003.

[9] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximation, " Journal of Constructive Approximation, vol. 13, no. 1, pp. 57-98, 1997.

[10] Baluja, S., “Population based incremental learning: A method for integrating genetic search based function optimization and competitive learning”, Technical Report, No. CMU-CS-94-163, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1994.

[11] Mendiburu, A., Miguel-Alonso, J., Lozano, J.A., “A Review on parallel estimation of distribution algorithms”, Studies in Computational Intelligence, vol.269, pp. 143-163, 2010.

[12] Pelikan, M., Goldberg, D.E., Lobo, F.G., “A survey of optimization by building and using probabilistic models”, Computational Optimization and Applications, vol. 21, no. 1, pp. 5-20, 2002.

[13] J.X. Cheng, G.X. Zhang, C.Z. Tang, "A novel approach of feature extraction for advanced radar emitter signals using time-frequency atom decomposition, " Journal of Xi’an Jiaotong University, vol. 44, no. 4, pp. 108-113, 2010.

[14] G.X., Zhang, “Quantum-inspired evolutionary algorithms: a survey and empirical study”, Journal of Heuristics, vol. 17, no. 3, pp. 303-351, 2011.

[15] J.X., Cheng, G.X., Zhang, X.X., Zeng, “A novel membrane algorithm based on differential evolution for numerical optimization”, International Journal of Unconventional Computing, vol. 7, no. 3, pp. 159-183, 2011.

[16] T. L. Carroll, "A nonlinear dynamics method for signal identification, " Chaos, vol. 17, no. 2, p. 023109, 2007.

[17] V. N. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.

[18] G. X. Zhang, "Support vector machines with Huffman tree architecture for multiclass classification, " in Lecture Notes in Computer Science. vol. 3773, pp. 24-33, 2005.

[19] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAG's for Multiclass Classification, " in Advances in Neural Information Processing Systems. vol. 12 Cambridge, MA: MIT Press, pp. 547-553, 2000.


**Haina Rong** was born in 1980, received B.S. degree in 2002 from Southwest University of Science and Technology; received M.S. and Ph.D. degrees in 2006 and 2010, respectively, from Southwest Jiaotong University, in the area of intelligent information processing. His major interests include natural computing, signal processing and intelligent information processing. She is a lecturer at the School of Electrical Engineering of Southwest Jiaotong University.

# Time Synchronization for Mobile Underwater Sensor Networks

Ying Guo
Qingdao University of Science & Technology, Qingdao, China
Email: qd.guoguo@gmail.com

Yutao Liu
Qingdao Yawei Software Co., Ltd, Qingdao, China
Email: ziyou2150@sina.com

***Abstract*—Time synchronization is very crucial for the implementation of energy constricted underwater wireless sensor networks (UWSN). The purpose of this paper is to present a time synchronization algorithm which is suitable to UWSN. Although several time synchronization protocols have been developed, most of them tend to break down when implemented on mobile underwater sensor networks. In this paper, we analyze the effect of node mobility, and propose a Mobile Counteracted Time Synchronization approach, called "Mc-Sync", which is a novel time synchronization scheme for mobile underwater acoustic sensor networks. It makes use of two mobile reference nodes to counteract the effect of node mobility. We also analyze and design the optimized trajectories of the two mobile reference nodes in underwater environment. We show through analysis and simulation that Mc-Sync provides much better performance than existing schemes.**

***Index Terms*—time synchronization, mobile reference node, optimized trajectory, node mobility, underwater sensor networks**

## I. INTRODUCTION

Time synchronization is an important issue for many distributed applications, especially for sensor networks [1]. It requires collaboratively processing of time sensitive data in the applications of environment supervision, target tracking and so on [2][3]. Recently many time synchronization protocols for terrestrial wireless sensor networks have been proposed [4][5][6], which provide a high degree of precision.

As terrestrial communication is based on Radio Frequency (RF) technology, all of these mechanisms assume that propagation latency is negligible and can be effectively factored out of design consideration. While underwater communication mainly uses acoustic communication technology. There are several fundamental differences between RF communication and acoustic communication, such as large propagation delay and node mobility [7][8]. Thus protocols for terrestrial sensor networks could not be directly applied to underwater acoustic sensor networks.

Underwater sensor networks (UWSN) inherit many different features. Node mobility is one of the most important effects [9]. From empirical observation, nodes without any self-propelling capability can move with wind and ocean current typically at the rate of 0.83-1.67m/s, and existing Autonomous Underwater Vehicles (AUV) typically move at a rate of up to 2.9m/s. In addition, node movement also brings Doppler shift [10][11], which adds the difficulty to estimate propagation delay exactly. Thus time synchronization algorithm must be able to cope with sensor node mobility.

Moreover, UWSN has long propagation delay, limited bandwidth, limited transmission rate, high bit error rate and so on [12]. The propagation speed of underwater acoustic channel is five orders slower than radio waves, thus resulting in significantly longer propagation delay [13], which makes relatively large Doppler Effect and inter symbol interference. Due to strong attenuation in high frequency band, acoustic communication has limited available bandwidth and low carrier frequency. Because of low cost hardware and limited power supply, most of sensor nodes have limited transmission rate. In addition, limited bandwidth and transmission rate make it suffers from high bit error [14]. All of these make terrestrial time synchronization different from underwater.

Recently, several underwater sensor network time synchronization algorithms have been proposed. TSHL [15] is the first time synchronization algorithm designed for high latency networks specifically. It uses one-way communication to estimate the clock skew and two-ways communication to estimate the clock offset. MU-Sync [10] runs two times of linear regression to estimate the clock skew and clock offset for cluster based UWSN. TSHL assumption of constant propagation delay, thus it cannot handle mobility issues. MU-Sync assumes that the one-way propagation delay can be estimated as the average round trip time which causes extra errors, and has relative high message overhead.

Mobi-Sync [7] and D-sync [8] consider more about node mobility. Mobi-Sync need the help of surface buoys with GPS and synchronized super nodes. Ordinary nodes use a correlation model to estimate their velocity and launch time synchronization. In order to improve time synchronization precision, D-sync exploits Doppler shift to provide an indication of the relative motion between


doi: 10.4304/jnw.8.1.116-123

nodes. But they require specialty deployment or depend on the precision of velocity measurement, which is difficulty to implement in underwater environment.

In this paper, we propose a mobile counteracted time synchronization algorithm for underwater acoustic sensor networks, called "Mc-Sync". As UWSN experiences node mobility, our design utilizes two mobile reference nodes to eliminate the bad effect. We also design the trajectories of the two mobile reference nodes to guarantee the synchronization precision. Major contributions of this paper are as follows:

- We analyze the effect of node mobility to the clock skew, and design a novel time synchronization technique, called "Mc-Sync", which exploits two mobile reference nodes to counteract the effect of node mobility.
- We design trajectories of the two mobile reference nodes in underwater sensor networks. It could compensate the effect of node mobility and improve time synchronization precision.
- We examine the performance of Mc-Sync carefully. Simulations results show that Mc-Sync performs better than other existing algorithms.

The rest of this paper is organized as follows. In Section II, we introduce the notion and error sources of time synchronization. In Section III, we analyze the effect of node mobility to the clock skew. Then provide Mc-Sync algorithm in details and design the trajectories of the two mobile reference nodes. We present simulation results and related work in Section IV and V. We offer our conclusion and future work in Section VI.

## II. BACK GROUND

The sensor node time is controlled by an internal clock, which is composed of a crystal oscillator and counter. The internal clock is updated with the frequency of the crystal oscillator, different hardware have different crystal oscillator frequency causing un-synchronization among node clocks. In general, we often model the local time of node i using two parameters, namely, clock skew and clock offset, as follows:

$$T_i(t) = a_i(t) + b_i$$

Where ai is the clock skew, bi is the clock offset, and T is the ideal time or Universal Time Coordinated (UTC).

The offset arises when the two sensor nodes have a different starting time. It causes constant error independent of time. The clock skew causes increasing error as time goes by. Thus, the synchronization algorithm must be able to estimate both the clock skew and offset.

A description of causes of synchronization error was first described by Koeptz and Schwabl, and extended by Horauer et. al. recently [16], in which incorporates physical layer jitter that cannot be over looked for high precision time synchronization. The sources of error could be summarized as below: (1) Send Time, (2) Access Time, (3) Transmission and Reception Time, (4) Propagation Time, and (5) Receive Time.

Moreover, the clock is also affected by the interaction of other components of the sensor system and underwater environment, for example long propagation delay, sensor node mobility, temperature, pressure, battery voltage and so on. Time synchronization schemes design in UWSN have to focus on eliminating or accounting for these sources of error and limiting energy consumption.

For mobile underwater sensor network, node mobility is one of the most important factors on time synchronization. In order to reduce its effect and improve synchronization precision, we analyze its influence in the following section, and design a new time synchronization method for the application of UWSN.

## III. ALGORITHM DESIGN

In this section, we analyze the effect of node mobility to the clock skew, and present design details of Mc-Sync, which makes use of two mobile reference nodes to counteract the impact of node mobility. After that, we introduce the trajectories of the two mobile reference nodes.

### *A. Effect of Node Mobility*

In order to compute out the clock skew, making use of one-way information exchange twice is the simplest method. In this process, we assume that the propagation delays for each information exchange are the same. As shown in figure 1, reference node A with standard time is fixed, and node B is a sensor node to be synchronized. Node A sends synchronization information to node B. The message contains sending time T1, which is a time stamp of MAC layer. Node B receives this message and records the information receiving time T2 with its local time. Then node A sends synchronization message again, which contains the sending time of MAC layer time stamp T3, node B receives this message and records the local receiving time T4.

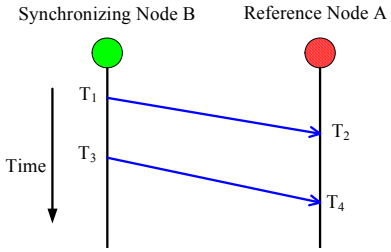



Figure 1. Compute the clock skew.

Assuming that every information propagation process costs the same time t, we could get

$$T_2 - t = aT_1 + b \,. \tag{1}$$

$$T_4 - t = aT_3 + b \,. \tag{2}$$

In which, a is the clock skew, b is the clock offset.

From equation (1) and (2), we could obtain:

$$a = \frac{T_2 - T_4}{T_1 - T_3} \tag{3}$$

Ideally, the clock skew is equal to the ratio of $T_2 - T_4$ and $T_1 - T_3$. However, in actual situation, a

computed by this method is not only the clock skew, but the clock skew with time variety caused by node mobility. Due to long underwater propagation delay, effect of the clock skew is much less than the node mobility in the time synchronization process.

We give an example to illustrate above conclusion. Assume a equal to 1.00001, and the sensor node to be synchronized moves along the direction away from reference node. The node moves with maximum speed 2.9m/s. The time interval between two successive reference packets is 1s. The average propagation speed in the simulated environment is 1500m/s with 1% fluctuates. Use equation (3), we could calculate that a is equal to 0.99982. That is to say, time synchronization makes the clock deviation bigger than before.

In order to solve this problem, and improve calculating precision of a, lots of existing algorithms adopt linear regression, which cannot eliminate the influence of node mobility essentially. Some algorithms involved node position, speed and other information to computation process, but most of them require complex deployment and hard calculation. In order to reduce the effect of node mobility, we use two reference nodes to get exact clock skew.

*B. Mc-Sync Design*

As in figure 2, Node A1 and A2 are reference nodes with standard time, which located at opposite sides of node B and keep still. Node B is the node to be synchronized, which floats with ocean current.

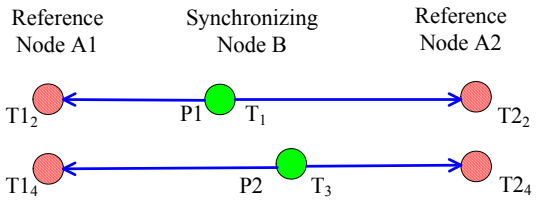



Figure 2. Effect of node mobility.

As in figure 2, Node B sends synchronization information at its local time T1, node A1 and A2 receive this information at time T12 and T22 respectively. We assume all the time used in time synchronization process is MAC layer time stamp. Node B sends synchronization information at its local time T3. Node A1 and A2 receive this information at T14 and T24 respectively. Due to the mobility of node B, it's information sending position changes from P1 to P2.

Adopt equation (3) to calculation a. There is a difference $\Delta$T between actual (T12-T14) / (T22-T24) and computed value, as the result of node mobility. The difference of A1 and A2 has the same value and opposite direction, as in figure 3.

For node A1, we could obtain:

$$a = \frac{T1_2 - T1_4 - \Delta T}{T_1 - T_3}. \quad (4)$$

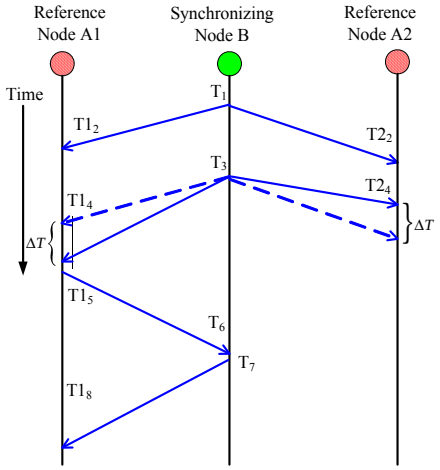



Figure 3. Mobile counteracted time synchronization.

In common, for node A2, we could obtain:

$$a = \frac{T2_2 - T2_4 + \Delta T}{T_1 - T_3}. \quad (5)$$

From equation (4) and (5), we could compute out exact clock as follow:

$$a = \frac{T1_2 - T1_4 + T2_2 - T2_4}{2(T_1 - T_3)}. \quad (6)$$

Then we use two-way information exchange to estimate the clock offset, as in figure 3. Node A1 transmits message to node B, when it received the second synchronization information. The massage includes the MAC layer time stamp T15. Node B receives this packet, records the receiving local time T6 and immediately sends synchronization information at time T7. Node A1 receives this message and records the receiving time T18. As node B immediately replies to node A1, the time interval is very short, node B is barely moving. So there is nearly no error of the clock offset caused by node mobility. We could get:

$$T1_5 + t = aT_6 + b. \quad (7)$$

$$T1_8 - t = aT_7 + b. \quad (8)$$

From equation (7) and (8), we could compute out the clock offset:

$$b = \frac{T1_8 + T1_5 - a(T_7 + T_6)}{2}. \quad (9)$$

As discussed above, we could get node's clock skew and offset from equation (6) and (9). We design the algorithm of reference nodes as follow:

```
Mc-Sync for reference nodes
1: program Ref_node ()
2:  while(time&&receive!=1)
3:   if (receive==1)        //first message exchange
4:    while(time&&receive!=2)
5:     if (receive==2)     //second message exchange
6:      send(sync);        //third message exchange
7:      while(time&&receive!=3)
8:       if (receive==3)   //forth message exchange
9:         compute (a,b);  //time synchronization
10:        broadcast(a,b);
11:       endif
12:     endwhile
13:    endif
14:   endwhile
15:  endif
16: endwhile
17: end
```

Figure 4. Algorithm of reference nodes.

### C. Trajectory of Reference Node

In order to implement Mc-Sync algorithm, two requirements have to be satisfied: (1) Assume that the node movement is mainly effected by ocean current, the two reference nodes should be deployed along the direction of ocean current. (2) The node should be located at the connecting line of the two reference nodes.

For the first requirement, we design new equipment composed by two reference nodes. One of them is a mobile reference node, it could move autonomously. Another reference node cannot move by itself, which is joined to the mobile reference node by light cable. It could move passively with the mobile reference node and ocean current.

After deployed, the mobile reference node keep still. Another reference node moves with ocean current, and the light cable always be tensed. We assume node movement is mainly effected by ocean current, thus the light cable of the two reference nodes parallels to the direction of ocean current. That is to say, the direction of the node motion is the same as the connecting line of the two reference nodes.

To satisfy the second requirement, the two reference nodes should be deployed at the opposite sides of deployment area, and their trajectories are perpendicular to the direction of ocean current, as shown in figure 5. Reference nodes move a little distance D along the trajectories, and keep still to do time synchronization. Reference nodes repeat this process until all the nodes have been synchronized.

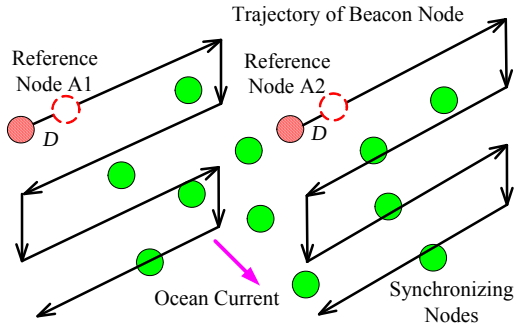



Figure 5. Trajectory of reference nodes.

In the time synchronization process, the node may receive more than one pair of information came from reference nodes. We use equation (3) to determine the most suitable positions of reference nodes.

The two reference nodes use equation (3) to calculate a respectively, the result that one a is greater than 1 and another one is less than 1 is retained, and others are invalid. Then we find out the most suitable reference node positions for Mc-Sync, which is the nearest location of the node to the reference nodes connecting line. The node computes out the sum distance from it to the two reference nodes in every time synchronization process. And compare these sum distances, the position with smallest distance value is the most suitable positions of reference nodes.

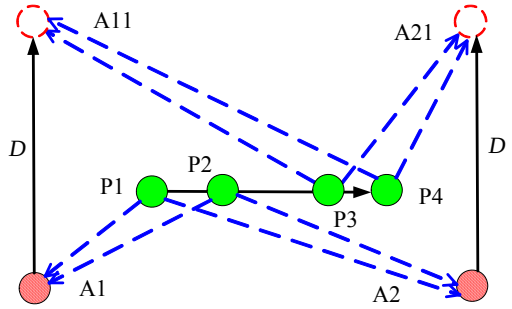



Figure 6. Error analysis.

As in figure 6, reference node A1 and A2 at position A1 and A2, node P broadcasts synchronization messages at position P1 and P2 respectively. Reference node A1 and A2 receive those messages, and then move to position A11 and A22. After that, node P broadcasts synchronization messages at position P3 and P4, which be received by reference node A1 and A2. They could compute out the distance from them to the node, and send the result to node P. Node P computes out the sum distance from it to the reference nodes at different position:

$$d_1 = d_{P1A1} + d_{P1A2} + d_{P2A1} + d_{P2A21}$$

$$d_2 = d_{P3A11} + d_{P3A21} + d_{P4A11} + d_{P4A21}$$

Then it compares d1 and d2, and chooses the minimum one of them to do time synchronization. As d1 < d2, node P use the data of positions A1 and A2.

Error exists when the node and reference nodes are no collinear, as in figure 6. The time cost by node mobility for reference nodes A1 and A2 are:

$$\Delta T1 = \frac{P2A1 - P1A1}{V}. \tag{10}$$

$$\Delta T2 = \frac{P2A2 - P1A2}{V}. \tag{11}$$

In which V is the speed of sensor node. When A1, P1, P2, A2 collinear, $\Delta T1 = -\Delta T2$.

As D is very little and $A1A2 >> P1P2$, we could get the conclusion that $\Delta T1 \approx -\Delta T2$. So equation (6) could be still used. Shorten D will cause higher precision, but longer synchronization time and energy cost.

The algorithm of synchronizing nodes is as follow:

```
Mc-Sync for synchronizing nodes
1: program Sync_node ()
2:  broadcast(sync);
3:  wait;                    //first message exchange
4:  broadcast(sync);
5:  wait;                    //second message exchange
6:  while(time&&receive==0)
7:   if (receive!=0)         // third message exchange
8:    broadcast(sync);   //forth message exchange
9:    while(time&&receive_result==0)
10:     if (receive_result!=0)  //time synchronization
11:      compare result;
12:      compute sum distance;
13:      compare sum distance;
14:      time synchronization;
15:     endif
16:    endwhile
17:  endif
18: endwhile
19: end.
```

Figure 7. Algorithm of synchronizing nodes.

## IV. SIMULATIONS

We analyze the preferences of Mc-Sync via simulations, and compare it with No-Sync, TSHL [15] and MU-Sync [10] in this section.

### A. Simulations Setup

In our simulations, we assume two reference nodes have standard time, which move along the trajectories setting in advance. The sensor nodes to be synchronized have their own inside clocks, and move randomly in the deployment area. The parameters used in simulations are as follow:

The initial clock skew is 8ppm.

The initial clock offset is 10ppm.

The maximum speed of sensor nodes to be synchronized is 2.9m/s (Vmax).

These nodes change their speed randomly within the range of [0, Vmax].

The time interval between two successive reference packets is 1s.

Clock granularity is 1μs.

Receive jitter is 1μs.

The propagation speed in simulated environment is 1500m/s with 1% fluctuates.

The number of reference packets used by TSHL and MU-Sync to perform linear regression is 20.

The time stamps in simulations are the time stamp of MAC layer.

### B. Result and Analyse

In the first simulation, we research on the error with time elapsed since synchronization and compare Mc-Sync with different algorithm, e.g. Mc-Sync, MU-Sync, TSHL and No-Sync. As in figure 8, with time goes by, errors grow after time synchronization. As the clock skew leads to error increasing after synchronization, this comparison result actually demonstrates different accuracies on the clock skew these algorithms can achieve. In the simulation, the error for Mc-Sync is 27.4% of TSHL, and 60.9% of MU-Sync.

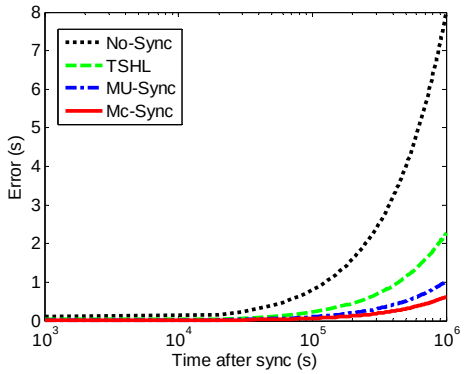



Figure 8. Error vs Time since synchronization.

The high precision of Mc-Sync arises from it takes node mobility into consideration. It utilizes two reference nodes to counteract the error caused by node mobility in time synchronization process. As a result, Mc-Sync performs better than No-Sync, TSHL and MU-Sync significantly.

The comparison of synchronization errors are listed in table 1. Both the average error and variance error of Mc-Sync are less than TSHL and MU-Sync.

TABLE I.
PREFERENCE COMPARISON

| Algorithm | TSHL | MU-Sync | Mc-Sync |
|---|---|---|---|
| Average error (s) | 1.15 | 0.52 | 0.31 |
| Variance error (s) | 0.61 | 0.16 | 0.07 |

Then we study the energy cost of these algorithms. In Mc-Sync, reference nodes do not always broadcast synchronization messages, which is not energy efficient. They only start the synchronization process when they arrive at the right positions. Figure 9 shows the number of packets needed in synchronization process, which represents the energy cost of TSHL, MU-Sync and Mc-Sync. As in figure 9, the packet number of Mc-Sync is the smallest compare with TSHL and MU-Sync.

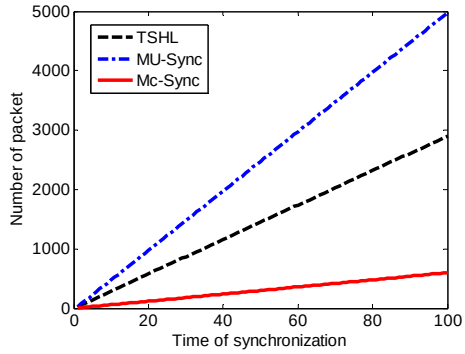



Figure 9. Number of packet.

MU-Sync and TSHL have higher packet number as they run linear regression to estimate the clock skew and clock offset. Mc-Sync only adopts five times of message exchange, and does not use linear regression, which reduces the energy cost greatly. As we can see, the packet numbers of Mc-Sync is 20.7% of TSHL, and 12.1% of MU-Sync.

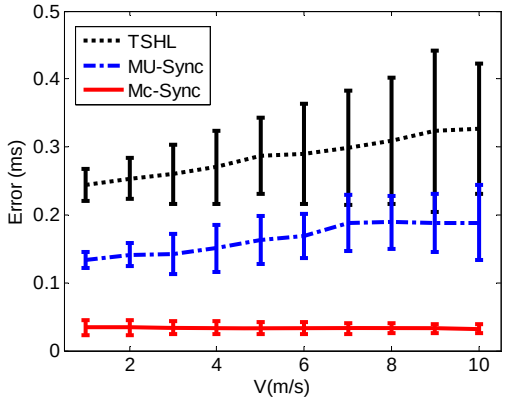



Figure 10. Effect of velocity.

As node mobility is one of the main effects to time synchronization, we also discuss the effect of node velocity in our simulation. We change the node speed from 0m/s to 10m/s. Figure 10 shows the impact of sensor node velocity compared Mc-Sync with TSHL and MU-Sync. The error of Mc-Sync is 11.5% of TSHL, and 17.9% of MU-Sync.

As in figure 10, with the parameter velocity's increase, the synchronization errors of TSHL and MU-Sync increase faster than Mc-Sync. This is because both TSHL and MU-Sync do not consider enough about propagation delay caused by node mobility. As Mc-Sync counteracts the effect of node mobility, its synchronization error is barely affected by velocity increase.

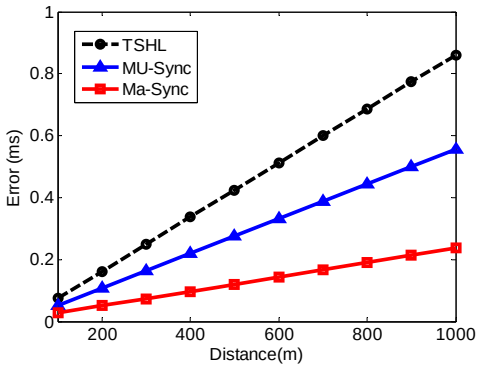



Figure 11. Effect of distance.

In order to study the affect of initial distance between nodes, we change the initial distance from 100m to 1000m, as in figure 11. The errors of all these algorithms are increasing with the distance's increase. It is because that long distance brings long propagation delay, which adds the error of propagation delay estimation. The error of Mc-Sync increases much smaller than TSHL and MU-Sync, as it reduces the effect of node mobility caused by distance increase.

Simulation results show that Mc-Sync preferences much better than resent synchronization algorithms. It has high synchronization precision and lest suffer node mobility's influence.

## V. Related Work

Synchronization protocols could be classified as Receiver-Receiver based approach, and Sender-Receiver based approach. In Receiver-Receiver based approach multiple nodes synchronize to a common event, and in Sender-Receiver based approach one node synchronizes with another. Both of them have their advantages and disadvantages.

Reference Broadcast Synchronization (RBS) [17] is a typical Receiver-Receiver based protocol. In RBS, nodes send reference beacons to their neighbors using physical layer broadcasts, receivers use broadcast message arrival time as a point of reference for comparing their clocks. It takes advantage of the broadcast channel to minimize synchronization error.

Sender-Receiver algorithms include many famous protocols. Timing-sync Protocol for Sensor Networks (TPSN) [18] works in two steps. A hierarchical structure is established in the network and then a pair wise synchronization is performed along the edges of this structure to establish a global timescale throughout the network. Eventually all nodes in the network synchronize their clocks to a reference node. Light weight Tree-based Synchronization (LTS) [19] introduces synchronization schemes that sacrifice accuracy by performing synchronization less frequently and between fewer nodes. The algorithm focuses on minimizing overhead while being robust and self-configuring. Tiny-Sync and Mini-Sync [20] proposed solution features minimal complexity in network bandwidth, storage and processing, which can achieve good accuracy. It also provides tight, deterministic bounds on both the clock skew and offset, and a method to synchronize the entire network in preparation for data fusion.

Delay measurement time synchronization (DMTS) [21] applicable for both single hop and multi-hop wireless sensor networks. As radio communication is a significant source of energy consumption, it adds minimum network traffic and is energy efficient. Flooding Time Synchronization Protocol (FTSP) [22] uses low communication bandwidth, and robust against node and link failures. FTSP achieves its robustness by utilizing periodic flooding of synchronization messages, and implicit dynamic topology update. The unique high precision performance is reached by utilizing MAC-layer timestamping and comprehensive error compensation including clock skew estimation. Global Clock Synchronization [23] proposes the all-node-based method, the cluster-based method, and the diffusion-based methods to solve the problem of time synchronization.

All of these terrestrial time synchronization methods cannot be directly applied to UWSN. Many unique characteristics of UWSN make underwater time synchronization challenging.

In the past several years, there are significantly growing interests in UWSN, but the research on underwater time synchronization is relatively limited [24]. TSHL [15] is the first time synchronization protocol designed for high latency networks. It performs linear regression over timing information to compute the clock skew. MU-Sync [10] is a cluster-based algorithm, which runs two times of linear regression to estimate the clock skew. It could account for the propagation delay

variability due to the nodes relative motion. Mobi-Sync [7] designed for mobile underwater acoustic sensor networks, which utilizes the spatial correlation of underwater mobile sensor nodes to estimate the long dynamic propagation delays. D-sync [8] exploits Doppler shift to time synchronization. It can handle substantial mobility without making any assumptions about the underlying motion and extensive signaling.

TSHL assume a static network which does not hold for most underwater systems. MU-Sync is not energy efficient, because it using a large number of two-way messages exchange. In Mobi-Sync, in order to estimate nodes velocity, each ordinary node has to maintain connectivity to at least three or more super nodes. D-sync has relatively complex computing process and susceptive with the measure precision of Doppler shift.

Not the same as these methods mentioned above, Mc-Sync utilizes two mobile reference nodes to eliminate the bad effect of node mobility and improve the synchronization precision.

## VI. Conclusion

In underwater sensor networks, node mobility has significant influence on the performance of time synchronization protocols. We research on the effect of node mobility, and make use of two mobile reference nodes to counteract the impact of node mobility. We present Mc-Sync as well as its trajectories, and analyze its preference though simulation.

In our future work, we plan to examine the applicability of our algorithm in more complex underwater environments. We also want to investigate other trajectories in underwater acoustic sensor networks of Mc-Sync to improve its accuracy.

## Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 61103196 and 61170258. Qingdao Science and Technology Development Program under Grant No. 12-1-4-3- (16) - jch.

## References

[1] J Heidemann, W Ye, J Wills, et al., “Research challenges and applications for underwater sensor networking”,IEEE Wireless Communication and Networking Conference, Washington, pp. 228-235, 2006.

[2] Wang Yifan, Zhang Zaichen, Bi Guangguo, “An overview on underwater acoustic sensor networks”, 17th International Conference on Telecommunications, 2010, pp. 779-783.

[3] Li Liu, Yang Xiao, Jingyuan Zhang, “A linear time synchronization algorithm for underwater wireless sensor networks”, IEEE International Conference on Communications (ICC '09), June 2009, pp. 14-18.

[4] Z. Zhou, J. H. Cui, and A. Bagtzoglou, “Scalable localization with mobility prediction for underwater sensor networks”, In Proceedings of IEEE INFOCOM’08, Mini-Conference, Phoenix, Arizona, USA, 2008

[5] Li Wang, Zhi Bin Wang, et al., “A survey of time synchronization of wireless sensor networks”, Conference on Wireless, Mobile and Sensor Networks (CCWMSN07) 2007.

[6] Lasassmeh, S.M., Conrad, J.M., “Time synchronization in wireless sensor networks: a survey”,IEEE SoutheastCon 2010.

[7] Yik Chung Wu Chaudhari, Q. Serpedin, E., “Clock synchronization of wireless sensor networks”, IEEE Signal Processing Magazine, vol. 8 (1), pp. 124-138, 2011.

[8] Jun Liu, Zhong Zhou, Zheng Peng and Jun-Hong Cui, “Mobi-Sync: efficient time synchronization for mobile underwater sensor networks”, IEEE Transactions on Parallel and Distributed Systems, issue 99, 2012.

[9] Feng Lu, Diba Mirza, Curt Schurger, “D-Sync: Doppler-based time synchronization for mobile underwater sensor networks”, The Fifth ACM International Workshop on Underwater Networks (WUWNET), Woods Hole, MA, 2010.

[10] Jun Liu, Zhaohui Wang, James Zheng Peng, Michael Zuba, Jun-Hong Cui, and Shengli Zhou, “TSMU: A Time Synchronization Scheme for Mobile Underwater Sensor Networks”, IEEE Global Telecommunications Conference (GLOBECOM 2011), pp. 1-6, 2011.

[11] Chirdchoo N, Soh W S, Chua K C., “MU-Sync: a time synchronization protocol for underwater mobile networks”, Proceedings of WUWNet’08. 2008, pp. 35-42.

[12] N. Parrish, S. Roy, and P. Arabshahi, “Symbol by symbol doppler rate estimation for highly mobile underwater OFDM”, Proceedings of the Fourth ACM International Workshop on Underwater Networks (WUWNet 2009), New York, NY, USA, 2009, pp. 1-8.

[13] Ismail Nor-Syahidatul N., Hussein Liban Abdullahi, Ariffin Sharifah H.S., “Analyzing the performance of acoustic channel in underwater wireless sensor network (UWSN)”,Asia Modeling Symposium 2010, pp. 550-555.

[14] Zhou Zhong, Peng Zheng, Cui Jun-Hong, Shi Zhijie, “Efficient multipath communication for time-critical applications in underwater acoustic sensor networks”, IEEE/ACM Transactions on Networking, August 5, 2010, pp. 28-41.

[15] M. Stojanovic, “On the relationship between capacity and distance in an underwater acoustic communication channel”, ACM SIGMOBILE Mobile Computing and Communications Review (MC2R), vol. 11, Issue 4, October 2007, pp. 34-43.

[16] Syed, A. A., Heidemann, J., “Time synchronization for high latency acoustic networks”, 25th IEEE International Conference on Computer Communications (INFOCOM 2006), pp. 1-12.

[17] M. Horauer, U. Schmid, K. Schossmaier, R. Holler, and N. Kero, “PSynUTC-evaluation of a high precision time synchronization prototype system for ethernet lans”, Proceedings of 34th Annual Precise Time and Time Interval Meeting (PTTI), December 2002, pp. 263-279.

[18] Elson J, Girod L, Estrin D., “Fine-grained network time synchronization using reference broadcasts”, Proceedings of OSDI 2002. Boston, MA, USA, 2002, pp. 147-163.

[19] Saurabh, G., Ram, K., Mani, B. S., “Timing-sync protocol for sensor networks”, Proceedings of the 1st International Conference on Embedded Networked Sensor Systems. Los Angeles, California, USA: ACM, pp. 138-149, 2003.

[20] Jana van Greunen, Jan Rabaey, “Lightweight time synchronization for sensor networks”, Proc 2nd ACM Int'l Conf Wireless Sensor Networks and Applications, pp. 11-19, 2003.

[21] Sichitiu ML, Veerarittiphan C., “Simple, accurate time synchronization for wireless sensor networks”, Proceeding

of the IEEE Wireless Communications and Networking Conference, 2003, pp. 16-20.
[22] Ping S., "Delay measurement time synchronization for wireless sensor networks", Intel Research Center: IR-TR-2003-64, 2003.
[23] Maroti M, Kusy B, Simon G., "The flooding time synchronization protocol", WCNC2004, Atlanta, GA, 2004, pp. 39-49.
[24] Qun Li, Daniela Rus, "Global clock synchronization in sensor networks", IEEE INFOCOM, 2004, pp. 19-28.
[25] Khandoker, T.U.I., Defeng Huang, Sreeram, V., "A low complexity linear regression approach to time synchronization in underwater networks", 8th International Conference on Information, Communications and Signal Processing (ICICS), pp. 1-5, 2011.

**Ying Guo** received her ME degree in Department of Automation from Qingdao University of Science and Technology in 2007, and received her PhD degree in the Department of Computer Science and Technology from Ocean University of China. Her research interests include wireless sensor networks, underwater acoustic networks. She is now working at Qingdao University of Science and Technology. Dr. Guo is a member of the IEEE.

**Yutao Liu** received his BE degree in Department of Computer Science and Technology from Qingdao University of Science and Technology in 2005. His research interests include computer networks, computer software and application. He is now working at Qingdao Yawei Software Co., Ltd.

# On Characterizing of Information Propagation in Online Social Networks

Xiaoting Han
School of Economics and Management, Beihang University, Beijing, China
Email: hanxiaoting@buaa.edu.cn

Li Niu
Key Laboratory of Ministry of Education for Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China
School of Information Resource Management, Renmin University of China, Beijing, China
Email: libraniu@foxmail.com

***Abstract*—Recent years have witnessed the explosive growth of online social networks (OSNs), which provide a perfect platform for observing the information propagation. Based on the theory of complex network analysis, considering each node in OSNs as an agent, an information propagation model called IP-OSN has been proposed in this paper, and numerical simulation experiments have been extensively conducted, which show the mechanism of information propagation in OSNs. It can be seen from the experimental results that along with the information spreading scope will be larger while time passing. Moreover, from the network structure perspective, some interesting findings are got. First, the number of initial known nodes has a certain impact on the information propagation speed; however it doesn't affect the number of final known and unknown nodes. Then, the number of neighbors has a great impact on the information propagation process and results. Finally, percentage of different user type is the decisive factor for information propagation either in final node number or propagation speed. This paper attempts to complement the theoretical framework of information propagation and complex networks, and thus to further promote the practice of information control in OSNs.**

***Index Terms*—Information Propagation/Diffusion; Online Social Networks; Network Structure; User Behavior**

## I. INTRODUCTION

Along with the development of Internet and Web2.0, online social networks sites such as Facebook, MySpace, LinkedIn and Twitter have become a popular social media platform0, while they have been developed massively for business and political purposes, such as viral marketing, targeted advertising, political campaigns, and even terrorist activities[2][3]. The users of these sites and the friendships among them constitute the so-called online social networks (OSNs).

Manuscript received April 1, 2012; revised May 1, 2012; accepted May 25, 2012.
Corresponding author: Li Niu, Email:libraniu@foxmail.com.

Recently study on information propagation rules in online social networks have increased. These studies usually focus on the topology of these social networks [4] and the mechanism of information propagation [5][6]. However, limited work has been done from the network structure perspective. Therefore, from the view of network structure, to study the rules of information propagation, and thus to further study how to control the information propagation process, has a very important theoretical value and practical significance.

Online social networks are focused on sharing information, and as such, have been studied extensively in the context of information propagation. Information propagation has been modeled in blogs [7], email [8], and sites such as Twitter, Digg, and Flickr [9][10].

Information propagation has been extensive and in-depth research in the field of epidemiology, sociology and marketing. Biology and epidemiology has conducted in-depth study on diffusion of virus within the group in early time [11], and the classical SIS model and SIR model are proposed. In sociology and marketing area, research on diffusion focuses on the problems of innovation diffusion. In the early 20th century, the BASS model [12] opened up new research directions for this research area, and derived a series of related models. Lopez-Pintado et al. [13] studied the product diffusion in complex social networks. He considered the mutual influence among individuals on the micro-level into the propagation equation based on mean-field theory, and found out that innovation diffusion in complex networks also exists a threshold which closely related to the degree distribution and propagation functions of the network. Westerman et al. [14] studied the effect of system generated reports of connectedness on credibility, and got that curvilinear effects for number of followers exist, such that having too many or too few connections results in lower judgments of expertise and trustworthiness.

Recent years, more and more scholars have modeled information propagation process in OSNs from the perspective of natural sciences and interdisciplinary. Agliari et al. [15] studied information spreading in a


doi: 10.4304/jnw.8.1.124-131

population of diffusing agents. Galstyan et al. [16] examined the problem of maximizing influence propagation in structured heterogeneous networks. Karsai et al. [17] studied the effects of different topological and temporal correlations on information spreading in complex communication networks. Jin et al. [18] provided HPC simulations method to study the behaviors of information propagation over complex social networks.

In the environment of Web 2.0, users who diffuse information in OSNs can be considered as agent, and OSNs can be considered as media; moreover, relation among users can be considered as links. Thus the information propagation process can be considered as the process of information copying, transferring, changing and diffusing in a social network whose nodes are agents and edges are links. In the process of information propagation, a node's attitude is often influenced by its neighbors. By the mutual effect between the neighbors, the information can be diffused in the OSNs. In accordance with different impact way between neighbors, models can be divided into threshold mode [19][20] and probability mode [21]. According to the different specific matters involved, researchers proposed different information propagation process model based on the above impact models, including infectious diseases spreading model [21], new products propagation model [22], rumors spreading model [23] and so on.

Based on the theory of complex network analysis, considering each node in OSNs as an agent, an information propagation model called IP-OSN has been proposed in this paper, and numerical simulation experiments have been extensively conducted, which show the mechanism of information propagation in OSNs. It can be seen from the experimental results that along with the information spreading scope will be larger while time passing. Moreover, from the network structure perspective, some interesting findings are got. First, the number of initial known nodes has a certain impact on the information propagation speed; however it doesn't affect the number of final known and unknown nodes. Then, the number of neighbors has a great impact on the information propagation process and results. Finally, percentage of different user type is the decisive factor for information propagation either in final node number or propagation speed. This paper attempts to complement the theoretical framework of information propagation and complex networks, and thus to further promote the practice of information control in OSNs.

## III. Modeling the Information Propagation in OSNs

### A. Model Description

In order to describe the information propagation model clearly, notations of parameters used in the model is shown in Table I.

In this paper, an information propagation model referred to communicable disease model SIR and SIS [24] named IP-OSN model is proposed in this section. The basic idea of this model is as follows:

TABLE I.
NOTATIONS OF PARAMETERS

| Notation | Description |
|---|---|
| $N$ | number of nodes in OSNs |
| $<k>$ | averaged degree of nodes in OSNs |
| $n_{in}$ | number of initial known nodes |
| $k_{in}$ | degree of source node |
| $k_i$ | degree of node $i$ |
| $p_1$ | percentage of type A users |
| $p_2$ | percentage of type B users |
| $p_3$ | percentage of type C users |
| $n_1$ | number of type A users |
| $n_2$ | number of type B users |
| $n_3$ | number of type C users |
| $p$ | probability of unknown node changed to known |
| $T$ | total running time |

- Information is diffused in OSNs, and users in OSNs are divided into three types based on user behavior:
  - Type A users: They don't accept information in OSNs, and they don't diffuse the information.
  - Type B users: They accept information in OSNs, but they don't diffuse the information.
  - Type C users: They accept information in OSNs, and they are willing to diffuse the information.
- At the initial time ($t$=0), there are few nodes in OSNs know the information, and most nodes haven't known the information.
- When $t$=$t$+1, the information propagation process starts until the total running time arrives.
- Nodes know the information and is type C will diffuse it to all its neighbors.
- If the neighbor node is type A, it will reject the information.
- If the neighbor node is type B or C, it will accept the information.

### B. Model Algorithm

The process algorithm of IP-OSN model is shown in Fig. 1:

*Step 1*: When $t$=0, Initialize information. Set $n_{in}$ nodes to known, and ($N$-$n_{in}$) nodes to unknown. After the initialization, the number of type A, B, C users are:

$$n_1 = N \times p_1 \quad (1)$$

$$n_2 = N \times p_2 \quad (2)$$

$$n_3 = N \times p_3 \quad (3)$$

*Step 2*: When $t$=$t$+1, visit each node (suppose node $i$) and do Step 2.1 to Step 2.5, and the information propagation process starts until $t$=T.

*Step 2.1*: If the status of node $i$ is unknown, the algorithm ends; else, go to Step 2.2.

*Step 2.2*: If node $i$ is type A or B, the algorithm ends; else, go to Step 2.3.

*Step 2.3*: Visit all the neighbors of node $i$, then do Step 2.4 $k_i$ times.

*Step 2.4*: Suppose the algorithm is visiting the neighbor node of node *i*, and we name the neighbor node *j*, then do Step 2.4.1 to Step 2.4.3.

*Step 2.4.1*: If node *j* is type A, the algorithm ends; else, go to Step 2.4.2.

*Step 2.4.2*: If status of node *j* is known, the algorithm ends; else, go to Step 2.4.3.

*Step 2.4.3*: Change the status of node *j* to known with the probability of *p*, and visit next neighbor of node *i*.

*Step 2.5*: Visit next node until all the nodes in OSNs are visited.

*Step 3*: End.

Suppose the number of different nodes at time t is as shown in Table II. According to mean-field theory [25], after an iteration of the above propagation process, when time = *t*+1, the number of each type of nodes is:

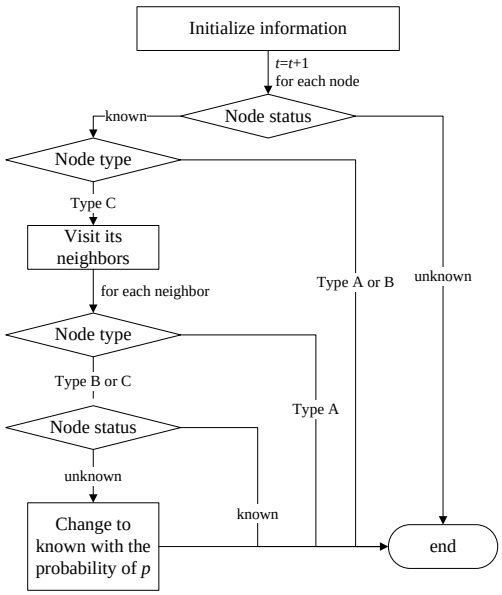



Figure 1. Information propagation process in OSNs

TABLE II.
NUMBER OF DIFFERENT NODES AT TIME *T* AND *T*+1

| Node Type | Status | Number at *t* | Number at *t*+1 | Number increased |
|---|---|---|---|---|
| A | known | $n_{k1}$ | $n'_{k1}$ | |
| | unknown | $n_{n1}$ | $n'_{n1}$ | |
| B | known | $n_{k2}$ | $n'_{k2}$ | |
| | unknown | $n_{n2}$ | $n'_{n2}$ | |
| C | known | $n_{k3}$ | $n'_{k3}$ | |
| | unknown | $n_{n3}$ | $n'_{n3}$ | |
| all | known | $n_k$ | $n'_k$ | $n_{kit}$ |
| | unknown | $n_n$ | $n'_n$ | $n_{nit}$ |

$$n'_{k1} = n_{k1} \tag{4}$$

$$n'_{n1} = n_{n1} \tag{5}$$

$$n'_{k2} = n_{k2} + n_{n2} - n_{n2} \times (1 - p_3 \times p)^{<k>} \tag{6}$$

$$n'_{n2} = n_{n2} \times (1 - p_3 \times p)^{<k>} \tag{7}$$

$$n'_{k3} = n_{k3} + n_{n3} - n_{n3} \times (1 - p_3 \times p)^{<k>} \tag{8}$$

$$n'_{n3} = n_{n3} \times (1 - p_3 \times p)^{<k>} \tag{9}$$

From Eq. (4)-(9), we can get that the number of known nodes increased from time *t* to *t*+1 is:

$$n_{kit} = n'_k - n_k = n'_{k2} - n_{k2} + n'_{k3} - n_{k3} = n_{n2} - n_{n2} \times (1 - p_3 \times p)^{<k>} + n_{n3} - n_{n3} \times (1 - p_3 \times p)^{<k>} \tag{10}$$

To be simple, that is:

$$n_{kit} = (n_{n2} + n_{n3}) \times (1 - (1 - p_3 \times p)^{<k>}) \tag{11}$$

Similarly, that the number of unknown nodes decreased from time *t* to *t*+1 is:

$$-n_{nit} = (n_{n2} + n_{n3}) \times (1 - (1 - p_3 \times p)^{<k>}) \tag{12}$$

From Eq. (11) and (12), we can see that the number of known nodes is increasing while the information is diffusing, but the number of increasing known nodes is gradually reduced. This phenomenon will be simulated in the next section.

## IV. EXPERIMENTS AND RESULTS ANALYSIS IN DEFAULT PARAMETER SETTING

### A. Methodology

There are several methods to study the information propagation in OSNs, such as complex network analysis [26], cellular automata [27] and agent based modeling [28]. In these three methods, we choose agent based modeling as the method to simulate the information propagation process because of its flexibility. Agent based modeling method can adjust the various factors effecting information propagation, therefore how the different combination of factors causes different information propagation effect can be compared easily, which can provide strong evidence for controlling negative information and spreading the positive information. In the simulation process, specific data of each agent can be easily obtained to quantitatively analysis how different user behavior effects the information propagation in real-world OSNs.

In order to prove the efficiency of the above model, a network simulating OSNs is conducted in Netlogo [29], which is simulation software based on agent. Nodes in OSNs are modeled by agents, and interaction among agents is used to simulate the information propagation mechanism in the proposed IP-OSN model. In this way, parameters in IP-OSN model can easily be controlled, which can facilitate the observation of efficiency and effectiveness the model and obtain the data of simulation results to do quantitative analysis.

### B. Experimental Setup

A randomly generated data set is used for the experiments. There are 2000 nodes in this data set, and the averaged degree is 6. Key features of this data set are summarized in Table III.

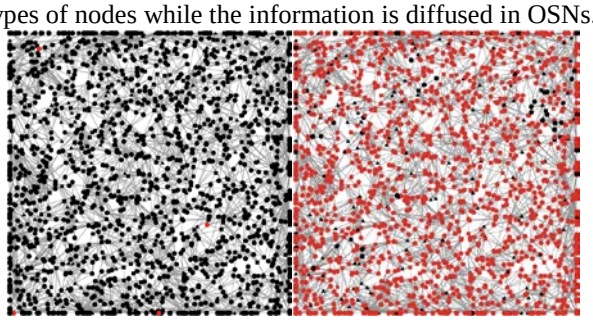
TABLE III.
PARAMETERS SETTING OF DATA SET

| Parameter | value |
|---|---|
| $N$ | 2000 |
| $<k>$ | 6 |
| $n_{in}$ | 4 |
| $k_{in}$ | 6 |
| $p_1$ | 10% |
| $p_2$ | 20% |
| $p_3$ | 70% |
| $p$ | 10% |
| $T$ | 400 |

The proposed IP-OSN model is implemented in Netlogo, on a Microsoft Windows 7 Professional platform with SP1 64bit edition. The experimental PC is with an Intel Core i7 2620M CPU, 4 GB DDRII 667 MHz RAM, and a Seagate Barracuda 7200.11 500GB hard disk.

### C. Experimental Results

First, the running effect of the IP-OSN model is given in Fig. 2 and Fig. 3. In Fig. 2, red nodes represent nodes who don't know the information, while black nodes represent nodes who have already known the information. Fig. 3 shows the percentage change process of the two types of nodes while the information is diffused in OSNs.



$t$=0 $t$=244

Figure 2. Initial and final status of information propagation

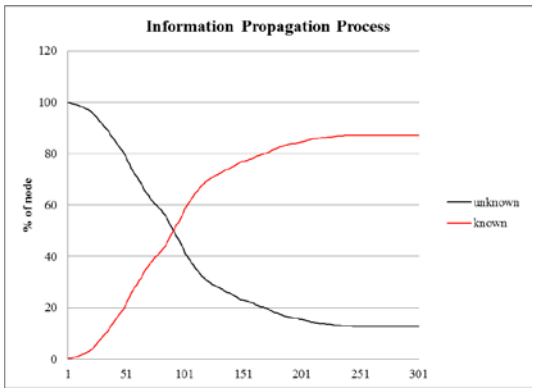



Figure 3. Evolution process of information propagation

As shown in the figures, at the initial time ($t$=0), there are only 4 nodes who know the information, while 99.8% of the initial nodes are unknown. As time passed, number of known nodes rapidly increases. At $t$=244, the number of known nodes achieves maximum of 1743. After that, number of known nodes and unknown nodes remains unchanged.

## V. SENSITIVITY ANALYSIS OF PATAMETERS FROM THE PERSPECTIVE OF NETWORK STRUCTURE

### A. Sensitivity Analysis of Number of Initial Known Nodes in OSNs

In order to examine how In order to examine how $n_{in}$ (number of initial known nodes in OSNs) affects the information propagation in OSNs, $n_{in}$ is changed to 2 in situation 1 and changed to 6 in situation 2 in the IP-OSN model.

The experimental results are shown in Fig. 4 - Fig. 6. Fig. 4 shows number of unknown and known nodes in different situations. Fig. 5 shows the final status in different situations. We can see from Fig. 4 and Fig. 5 that the number of final known nodes and unknown nodes doesn't change much along with $n_{in}$ changes. In other words, the number of initial known nodes is not a decisive factor of the final number of known and unknown nodes.

Different propagation process based on the above three situations is shown in Fig. 6. It can be seen that the propagation changes a lot when $n_{in}$ (the number of initial known nodes in OSNs) changes. Clearly the propagation process goes faster when $n_{in}$ increases. It means that when the initial known nodes get more, the propagation speed among nodes will be faster.

By sensitivity analysis of number of initial known nodes in OSNs, we find that $n_{in}$ has a certain impact on the information propagation speed, however it doesn't affect the number of final known and unknown nodes.

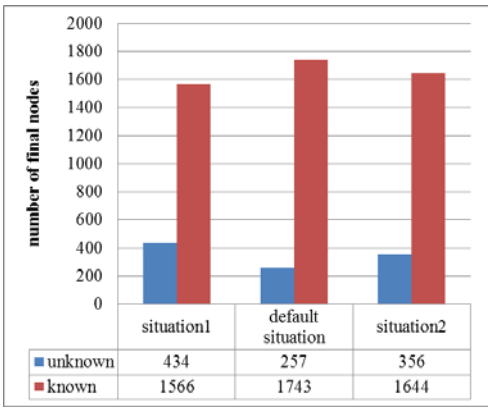



Figure 4. Number of final nodes in different situations

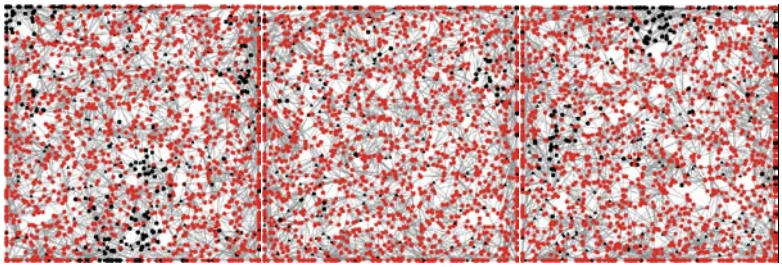


Figure 5. Final status in different situations

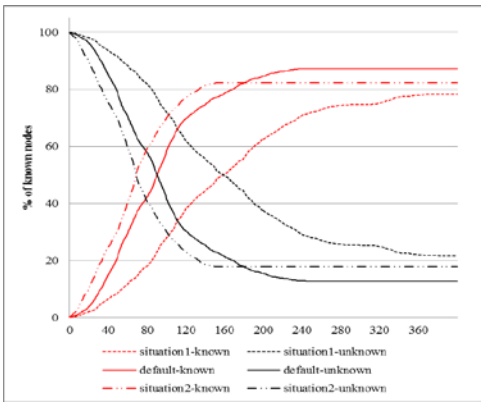



Figure 6. Changing process in different situations

## B. *Sensitivity Analysis of Averaged Degree of Nodes in OSNs*

In order to examine how In order to examine how <*k*> (the averaged degree of nodes in OSNs) affects the information propagation in OSNs, <*k*> is changed from 4 to 8 in the IP-OSN model as shown in Table 4.

The experimental results are shown in Fig. 7 - Fig. 10. Fig. 7 shows the final status in different situations. Fig. 8 shows number of unknown and known nodes in different situations. We can see from Fig. 7 and Fig. 8 that the number of final known nodes increases with <*k*> increasing. In other words, when nodes in OSNs have more neighbors, the same information will diffused to more nodes in the end. However, the number of final known nodes in situation 3 is as much as that in situation 4. That is to say, there is a threshold for <*k*>, and in this experiment, the threshold is 7. When $< k >\geq 7$ , the number of final known nodes will be almost the same.

TABLE IV.
DIFFERENT PARAMETER SETTINGS OF <*k*>

| | <*k*> |
|---|---|
| Default Situation | 6 |
| Situation 1 | 4 |
| Situation 2 | 5 |
| Situation 3 | 7 |
| Situation4 | 8 |

Different propagation process based on the above five situations is shown in Fig. 9 and Fig. 10. It can be seen that the propagation changes a lot when <*k*> (the averaged degree of nodes in OSNs) changes. Clearly the propagation process goes faster when <*k*> increases. It means that when the degree of nodes increases, the propagation speed among nodes will be faster.

By sensitivity analysis of averaged degree of nodes in OSNs, we find that <*k*> has a certain impact on the information propagation process and results. When <*k*> increases, the nodes who know the information will be more and the propagation process will be faster.

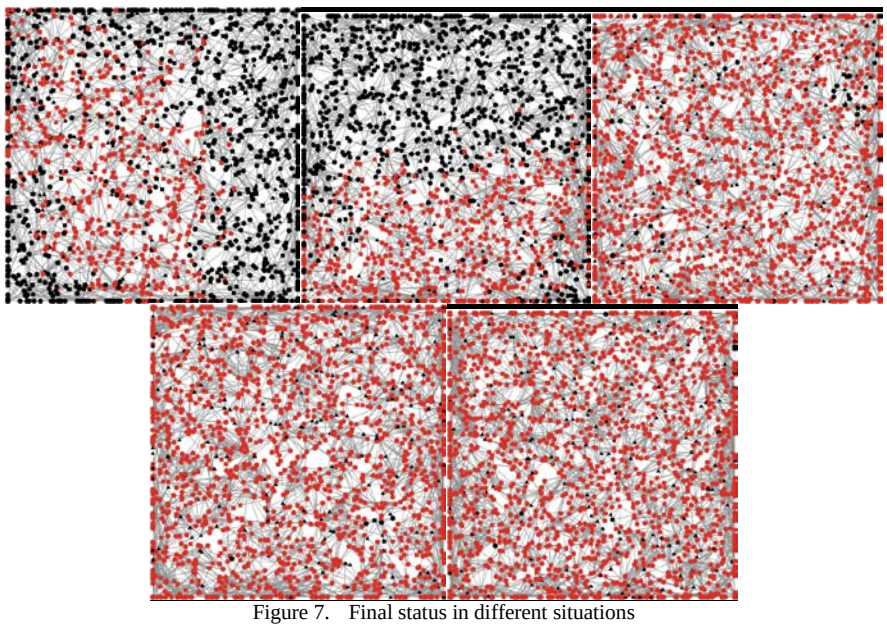


Figure 7. Final status in different situations

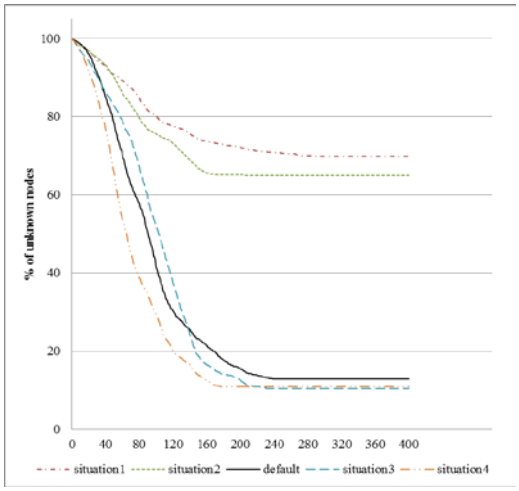



Figure 8. Changing process of unknown nodes

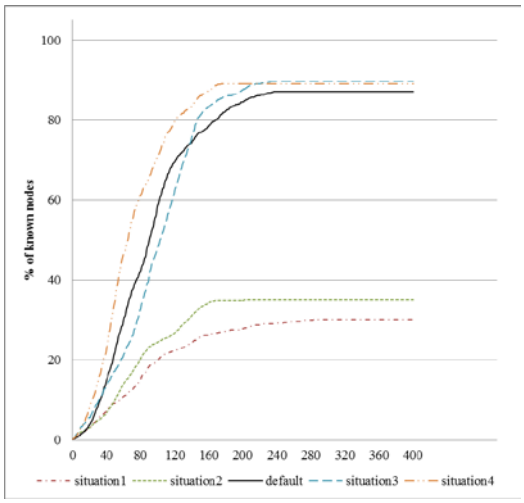



Figure 9. Changing process of known nodes

### C. *Analysis of Different User Types*

In this section, we change the percentage of different user types in order to analysis how different user behavior affects information propagation in OSNs. Four more situations are assumed. In situation 1 and situation 2, we decreases $p_1$ and $p_2$ (the percentage of Type A and Type B user), and in contrast, in situation 3 and situation 4, $p_1$ and $p_2$ are increased. The parameter setting is shown in Table V.

The experimental results are shown in Fig. 11 - Fig. 14 Fig. 11 shows number of unknown and known nodes in different situations. Fig. 12 shows the final status in different situations. We can see from Fig. 11 and Fig. 12 that the number of final known nodes increases with $p_3$ increasing and $p_1$ and $p_2$ decreasing. In other words, when there are more nodes who want to accept and diffuse information in OSNs, the information will be diffused to more nodes.

TABLE V.
DIFFERENT PARAMETER SETTINGS OF USER TYPES

| | $p_1$ | $p_2$ | $p_3$ |
|---|---|---|---|
| Default Situation | 10% | 20% | 70% |
| Situation 1 | 0% | 5% | 95% |
| Situation 2 | 5% | 10% | 85% |
| Situation 3 | 15% | 30% | 55% |
| Situation4 | 20% | 40% | 40% |

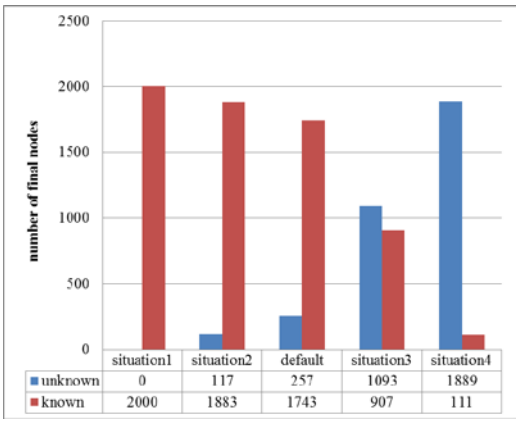



Figure 10. Number of final nodes in different situations

Different propagation process based on the above three situations is shown in Fig. 13 and Fig. 14. We can see in the figures that the propagation process changes in the five situations. Its changing trend goes steeper in situation 1 and situation 2 while gentler in situation 3 and situation 4. It shows that when the number of users who want to diffuse information in OSNs increases, the information speeding speed will be faster, and in contrast, when the number of users who want to diffuse information in OSNs decreases, the information speeding speed will be slower.

By analysis of different user types affecting the information propagation in OSNs, we find out that the percentage of different users has a certain impact on the propagation process and propagation result. It is clearly seen from the experimental results that when there are more nodes who are willing to accept and diffuse information in OSNs, nodes who will finally know the information will be more, and the propagation process will be faster.

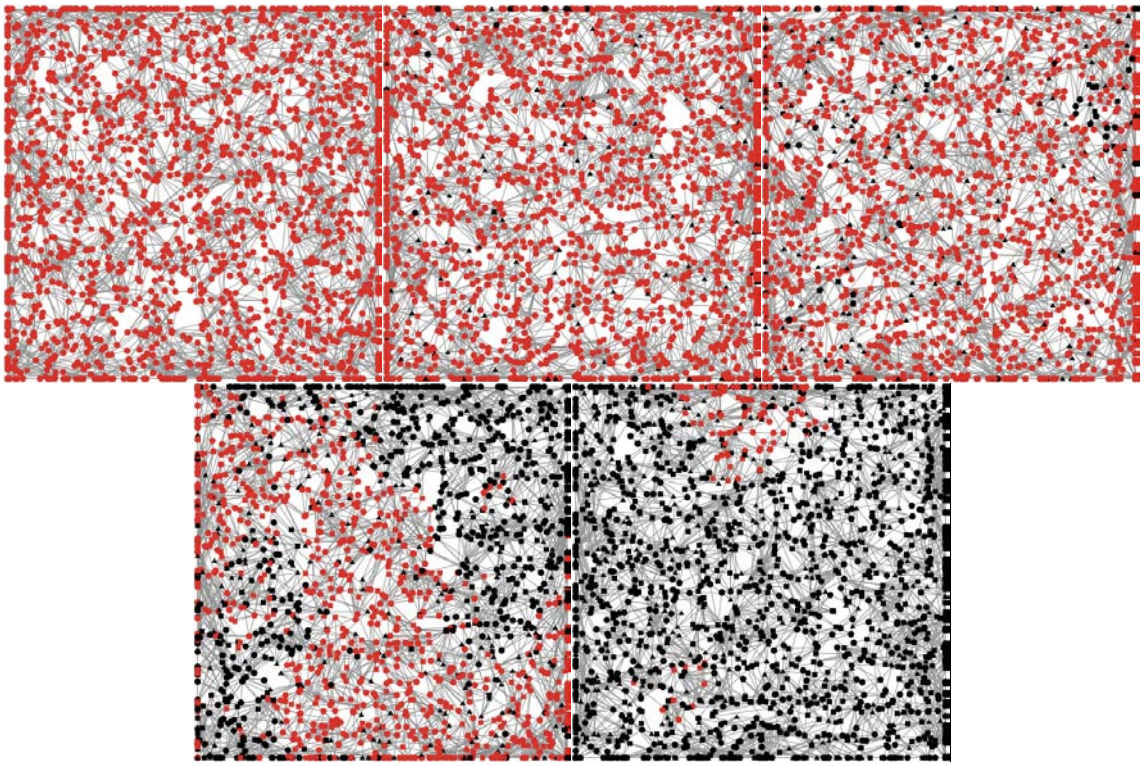


Figure 11. Final status in different situations

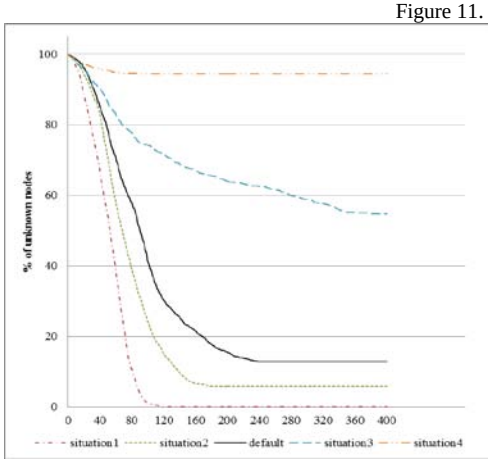



Figure 12. Changing process of unknown nodes

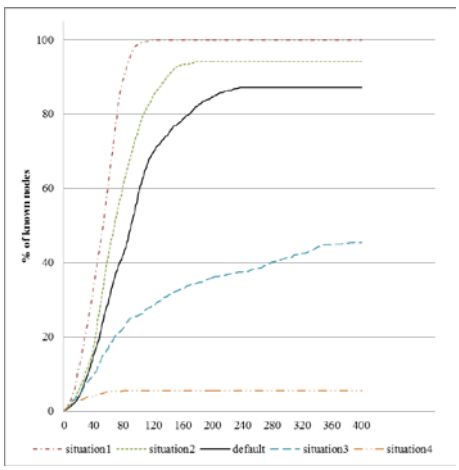



Figure 13. Changing process of known nodes

## VI. SUMMARY

An information propagation model named IP-OSN is proposed in this paper. From the model, we can investigate some information propagation rules in OSNs. From the experimental results, it is can be seen that along with the information propagation, the number of known nodes increases and reaches its maximum, then keep an unchanging status. Moreover, numerical simulations are conducted from the network structure perspective. Simulation results show some meaningful findings for information control in OSNs. First, the number of initial known nodes is a decisive factor on the information propagation speed but not on the number of final known and unknown nodes. Also, the number of neighbors has a great impact on the information propagation process and results either in final node number or propagation speed. Furthermore, we find out that the more nodes who are willing to accept and diffuse information in OSNs, the more nodes who will finally know the information, and the faster the information will be diffused.

## REFERENCES

[1] Y.Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, "Analysis of topological characteristics of huge online social networking services," *Proceedings of the 16th international conference on world wide web*, WWW '07: pp. 835–844, 2007.

[2] A. Dunne, M. A. Lawlor, J. Rowleym, "Young people's use of online social networking sites - a uses and gratifications perspective," *Journal of Research in Interactive Marketing*, vol. 4, pp.46-58, 2010.

[3] C. M. K. Cheung, P. Y. Chiu, M. K. O. Lee, "Online social networks: Why do students use facebook?," *Computers in Human Behavior*, vol. 27, pp. 1337–1343, 2011.

[4] L. A. N. Amaral, B. Uzzi, "Complex systems - a new paradigm for the integrative study of management, physical, and technological systems," *Management Science*, vol. 53, pp. 1033-1035, 2007.
[5] V. Bellini1, G. Lorusso, A. Candini, W. Wernsdorfer, T. B. Faust, G. A. Timco, R. E. P. Winpenny, M, "Affronte. Propagation of Spin Information at the Supramolecular Scale through Heteroaromatic Linkers," *Physical Review Letters*, vol. 106, pp. 227205, 2011.
[6] J. L. Iribarrena, E. Moro, "Affinity Paths and information diffusion in social networks," *Social Networks*, vol. 33, pp. 134–142, 2011.
[7] F. Fu, L. Liu, L. Wang, "Empirical analysis of online social networks in the age of Web 2.0," *Physica A: Statistical Mechanics and its Applications*, vol. 387, pp. 675–684, 2008.
[8] L. A. Thompson, K. Dawson, R. Ferdig, E. W. Black, J. Boyer, J. Coutts, N. P. Black, "The Intersection of Online Social Networking with Medical Professionalism," *Journal of General Internal Medicine*, vol. 23, pp. 954-957, 2008.
[9] A. M. Kaplan, , M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, pp. 59–68, 2010.
[10] D. Scanfeld, V. Scanfeld, E. L. Larson, "Dissemination of health information through socialnetworks: Twitter and antibiotics," *American Journal of Infection Control*, vol. 38, pp. 182–188, 2010.
[11] R. M. Anderson, R. M. May. "Infectious Diseases of Humans: Dynamics and Control." Oxford University Press, USA, 1992.
[12] R. Albert, A. L. Barabasi. "Statistical mechanics of complex networks." *Reviews of Modern Physics*, vol. 74(1), pp. 47-97, 2002.
[13] D. Lopez-Pintado. "Diffusion in Complex Social Networks." *Games and Economics Behavior*, vol. 62(2), pp. 573-590, 2008.
[14] D. Westermana, P. R. Spenceb, B. V. D. Heide. "A social network as information: The effect of system generated reports of connectedness on credibility on Twitter." *Computers in Human Behavior*, vol. 28(1), pp. 199-206, 2012.
[15] E. Agliari, R. Burioni, D. Cassi, F. M. Neri. "Efficiency of information spreading in a population of diffusing agents." *Physical Review E*, vol. 73, pp. 046138, 2006.
[16] A. Galstyan, V. Musoyan, P. Cohen. "Maximizing influence propagation in networks with community structure." *Physical Review E*, vol. 79, pp. 056102, 2009.
[17] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, J. Saramäki. "Small But Slow World: How Network Topology and Burstiness Slow Down Spreading." *Physical Review E*, vol. 83, pp. 025102, 2011.
[18] J. Jin, S. J. Turner, B. S. Lee, J. Zhong, B. He. "HPC Simulations of Information Propagation Over Social Networks." *Procedia Computer Science*, vol. 9, pp. 292-301, 2012.
[19] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, pp. 1360-1380, 1978.
[20] D. J. Watts, "A simple Model of Information Cascades Random Networks," *Proceedings of the National Academy of Science*, vol. 99, pp. 5766-5771, 2002.
[21] H.W. Hetheote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, pp. 599-653, 2000.
[22] N. Meade, T. Islam, "Modeling and forecasting the diffusion of innovation: a 25-year review," *International Journal of Forecasting*, vol. 22, 519-545, 2006.
[23] M. Nekovee, Y. Moreno, G. Bianconi et al, "Theory of rumor spreading in complex social Networks, " *Physica A*, vol. 374, pp. 457-470, 2007.
[24] H. W. Hethcote, "Qualitative Analyses of Communicable Disease Models," *Mathematical Biosciences*, vol. 28, pp.335-356, 1976.
[25] Y. Roudi, J. A. Hertz, "Mean field theory for nonequilibrium network reconstruction," *Physical Review Letters*, vol. 106, pp. 048702, 2011.
[26] R.V. Kozinets, K. De Valck, A.C. Wojnicki, S.J.S. Wilner, "Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities," *Journal of Marketing*, vol. 74, pp. 71-89, 2010.
[27] J. Goldenberg, B. Libai, "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters*, vol. 12, pp. 211-223, 2001.
[28] T. Smith, J. R. Coyle, E. Lightfoot, A. Scott, "Reconsidering Models of Influence: The Relationship between Consumer Social Networks and Word-of-Mouth Effectiveness," *Journal of Advertising Research*, vol. 47, pp. 387-397, 2007.
[29] U. Wilensky, "http://ccl.northwestern.edu/netlogo/," 2012.

**Xiaoting Han** was born in Shandong province of China at 7th April, 1983. She received her bachelor degree in information management and information system at Beihang University in Beijing of China in 2003, and then received her master degree in management science and engineering at Beihang University in Beijing of China in 2006.

She is now a technician at School of Economics and Management and dean of Lab of Economics and Management of Beihang University. She is also a Ph.D. candidate currently focused on information diffusion in online social networks, complex networks analysis, web data mining and statistical physics.

**Li Niu** was born in Henan province of China at 23rd October, 1982. He received his bachelor degree in information management and information system at Beihang University in Beijing of China in 2005, and then received his doctor degree in system engineering at Beihang University in Beijing of China in 2010.

He is now an assistant professor at School of Information Resource Management of Renmin University of China. His research interests include information resource management, social network analysis, agent-based modeling and decision support system.

# QoS Evaluation of VANET Routing Protocols

Shouzhi Xu
Collage of Computer and Information Technology, China Three Gorges University, Yichang, China
Email: xsz@ctgu.edu.cn

Pengfei Guo, Bo Xu, Huan Zhou
Collage of Computer and Information Technology, China Three Gorges University, Yichang, China

***Abstract*—Traffic jams and traffic accidents have become a major concern in current society. VANET (vehicle ad hoc network) is an emerging attractive application to solve such problems. Quality of service (QoS) in VANET becomes a hot topic own to its increasing challenge about unique features, such as limited transporting distance, high mobility, and poor link quality. The main goal of this paper is to analyze the main quality criteria among popular routing protocols with an integrated VANET test bed. Typical topology-based routing protocols are reviewed. To study QoS performance of different protocols, evaluating models of frame loss ratio, PSNR and connectivity probability are illustrated. Three typical routing protocols: DSDV, AODV and GPSR are chosed to testify the QoS according to the statistics result of video transmission over VANET testbed. QoS performance is analyzed under several conditions of different distance of routing data transmission and vehicles' arriving rate. Test results show that Pro-active protocol is not suitable for high mobility VANET, and Position-based hybrid protocol is more suitable for video transmission over VANET than Re-active protocol. Comparing with other research, the result shows that our model is better and more efficient for the evaluation platform.**

***Index Terms*—VANET, QoS Routing, Evaluation Model**

## I. INTRODUCTION

Two urgent traffic problems have become a major concern in current society. Firstly, traffic jams have global economic and environmental impacts since that cause more and more attention to traffic delays and fuel wastage, which need intelligent transport management technology to optimize traffic. Additionally, road traffic accidents have become as a killer worldwide, which make transport safety warning technology very important. As a solution, 5.9 GHz Dedicated Short Range Communication (DSRC) technology is applied for vehicular communications, and Vehicular Ad Hoc Network (VANET) is developed widely for traffic safety, transport efficiency and information service [1].

VANET is a form of Mobile ad-hoc network (MANET), to provide communications with DSRC among vehicles and nearby roadside fixed equipment [2]. These roadside units open up a wide variety of services for vehicular networks such as acting as a drop point for messages on sparsely populated roads, serving up geographically-relevant data, or serving as a gateway to the Internet. VANET can play a key role and broadcast useful information to the vehicles in the vicinity in the future. It can be used to not only solve traffic safety warning, but also traffic information inquiry, commercial advertisement and so on.

The most of the important applications is driving assistance. Adjacent vehicles can share road information video and traffic information video with each other, which are got from inter-vehicle equipment and roadside facilities. With these videos, moving vehicles can understand road information and traffic information around. In case of a car accident in the distance, accident avoidance warnings could quickly notify drivers about conditions that could cause a collision. Vehicles that are driving towards this area can receive live video about this area, and then drive with alert or simply choose another road. Cooperative driving would allow vehicles to navigate without driver intervention by communicating with other vehicles about velocity, proximity, and other factors. Communication made to other vehicles prior to collision may allow the accident to be reconstructed more easily [3]. Rescue vehicles could instantly receive exact coordinates of the location of an accident to reach the scene of the emergency faster. Furthermore, information about traffic and road hazards could be acquired and fed into vehicle navigation systems in real-time to provide alternate driving routes.

Furthermore, commercial and entertainment applications become an attractive tendency. Road side businesses, such as hotels and restaurants, can use content-rich video streams to broadcast advertisements to drivers on the road. Passengers in nearby cars can setup a video conversation by using the inter-vehicle streaming technology [4]. Drivers or passengers could also enjoy watching live news or football match, while the video data is conveyed by other relay vehicles.

In whatever applications, VANET routing has been a very crucial research topic. Traffic management applications require data dissemination in a multi-hop network in different geographic locations to alert other vehicles regarding traffic situations. On the other hand, commercial applications require unicast routing. However, VANET has faced the big challenges on stability, efficiency, scalability of network since that routing protocols developed for MANET show degraded performance in vehicular scenarios. VANET has its unique characteristics which may make data link disable


doi: 10.4304/jnw.8.1.132-139

[5], such as limited transporting distance, high mobility, distributed nature of operation, poor link quality and varied channel conditions. Such particular features often make standard networking protocols inefficient or unusable in VANETs, whence the growing effort in the development of communication protocols which are specific to vehicular networks. In a VANET, quality of multimedia data transmission is more difficult than other wireless networks, such as Ad Hoc network, wireless sensor network and general MANET.

The main goal of this paper is to analyze the main quality criteria among popular QoS routing protocols and to find a QoS validated route between highly "on mobile" vehicles with an integrated VANET test bed. its evaluation models will be proposed and a simulation tool, VANET-Evalvid will be designed to evaluate the quality of video transmitted over VANET in a more realistic environment. With analysis of network performance of video communication in dense traffic scenario or in sparse traffic scenario, we try to decide which is the key factors of QoS performance in terms of frame loss ratio, PSNR and connectivity probability.

The rest of this paper is organized as follows. Three kinds of topology based VANET routing protocols: Proactive, Reactive and Hybrid Protocols are reviewed in Section 2. Section 3 proposes three main evaluation models and designs a simulation platform for evaluating video transmission performance in VANET. Section 4 simulates video communication and analyzes the experimental results. Section 5 gives the concluding remarks finally.

## II. Typical VANET Routing Protocols

VANET involves vehicles travelling in a wide area at high speeds resulting in rapid network topology changing, and so presents challenges to efficient and reliable message dissemination. The special features of VANET include: non-uniform node distribution, lack of a centralized administrative entity, fragile link topology, dynamic changes in node density, large scale and stringent real time delay requirements, and the impact of driver's behavior on network topology [6]. Those features make it is very hard to provide suitable QoS service for VANET, the key technology is about routing.

A vast number of protocols have been developed to cater for VANET specific reliability requirements. In VANET, the routing protocols are classified into five categories on the basis of area/application where they are most suitable: Topology based, Position based, Cluster based, Geocast and Broadcast. The definitions of these fives categories may overlap.

Topology based routing protocols which use links information of network to perform packet forwarding are mainly studied in this paper. They are further divided into Proactive, Reactive and Hybrid Protocols [7]. Typical protocols are classified as shown in Tab. 1.

In this section, three routing protocols: DSDV, AODV, GPRS are analyzed, which are chosen from typical classes respectively.

TABLE I.
TYPICAL VANET ROUTING PROTOCOLS

| Proactive routing | Reactive routing | Hybrid routing |
|---|---|---|
| DSDV | AODV | GPSR |
| OLSR | S-AODV | GPSR-L |
| MOPR | PAODV | GPCR |
| GSR | DSR | GpsrJ+ |

### A. Proactive Routing Protocols

The proactive routing means that the routing information, like next forwarding hop is maintained in the background irrespective of communication requests. The various types of proactive routing protocols are: OLSR MOPR, DSR, DSDV.

OLSR [8] (Optimized Link State Routing Protocol) is an optimization of a pure link state protocol for mobile ad hoc networks. Each node in the network selects a set of neighbor nodes called as multi-point relays (MPR) which retransmits its packets. The neighbor nodes which are not in its MPR set can only read and process the packet.

MOPR [9] is a movement prediction-based routing, which optimizes the procedure of selecting the MPR (Multi-Point Relay) sets in OLSR routing protocol as well as that of determining the optimal path from each pair of vehicles.

GSR [10] (Geographic Source Routing) is improved to use in VANET scenario by incorporating in to it greedy forwarding of messages toward the destination. If at any hop there are no nodes in the direction of destination then GPSR utilizes a recovery strategy known as perimeter mode.

DSDV [11] is a table driven algorithm based on Bellmen-ford routing mechanism. In this protocol, every mobile node maintains a routing table in which all of the possible destinations within the network and the number of hops to each destination are recorded. Each entry is assigned a sequence number by the destination node. The sequence numbers enable the mobile nodes to distinguish stale routes from new ones, thereby avoiding the formation of routing loops.

Proactive routing protocols tend to update routing information periodically, which perform satisfactorily in city environments (with low mobility) but shows degraded performance in highly mobile and dense scenarios when compared with reactive routing protocols. The advantage of proactive routing protocols is that there is no route discovery since the destination route is stored in the background, but the disadvantage of this kind of protocols is that it provides low latency for real time application.

### B. Reactive Routing Protocols

Reactive routing opens the route only when it is necessary for a node to communicate with each other. Reactive routing consists of route discovery phase in which the query packets are flooded into the network for the path search and this phase completes when route is found. Reactive routing protocols are very popular in mobile ad hoc network. The various types of reactive routing protocols are AODV, S-AODV, PAODV, DSR.

DSR (Johnson, 1996) uses source routing to indicate the sequence of intermediate nodes on the routing path, which is not suitable to high mobility VANET.

AODV minimizes the number of required broadcast by creating routes on an on-demand basis to improve the DSDV based protocols. When a node wants to send a message to another node with an invalid route, it initiates a Path Discovery process to locate the destination node at first [12]. It broadcasts a route request (RREQ) packets to its neighbors, which then forward the request to their neighbors, and so on, until either the destination or an intermediate node with a newly route to the destination node.

Some VANET specific AODV extensions have been also proposed. The S-AOMDV [13] protocol combines speed with hop-count as a routing metric and shows improved performance in VANET. Prior AODV (PAODV) [14] routing protocol restricts the number of discovered routes to nodes between threshold distance and transmission range in order to reduce routing overhead.

### *C. Hybrid Routing Protocol*

The hybrid protocols are introduced to reduce the control overhead of proactive routing protocols and decrease the initial route discovery delay in reactive routing protocols. The various types of reactive routing protocols are: GPSR, GPSR-L, GPCR, GpsrJ+.

Greedy Perimeter Stateless Routing (GPSR) [15] uses a greedy forwarding technique that selects next hop which is closest to the destination. Each node in GPSR keeps states from immediate neighbors and uses only those states for data forwarding. Forwarding nodes run greedy mode routing, which firstly selects a node whose distance to a destination is shortest among all immediate neighbors and then drives the routing nodes to forward data to the destination node. If there is no neighbor whose distance to destination is greater than distance from forwarding node to destination, forwarding node runs perimeter mode routing. The state is geographic position that all sensor nodes can self-configure through GPS or others location devices. Source node propagates data with the position of destination to vehicle ad hoc network.

GPSR with Life time (GPSR-L) [16] improves the performance of GPSR by selecting the next hop neighbor with good link quality and a nonzero life-time. Greedy Perimeter Coordinator Routing (GPCR) [17] further improves the repair strategy of GPSR. GPCR uses nodes situated at junctions to decide which street the information should traverse in order to reach next junction and finally destination.

Subsequently, the furthest node is selected as the next hop to forward the information. This way GPCR outperforms GPSR in city environments where interference is higher. Authors in [18] proposed prediction based GpsrJ+ to improve the recovery strategy of GPCR. GpsrJ+ outperforms GPCR and GPSR in terms of packet delivery ratio and hop count.

In order to apply a suitable routing protocol for VANET, or design an efficient routing protocol, we choose these three typical protocols to study their performance of multimedia data transmission in highway scene.

## III. EVALUATION MODEL AND TEST BED

There are two approaches to support video transmitted over VANET: V2V (Vehicle to Vehicle) approach and V2I (Vehicle to Infrastructure) approach. In this paper, we are interested in investigating the problem of supporting video transmitted over VANET using the first approach, because the first approach doesn't need any roadside infrastructure, it's easier to deploy. In this section, we mainly present an evaluation model, and design a simulating platform to evaluate performance of quality of multimedia under different routing protocols.

Many literatures have done some research on video transmission over VANET ( [19]). Since our goal is to evaluate the quality of video transmitted over VANET under different traffic conditions and different routing protocols, we mainly consider following characters:

1) Wireless channel quality can be easily affected by many factors, including street construction, road conditions, vehicle type and so on;

2) High dynamics of mobile nodes, which may incur frequent link disconnection and even network partition;

3) Application background is totally different from general Ad hoc network, while VANET is specially designed to achieve multi-hop communication between vehicle-to-vehicle and vehicle-to-Infrastructure on road.

So, we have to set a suitable evaluation model and then evaluate the quality of video transmitted in a realistic simulation environment [20]. In order to better match the reality, we use real video data and realistic vehicle mobility traces to evaluate the performance of video transmitted over VANET.

### *A. Evaluation Models*

In this paper, we use two important performance parameters to evaluate the quality of video transmitted over VANET: Frame Loss Rate and PSNR.

1) Frame Loss Ratio Model

The frame loss rate is a fundamental criterion of performance evaluation. In video streaming, a single video frame is decomposed into many smaller packets and sent into the network, so decoded video quality at the receiver is affected by two factors: encoder compression performance and distortion due to the packet loss or late arrivals.

Based on the model in [13], the video distortion can be modeled as:

$$D_{dec} = D_{enc} + D_{loss} . \quad (1)$$

The encoder distortion can be modeled by:

$$D_{enc} = D_0 + \theta / (R - R_0) . \quad (2)$$

Where R is the rate of the video stream, and the parameters, $D_0$ and $R_0$ are estimated from empirical rate-distortion curves via regression techniques.

In this paper, we mainly care about the packets lost rate, represented by criterion $D_{loss}$. If the percentage of

lost packets excesses the bound of error correction, the receiver cannot playback this frame. Similarly, if the packet arrival time is later than the playback deadline of the corresponding frame, it will also be dropped by the decoder in receiver cache. Therefore $D_{loss}$ can be modeled by:

$$D_{lose} = P_{loss} + P_{delay} . \quad (3)$$

In our study, we consider the Frame Loss Rate as the performance parameter based on (1). We use a video file "foreman.yuv" which has 400 frames, which contains 45 I frames, 89 P frames and 266 B frames. For each frame type, we measure the lost frames separately, and then we calculate the percentage of the overall frame loss. Overall Frame Loss Rate %= (Lost I frames + Lost P frames + Lost B frames) *100 /Total Number of Frames.

2) PSNR model

Peak Signal-to-Noise Ratio (PSNR) is an important criterion to measure the error of image between the reconstructed and the original frame by frame. The PSNR has become the most widespread objective metric used to assess the application-level QoS of video transmissions [21].

For frame n, its PSNR between the source image S and destination image D is defined by PSNR model as:

$$PSNR(n)_{dB} = 20\lg\left(\frac{V_{peak}}{\mathrm{Q_{SD}}}\right). \quad (4)$$

$$\mathrm{Q_{SD}} = \sqrt{\frac{1}{\mathrm{N}_{col}\mathrm{N}_{row}}\sum_{i=0}^{N_{col}}\sum_{j=0}^{N_{row}}\left[Y_S(n,i,j) - Y_D(n,i,j)\right]^2} \quad (5)$$

Where Y denotes the luminance component, $V_{peak} = 2^k - 1$ and n = number of bits per pixel (luminance component). $Y_S(n,i,j)$ and $Y_D(n,i,j)$ are the values of the luminance component of nth frame at pixel for the source and destination images respectively. $\mathrm{N}_{col}$ and $\mathrm{N}_{row}$ are the dimensions of the frame. QSD indicates the difference between the source and destination images. Tab.II denotes the relationship of PSNR and the quality of images.

Table II.
QUALITY LEVEL OF IMAGES TO PSNR

| PSNR | >37 | 31~37 | 25~31 | 20~25 | <20 |
|---|---|---|---|---|---|
| Level | Excellent | Good | Fair | Poor | Bad |

3) Movement model of vehicle

In order to simulate real traffic scenario, we suppose that the car arriving rate and the distribution of cars' velocity Obey the law of normal distribution:

(1) The car arriving rate: it is under the parameter $\lambda_t$ (vehicles/second). So the distribution of separation distance of different cars Obeys the law of normal distribution under the parameter $\lambda_s$ (vehicles/meter).

(2) The distribution of cars' velocity: in free traffic status, the distribution of cars' velocity is under the probability density function:

$$f_V(v) = (1/\sigma\sqrt{2\pi})e^{\frac{-(v-u)^2}{2\sigma^2}} . \quad (6)$$

in which, u is the average velocity, σ is the shift of velocity.

So, the distribution of separation distance of different cars satisfies:

$$f_X(x) = \lambda_s e^{-\lambda_s x} . \quad (7)$$

Where, $\lambda_s \approx \lambda_t / \overline{V}$, $\overline{V}$ is the average velocity of cars on the road.

4) Connectivity model

Each vehicle is a node in the VANET as shown in fig.1.

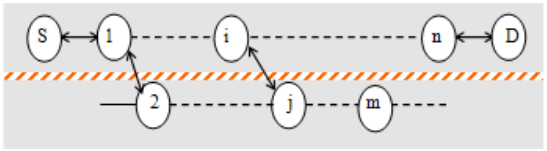



Figure 1. Link stability in terms of communication life time

From the above movement model of vehicle, $u_{eff}$ can be introduced as indication of $\overline{V}$ and presented as:

$$u_{eff} = [\mathrm{erf}(\frac{V_{\max}-u}{\sigma\sqrt{2}}) - \mathrm{erf}(\frac{V_{\min}-u}{\sigma\sqrt{2}})] / \int_{V_{\min}}^{V_{\max}} \frac{2 f_V(v)}{v} dv . \quad (8)$$

erf () is a function of error. So equation (7) changes to:

$$f_X(x) = (\lambda_t / u_{eff}) e^{-\frac{\lambda_t x}{u_{eff}}} \quad (9)$$

So, probability of a state when the distance of two adjacent cars is less than transmission range r is:

$$P(X \le r) = 1 - e^{-\lambda_s r} = 1 - e^{-\frac{\lambda_t r}{u_{eff}}} . \quad (10)$$

Then the Connectivity between source vehicle and destination vehicle d is presented:

$$P_c = (1 - e^{-\lambda_s r})^{N-1} = (1 - e^{-\frac{\lambda_t r}{u_{eff}}})^{N-1} . \quad (11)$$

where N is the hops of distance from the source to the destination.

*B. Test Bed*

To evaluate the performance of video transmitted over VANET, we propose a simulation testbed called VANET-EvalVid as shown in fig. 2. This tool-set integrates myEvalvid [21], and VanetMobiSim [22]. MyEvalvid is a tool-set for evaluating the quality of video transmitted over a real or simulated communication network, which combines Evalvid and NS2. VanetMobiSim is a tool for generating realistic vehicle mobility trace file.

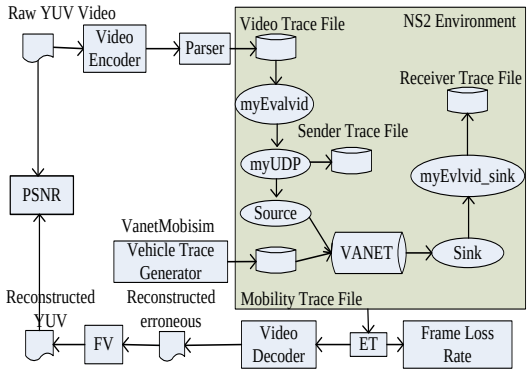



Figure 2. Framework of VANET-EvalVid

Main components of the VANET-EvalVid framework are as follows:

1) Video Source: The video source is responsible for generating video streams. The source video format can be either in YUV CIF (352 x 288) or YUV QCIF (176*144).

2) Video Encoder and Video Decoder: These coders are used to convert the video file from YUV format to MPEG4 format at the sender side and transfer it back to YUV format at the receiver side. Presently, EvalVid supports two MPEG-4 codecs: NCTU codec and ffmpeg. In this study, we apply ffmpeg for video coding.

3) Fix Video (FV): since video frames are compared frame by frame, the total number of video frames at the receiver end must match that of the original video at the sender end. If there are missing frames that video codec cannot be decoded, the FV is responsible for fixing them by substituting the last successfully decoded frame for each lost frame.

4) Evaluate Trace (ET): the ET component is responsible for the evaluation task. A evaluation begins at the sender side when the video transmission finishes. Information for delivering video packets is transported back to the sender side on a certain route. The ET generates a report of video frame/packet delay, frame/packet loss, and frame/packet jitter by comparing the trace files, which includes the original encoded video file, the video trace file, the sender trace file and the receiver trace file. In addition, the ET also creates a reconstructed video file, corresponding to the possible corrupted video frame at the receiver side. Generally, the generation of the possible corrupted video is a process of copying the original video trance file frame by frame and omitting frames indicated as corrupted or lost at the receiver side.

5) VanetMobiSim: this component is responsible for generating vehicle mobility trace file used as the mobility trace file in NS2. it is a simulation tool designed for generating vehicle mobility trace file, which supports both macro-mobility and micro-mobility representation to defined mobility models. The macro-mobility models is mainly used to define the characteristic of the roads, as lines, speed limit, traffic signs, etc.. Using the different options VanetMobiSim offers, we can define different types of scene about road status, vehicle number, driving speed and so on. The micro-mobility models include Intelligent Driver Model with Intersection Management (IDM-IM) and Intelligent Driver Model with Lane Changes (IDM-LC). IDM-IM is used to define the behavior when the driver is at the intersection, while IDM-LC is used to define the behavior when the driver is changing lane.

The communication interfaces: the interfaces between myEvalvid and NS2 are given by MyTrafficTrace MyUDP and MyUDPSink in test bed. (1) MyTrafficTrace is response to read the video trace file and extract the video frame type and the video frame size. Furthermore, MyTrafficTrace fragments the video frames into smaller video packets and sends these packets to the lower layer at the appropriate time according to the time settings in the simulation script file. (2) MyUDP enhances the original UDP component in NS2. This interface allows users to designate the output file name of the sender trace file; moreover, it also records the information of transmitted video packets, like the packet id, the timestamp, and the payload size. (3) MyUDPSink is a receiving agent for video packets sent by MyUDP. When receiving a video packet, it records information from the transmitted video packet, such as the packet id, the timestamp, and the payload size.

## IV. Simulation and Analysis

### *A. Simulation Scenario Setup*

We setup a scenario of transmitting real video data over a 2km road with a single-directional double lanes, which is generated by VanetMobiSim. Vehicles on the road can react according to the vehicles ahead, like changing lanes. Two different traffic environments is constructed for well-distributed traffic scenario and random-distributed traffic scenario, which are sparse and dens. Then we use the simulation tool-set introduced in section 3 to evaluate the quality of video transmitted under three different routing protocols. Main simulation parameters list in Tab. III.

TABLE III.
Simulation Parameters

| | |
|---|---|
| Simulator | NS2 2.28 |
| Routing protocols | DSDV, AODV, GPSR |
| Simulation time | 100s |
| Video file | Foreman.yuv |
| Communication Range | 300m |
| bandwidth | 6Mbps |
| MAC Protocol | IEEE 802.11DCF |
| Packet size | 1024 Bytes |

The video file used in this test is foreman.yuv containing 400 frames, which average PSNR is 34.89. According to DSRC in [23], we set the bandwidth 6Mbps and the communication range 300m. In our experiments, we mainly compare three different routing protocols: DSDV, AODV and GPSR, which introduced in section 2.

The main simulation step is as follows: video data is transmitted over a sparse traffic environment and a dense traffic environment separately to test performance in different scenarios. In each test, by changing the speed of

vehicles, the quality of video transmitted under different routing protocols can be measured. The video starts at 0.0 second and the simulation lasts 100 seconds.

*B. Simulation Results and Analysis*

1) In well-distributed traffic scenario

In this simulating scenario, all cars are well-distributed on the road. Figure 3 and Figure 4 shows the performance comparisons of 3 typical protocols when distance between sender and receiver has different hops. From this test result, DSDV is not suitable to VANET, since its proactive route does not adapt for the high mobility of vehicles.

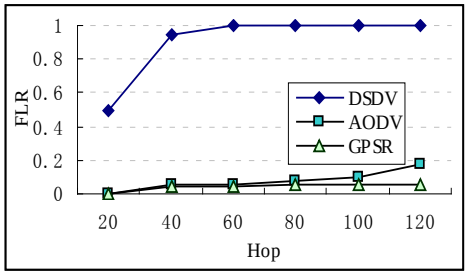



Figure 3. Comparison of Lost Packet Rate (LPR) & Hop

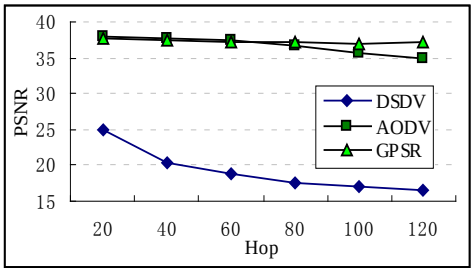



Figure 4. Comparison of PSNR & Hop

2) In random-distributed traffic scenario

In this scenario, we test the network performance when the arriving rate of car $\lambda_t$ is different.

Connectivity Probability is mainly related to distance of adjacency cars and transmission area owning to transmitting power.

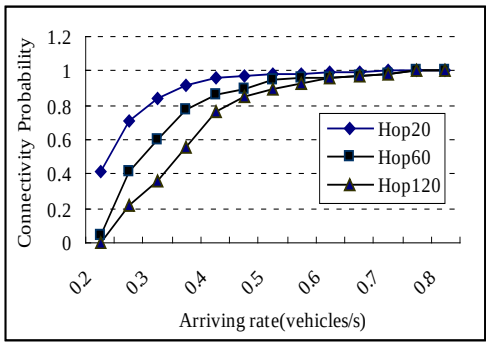



Figure 5. Comparison of Connectivity & Arriving rate

Fig. 5 shows that rising of hops between the source (the sender) and the destination (the receiver) results in lower connectivity at the same arriving rate, but the connectivity probability increases with the arriving rate of vehicles, since the density of vehicles in the network increases with the rate, which assures the link state. And if the arriving rate is high enough, the connectivity probability will get stable.

When the traffic is sparse, we suppose that: the average velocity of cars $\overline{V}$ is 20m/s, and the arriving rate λt=0.3 vehicles/s, all cars distribute on a 2km road. From the results in fig.6 and fig.7, the Frame Loss Rate of DSDV increase greatly with the increase in vehicle speed and the video in DSDV cannot be reconstructed when the distance is more than 40 hops, while GPSR protocol plays best. The network is more likely to disconnect with the increase in vehicle speed, so the Frame Lost Ratio of different protocols is the biggest. Figure 7 shows the average PSNR in AODV and GPSR are all greater than 32dB. it indicates that the quality of received video is at the excellent level.

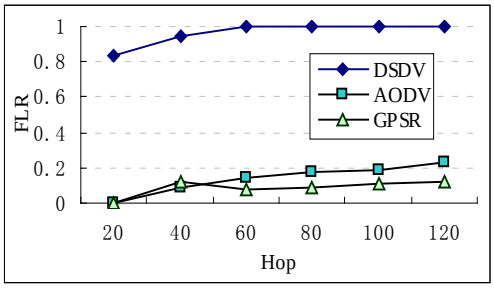



Figure 6. Frame Lost Rate in different protocols

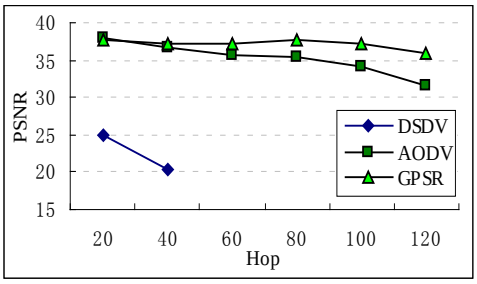



Figure 7. Average PSNR in different protocols

When the traffic is dense, we suppose that: the average velocity of cars $\overline{V}$ is 20m/s, and the arriving rate λt=0.9 vehicles/s, all cars distribute on a 2km road. From fig. 8 and fig. 9, DSDV protocol can't serve the video transmission when the distance is more than 40 hops. GPSR protocol plays best, with the increase in vehicle speed. The simulation result shows that Position-based protocol can better handle the video transmission over high mobility scenario. The main reason is that the position-based protocol does not rely on the maintained neighboring information which is likely inaccurate in high mobility scenario.

From the above result and analysis, three class routing protocols play quite different in VNET. Proactive protocols such as DSDV have considerable difficulties in maintaining valid routes, and lose many packets because of that. With increasing mobility, it strives to continuously maintain routes to every node increases network load as updates become larger. The rapidly

changing routes through the fast vehicle nodes are required for inter-group traffic and are fairly long. So such protocol cannot adapt well to such fast route changes.

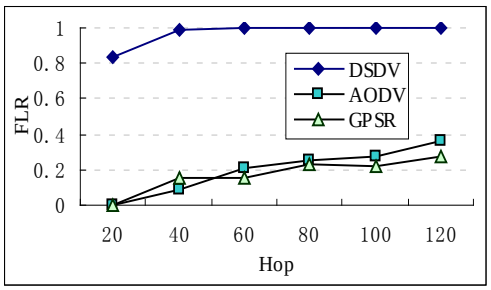



Figure 8. Frame Lost Rate in different protocols

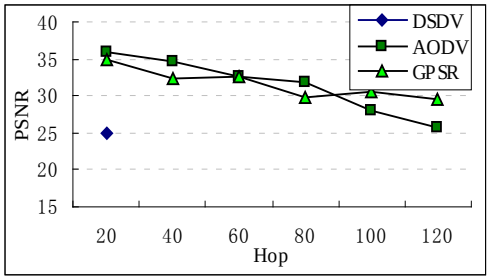



Figure 9. Average PSNR in different protocols

In reactive routing protocols like AODV, uncontrolled flooding messages generate redundant transmissions, which may cause broadcast storm problem, and the network scalability suffers from the increasing administrative load as the number of vehicle nodes grows. Since the forwarding node tries to find a node that is as close as possible to the location stored in the packet header if a routing is in greedy mode, where it simplifies the check if the destination node is in its neighborhood. The simulation results also show that the quality of video transmitted in dense traffic scenario is better than that in sparse traffic scenario, which is similar within [24]. This result is quite similar with other research, and it testifies our model correct and efficiency of evaluation.

## V. Conclusion and Future Work

In this paper, we presented three main evaluation models of QoS performance: frame loss ratio, PSNR and connectivity probability, and we designed a simulation platform of video communication over VANET. With the integration of simulation platform, researchers can easily evaluate the quality of video transmitted over VANET and analyze their designed mechanisms, such as network protocols or QoS control schemes in a realistic simulation environment. We created two different scenarios to test the performance of multimedia data transmission under different routing protocols. From the test result and its analysis, Pro-active protocol is not suitable for video transmission over VANET, and Position-based protocol is more suitable for video transmission over VANET than Re-active protocol, and it also illustrates scalability of different protocols. In the future work, we'll compare more protocols to choose a better routing protocol for video transmission over VANET. The result testifies the correctness of our model and the efficiency of evaluation platform.

## Acknowledgment

This work was partially supported by National Natural Science Foundation of China (61174177) and Yichang Science and Technology Foundation (A2011-302-13).

## References

[1] H. Hartenstein, et al., "A tutorial survey on vehicular ad hoc networks", in IEEE Communications Magazine, vol. 46, no. 6, pp. 164-171, 2008.

[2] Xin Su, "A comparative survey of routing protocol for vehicular sensor networks", Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference, pp. 311-316, 25-27 June 2010.

[3] R. Fracchia, et al., "Alert service in VANET: Analysis and design", Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, 2006 4th International Symposium on, pp. 1- 8, 03-06 April 2006.

[4] Y. Toor, et al., "Vehicle Ad Hoc networks: applications and related technical issues", Communications Surveys & Tutorials, IEEE, vol. 10, no. 3, pp. 74-88, Third Quarter 2008.

[5] M. Meenakshi, "A Study of Live Video Streaming over Highway Vehicular Ad hoc Networks", International Journal of Computer Applications, 2010, 1 (21), pp. 86-90.

[6] Y. Saleh, et al., "Vehicular ad hoc networks (VANETs): Challenges and perspectives", in ITS Telecommunications Proceedings, 2006 6th International Conference. pp. 761-766, June 2006.

[7] H. Trivedi, et al., "Routing Mechanisms and Cross-Layer Design for Vehicular Ad Hoc Networks: A Survey", in Computers & Informatics (ISCI), 2011 IEEE Symposium, pp. 243 – 248, 20-23 March 2011.

[8] T. Jamal, et al., "Optimizing OLSR in VANETS with differential evolution: A comprehensive study", in DIVANet '11 Proceedings of the first ACM international symposium on Design and analysis of intelligent vehicular networks and applications, pp. 1-8, 2011.

[9] H. Menouar, et al., "Improving Proactive Routing in VANETs with the MOPR Movement Prediction Framework", in Telecommunications, 2007. ITST '07. 7th International Conference on ITS, pp. 1-6, 6-8 June 2007.

[10] L. Lichuan, et al., "A Geographic Source Routing Protocol for Traffic Sensing in Urban Environment", in Automation Science and Engineering, August 2008, pp. 347 – 352.

[11] Spaho, Evjola, Wajiro-Higashi, Higashi-Ku, "Performance comparison of DSDV and DYMO protocols for vehicular ad hoc networks", International Conference on Advanced Information Networking and Applications, AINA, p 629-634, 2012,

[12] C. Perkins, E. Royer, "Ad-hoc on-demand distance vector routing", In: Proc.of IEEE WMCSA'99, New Orleans, Louisiana, Feb.1999, 90-100.Services, 2009, pp. 518-521.

[13] C. Yufeng, et al., "An improved AOMDV routing protocol for V2V communication", in Proc. IEEE Intell. Vehicles Symposium, 2009, pp. 1115-1120.

[14] O. Abedi, et al., "Improving route stability and overhead on AODV routing protocol and make it usable for VANET, " in Proc.29th IEEE Int. Conf. on Distributed Computing Systems Workshops, 2009, pp. 464-467.

[15] Zhang Ning, Yunho, Jung, Yan, Jin, Kim, Kee-Cheon, "Route optimization for GPSR in VANET", 2009 IEEE International Advance Computing Conference, p 569-573, 2009
[16] S. A. Rao, et al., "GPSR-L: Greedy perimeter stateless routing with lifetime for VANETS", in ITS Telecommunications, 2008. ITST 2008. 8th International Conference, pp. 299 – 304, 24-24 Oct. 2008.
[17] C. Lochert, et al., "Geographic routing in city scenarios", ACM SIGMOBILE Mobile Computing and Communications Review, vol. 9, pp. 69-72, 2005.
[18] K. C. Lee, et al., "Enhanced Perimeter Routing for geographic forwarding protocols in urban vehicular scenarios, " in Proc. IEEE Globecom Workshops, 2007, pp. 1-10, 26-30 Nov. 2007.
[19] M. Guo, et al., "V3: A vehicle-to-vehicle live video streaming architecture", In Pervasive Computing and Communications, 2005. PerCom 2005. Third IEEE International Conference, pp. 171 – 180, 8-12 March 2005.
[20] K. Chih-Heng, et al., "A Novel Realistic Simulation Tool for Video Transmission over Wireless Network", In Sensor Networks, Ubiquitous, and Trustworthy Computing, 2006. IEEE International Conference, 5-7 June 2006.
[21] J.Klaue, et al., "EvalVid - A Framework for Video Transmission and Quality Evaluation", In Proc. of the 13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, pp. 255-272, Illinois, USA, September 2003.
[22] VanetMobiSim, available at http: //vanet.eurecom.fr
[23] He Jianhua, Tang Zuoyin, O'Farrell Tim, Chen Thomas M, "Performance analysis of DSRC priority mechanism for road safety applications in vehicular networks", Wireless Communications and Mobile Computing, v 11, n 7, p 980-990, July 2011
[24] G. Wenyang "Adaptive Rate Control of Dedicated Short Range Communications Based Vehicle Networks for Road Safety Applications", in Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd, p.1-5, 15-18 May 2011.

**Shouzhi Xu** is a Professor of System Architecture at the College of Computer and Information Technology at China Three Gorges University, Yichang, Hubei Province, China. He received a Ph.D. in Information and Communication Engineering from the Huazhong University of Science and Technology in China. He is a Fellow of the China Computer Federation.

His Current research interests include wireless sensor network, vehicle ad hoc network and network security. His research has been founded by National Natural Science Foundation of China (NSFC 61174177), Foundation of Hubei Provincial Department of Education and Science and Technology Foundation of Yichang.

**Pengfei Guo** has received his B.A's degree in computer science and technology in China Three Gorges University, Yichang, Hubei Province, China, in 2010.

He has been a graduate student of China Three Gorges University since 2010. His main research interests are wireless sensor network and vehicle ad hoc network.

# Cloud Computing for Network Security Intrusion Detection System

Jin Yang
[1] School of Information Science & Technology, Southwest Jiaotong Univ., Chengdu, China
[2] Department of Computer Science, LeShan Normal Univ., LeShan 614000, China
Email: jinnyang@163.com

[1,] Cilin Wang, [2,] Caiming Liu, [3,] Le Yu
[1, 2,] Department of Computer Science, LeShan Normal Univ., LeShan 614000, China
[3,] Military Representative Office of PLA, Guiyang, China
Email: bigluckboy@163.com

***Abstract*—In recent years, as a new distributed computing model, cloud computing has developed rapidly and become the focus of academia and industry. But now the security issue of cloud computing is a main critical problem of most enterprise customers faced. In the current network environment, that relying on a single terminal to check the Trojan virus is considered increasingly unreliable. This paper analyzes the characteristics of current cloud computing, and then proposes a comprehensive real-time network risk evaluation model for cloud computing based on the correspondence between the artificial immune system antibody and pathogen invasion intensity. The paper also combines assets evaluation system and network integration evaluation system, considering from the application layer, the host layer, network layer may be factors that affect the network risks. The experimental results show that this model improves the ability of intrusion detection and can support for the security of current cloud computing.**

***Index Terms*—Cloud Computing; Artificial Immune Systems; Network Security**

## I. INTRODUCTION

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software, and information are provided to computers and other devices as a metered service over a network. It has so many advantages such as economy, complex calculations, agility, high scalability, high reliability, easy maintenance. The concept of cloud computing was born in the 1960s from the ideas of American computer scientist J.C.R. Licklider and John McCarthy stated that computing will become a publicly available service in the future[1]. In 1983, Sun Microsystems bring forward a singular vision that "the network is the computer" [2]. On August 09, 2006, the CEO Google, Eric Schmidt, firstly mentioned the concept of Cloud computing on SES San Jose 2006. On Jan.30th, 2008, Google declared "Cloud Computing Research Plan" in Taiwan and will promote the advanced technology in Taiwan's colleges. Feb 1th 2008, IBM (NYSE: IBM) announced it will establish the first Cloud Computing Center for software companies in China, which will be situated at the Wuxi-TaiHu New Town Science and Education Industrial Park, China. On March 5, 2010, Novell and CSA released a supplier neutral plane, named as Trusted Cloud Initiative. May 22, 2009, China's first Cloud Computing Conference held in Beijing China World Hotel. January 22, 2010, China cloud computing technology and industry alliance (CCCTIA) announced in Beijing.

Cloud computing through the network environment make the complex computational processing program to split into numerous smaller subroutines, and then handed over the analyzed results back to the user. It is a kind of typical network computing mode, which emphasizes on large-scale virtual computing environment to run application scalability and availability. It has become the focus of great concern to the industry, the academia, and even the government. But now, with the increasing popularity of cloud computing, the importance of network security in cloud computing is rising, which has become the important factor of restricting its development. The traditional network security approaches include virus detection, frangibility evaluation, and firewall etc., e.g., the Intrusion Detection System (IDS) [3]. They rely upon collecting and analyzing the viruses' specimens or intrusion signatures with some traditional techniques. Moreover, being lack of self-learning and self-adapting abilities, they can only prevent those known network intrusions, and can do nothing for those variety intrusions.

Recent years, the artificial immune system has the features of dynamic, self-adaptation and diversity [4-7] that just meet the constraints derived from the characteristics of the grid environment, and mobile agent has many same appealing properties as that of artificial immune system. Negative Selection Algorithm and the

This work was supported by China Postdoctoral Science Foundation (No.2011M501419), and the National Natural Science Foundation of China (No.61003310, No.61103249) and the Scientific Research Fund of Sichuan Provincial Education Department (No. 10ZB005).
† Corresponding author.


doi: 10.4304/jnw.8.1.140-147

concept of computer immunity proposed by Forrest in 1994 [8-9]. In contrast, the AIS theory adaptively generates new immune cells so that it is able to detect previously unknown and rapidly evolving harmful antigens [10]. However, much theoretical groundwork in immunological computation has been taken up, but there is a lack of perfectly systems based AIS of dynamical immunological surveillance for network security [11, 12]. Based on the correspondence between the artificial immune system antibody in the artificial immune systems and pathogen invasion intensity, this paper is to establish a network risk evaluation model [13, 15]. We built a hierarchical, quantitative measurement indicator system, and a unified evaluation information base and knowledge base. This model will help the network managers evaluate the possibility and the graveness degree of the network dangerous quickly, ease the pressure of recognition, to get targeted immediate defense strategy of the strength and risk level of the current network attacks.

This article applied AIS technique in the field of network security situation awareness, designed and established an immune network security situation awareness system. It is aimed to carry out in cloud computing environment on real-time monitor the network security situation, realize real-time and quantitative awareness of network security situation before malicious network behavior becomes out of control, and help make timely and effective network security strategy adjustment for better general security safeguard of system.

## II. THE IMMUNE-BASED INTRUSION DETECTION

Biological Immune System (BIS) is a complicated system with the ability of self-adapting, self-learning, self-organizing, parallel processing and distributed coordinating, and it also has the basic function to distinguish self and non-self and clean non-self. The problems in the field of computer security and Artificial Immune Systems have the astonishing similarity of keeping the system stable in a continuous changing environment. Artificial Immune System can use biological immune theoretic for references to search and design relevant models and algorithms to solve the various problems occurred in the field of computer security. Technology for Network Security Situational Awareness, which is a positive defense technology, has become the orientation of research in the field of network security. Based on the analysis of the papers from domestic and foreign on technologies for network security situational awareness, this paper designs and builds a network security situational awareness system based on the profound research of BIS. The system uses network intrusion detection, which based on the theory of biological immunity as the base of situational awareness, to detect known and unknown intrusions with the help of biological technology such as self/non-self discrimination, self-tolerance, self-learning, evolution mechanism, immunological surveillance, etc. According to correspondence relations of density change of antibody in the artificial immune systems and pathogen invasion intensity, a novel network risk evaluation model is also established. Based on the current real-time network risk evaluation, the thesis also makes risk evaluation on short-term, medium-term、long-term network in different span. These methods make overall and quantitative identification about network security, and it is also helpful to resemble network security tragedy effectively, therefore, protect network infrastructure greatly.

Simulating creatural immune system, we place a certain amount of immune cells into the network, and perceive the surrounding environment of the detectors. As soon as the immune detectors detect an attack, the detectors begin clone and generate a mass of similar detectors in order to defend from fiercer network attacks and warn the dangerous level of the network. The network security situation awareness agent in cloud computing environment shown in Figure 1 is itself a sub-network security situation awareness system, defined by recursion, and it mainly monitors on the sub-network security situation within its control, and specifically speaking, real-time monitor on the type, strength and harmfulness of attacks suffered by sub-network. Because there might as well be subnets under subnets, sub-network security situation awareness agents may be composed of sub-network security situation awareness agents at lower level. Eventually, security situation awareness agents, that monitor the specific host computer, are made up of intrusion detection and security situation evaluation of the host computer.

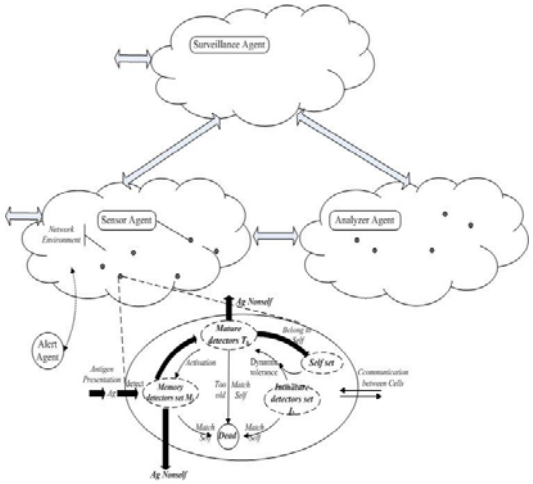



Figure 1. Architecture of the Evaluation of the Network Danger

TABLE I.
THE RELATIONSHIP BETWEEN THE BIS AND THE OUR MODEL

| Biological immune system | Network intrusion detection system |
|---|---|
| Organism | Network |
| Organ | Network segment |
| Cell | Host computer |
| Vaccine distribution | The transmission of intrusion information |
| Antigen | The binary character string feature-extracted from IP packets |
| B Cell, Antibody | Antibodies represented in binary character string |
| Cell clone | Duplication of antibody |

Combined with multi-agent and AIS technology, the detection model constitutes a multi-direction and multi-level intelligent network security model, with its mapping relationship with BIS model as shown in Table 1, and its system structure diagram as shown in Figure 1.

Agent of intrusion behavior distilling use vector space model and present the received datagram in discrete characters. Agent of Training generates various immature detectors from gene library to distinguish self and Non-self. According to immune principle, some of these new immature detectors are false detectors and they will be removed by the negative selection Agent, which matches them to the training datagram. If the match strength between an immature detector and one of the training datagram is over the pre-defined threshold, this new immature detector is considered as a false detector. Agent of intrusion surveillance matches the received datagram to the mature detectors. If the match strength between a received datagram and one of detectors, the behavior will be consider as the intrusion. The detail training phases are as following.

*A. Definition of Antigen, Antibody, Self and Non-self*

Definition: Antigens ( $Ag$ , $Ag \subset U$ , $D = \{0,1\}^l$ ) are fixed-length binary strings extracted from the Internet Protocol (IP) packets transferred in the network. The antigen consists of the source and destination IP addresses, port number, protocol type, IP flags, IP overall packet length, TCP/UDP/ICMP fields [16], etc. The structure of an antibody is the same as that of an antigen. For virus detection, the oneself set (Nonself) represents IP packets from a computer network attack, while the self set (Self) is normal sanctioned network service transactions and nonmalicious background clutter. Set $Ag$ contains two subsets, $Self \subseteq Ag$ and $Nonself \subseteq Ag$ such that

$$Self \cup Nonself = Ag \qquad Self \cap Nonself = \Phi$$

*B. The Dynamic Equations of the Mature Cells*

Let $G_b$ represent the set of all the mature cells, and G the amount of the mature cells in the set at some time. Then the dynamic equation of the mature-cell set is formulated as following:

$$G(t+\Delta t) = G(t) + g_{new} \cdot \Delta t - (\frac{\partial G_{active}}{\partial x_{active}} + \frac{\partial G_{death}}{\partial x_{death}}) \cdot \Delta t .$$

It implies that the changing process of the set $G_b$ is separated into two stage: one is "flow into", the other is "flow out". The 1st stage when mature cells come into the set, $G = g_{new} \cdot \Delta t$ (where $g_{new}$ is the new mature cells in a unit time interval), the 1st stage shows the amount of mature cells which have flew into the set $G_b$ in the time interval $\Delta t$ ; the 2nd stage when mature cells flow out from the set, including two aspects: the mature cells which have been stimulated to be active and the dead mature cells. We denote $\partial G_{active} / \partial x_{active} \cdot \Delta t$ the differential amount of activated cells and $\partial G_{death} / \partial x_{death} \cdot \Delta t$ the differential amount of dead cells.

$G_{dead} = \{x \mid x.age > \lambda, x.cout < \beta\}$

During the course of the mature cells activated and dying, the following course is happening

$$age(t+1) = age(t) + 1 \qquad t < \lambda$$
$$affinity(t+1) = affinity(t) + 1 \qquad t < \lambda$$

where $\lambda$ means the affinity accumulating cycle.

Equation implies that mature cells must accumulate affinity with the antigen Ag. In one cycle $\lambda$ , with one unit of time interval, the age of the mature cell is added 1; if the matching of affinity is successful, that is, the function $f_{match}(x, y)$ holds, the *affinity* of the mature cell is added 1, Therefore, there must be one of following three cases:

① If $affinity \geq \theta \wedge t < \lambda$ , the mature cell is stimulated to be active and become the memory cell.

② If $affinity \geq \theta \wedge t < \lambda$ , the affinity of the mature cell isn't sufficient, the cell needs to go on accumulating affinity.

③ If $affinity < \theta \wedge t > \lambda$ , the cell can't accumulated enough affinity in one cycle $\lambda$ , so it has to die.

In the course, $\theta$ is the threshold of the affinity for the activated cells. The affinity function $f_{mathch}(x, y)$ may be any kind of Hamming, Manhattan, Euclidean, and r-continuous matching, etc. In this model, we take r-continuous matching algorithm to compute the affinity of mature cells. The matching functions utilize the following definitions:

$$f_{\text{math}}(x, y) = \begin{cases} 1 & \exists i, j, j - i \geq r \wedge 0 < i < j \leq l, \\ & x_i = y_i, x_{i+1} = y_{i+1}, \cdots, x_j = y_j \\ 0 & otherwise \end{cases}$$

The r-continuous matching is commonly used method for measuring the distance between bit strings with the goal of producing a better similarity coefficient.

*C. The Dynamic Equation of Memory Cells*

Let $M_b$ the set of memory cells, M the amount of the memory cells in the set at some time. Because memory cells are more difficult to come into being, in this paper, the changing process of the set $M_b$ only includes stage "flow into" except the dying stage. Memory cells totally come from the activated mature cells $G_{active}$ , that is, the dynamic equation of memory cells is:

$$M(t+\Delta t) = M(t) + \frac{\partial M_{active}}{\partial x_{active}} \cdot \Delta t + M_{other\_host}(\Delta t) - M_{dead}(\Delta t)$$

$$M_{dead}(t) = \{x \mid x \in M(t), f_{match}(x, Self(t-1)) = 1\}$$

Equation describes the dynamic process of the memory cell set. (where $\frac{\partial M_{active}}{\partial x_{active}} \cdot \Delta t = \frac{\partial G_{active}}{\partial x_{active}} \cdot \Delta t$ )

$$M(t+\Delta t) = M(t) + M_{new}(\Delta t) + M_{from\_other}(\Delta t) - M_{dead}(\Delta t),$$
$$when \quad f_{match}(M(t), Ag(t)) \neq 1 \quad , \quad t > 1,$$

$$M(t+\Delta t)=M(t)+M_{clone}(t)+M_{new}(\Delta t)+M_{from\_other}(\Delta t)$$
$$-M_{dead}(\Delta t), when \quad f_{match}(M(t),Ag(t))=1$$
$$M_{clone}(t)=\frac{\partial M_{clone}}{\partial x_{clone}}\cdot\frac{\partial M_{active}}{\partial x_{active}}\cdot\Delta(t-1),$$
$$when \quad f_{match}(M(t),Ag(t))=1$$
$$M_{clone}(t+\Delta t)=M_{clone}(t), M.\rho(t+\Delta t)=M.\rho(t)+V_p\cdot\Delta t,$$
$$M.count(t+\Delta t)=M.count(t)+1$$
$$M.\rho(t+\Delta t)=\frac{1}{2}\cdot M.\rho(t), M.age(t+\Delta t)=M.age(t)+1,$$
$$when \quad f_{match}(M(t),Ag(t))\neq 1$$
$$M_{new}(\Delta t)=\frac{\partial M_{new}}{\partial x_{new}}\cdot\Delta t=\frac{\partial T_{active}}{\partial x_{active}}\cdot\Delta(t-1)$$
$$M_{new}.\rho(t)=\rho_0$$
$$M_{dead}(\Delta t)=\frac{\partial M_{death}}{\partial x_{death}}\cdot\Delta t, when \quad f_{match}(M(t-1),Self(t-1))=1$$
$$M_{from\_other}(\Delta t)=\sum_{i=1}^{k}(\frac{\partial M^{i}_{from\_other}}{\partial x_{from\_other}}\cdot\Delta t)$$

Equations depict the dynamic evolution of memory detector. $M(t+\Delta t)$ simulates the process that the memory detector evolve into the next generation ones. $M_{new}(t)$ is the set of memory detector that are activated by antigens lately. These mature detector matched by an antigen will be activated immediately and turn to a memory detector. $M_{dead}(t)$ is the memory detector that be deleted if it matches a known self antigen. $M_{clone}(t)$ is the reproduced memory detector when the detector distinguish a antigens. $M_{from\_other}(t)$ is the memory detector that transformed from other computers. The k indicates that the ID number of the computer. Therefore, dynamic model of immune is to generate more antibodies and enhance the ability of self-adaptation for the system.

### *D. The Dynamic Model of Self*

In a real-network environment some network services and activities are often change, which were permitted in the past but may be forbidden at the next time.

$$I(t)=I(0)=\{x_1,x_2,...,x_n\}, \qquad t=0$$
$$I(t+\Delta t)=I(t)+I_{new}(\Delta t)-I_{match\_self}(\Delta t)-I_{maturation}\cdot\Delta t, t>1$$
$$I.age(t+\Delta t)=I.age(t)+1$$
$$I_{match\_self}(t+\Delta t)=I(t), when \quad f_{match}(I(t-1),Self(t-1))=1$$
$$I_{maturatiio\,n}(t+\Delta t)=I(t), \qquad I.age(t+\Delta t)>\alpha$$
$$I_{new}(\Delta t)=(\xi_1\cdot\frac{\partial I_{\text{random}}}{\partial x})\cdot\Delta t+(\xi_2\cdot\frac{\partial I_{\text{inherit}}}{\partial x})\cdot\Delta t$$

Equation stimulates the dynamic evolution of self-antigens, where $x_i\in\Re(i\geq 1, i\in N)$ is the initial self element defined. Inew is the set of newly defined elements at time t, and Imaturation is the set of mutated elements. $f_{match}(y,x)$ is used to classify antigens as either self or nonself: if x is a self-antigen, return 0; if x is a nonself one, return 1; if x is detected as nonself but was detected as a self-antigen before, then it may be a nonself antigen (needs to be confirmed), and return 2. There are two advantages in this model. (1) Self immune surveillance: The model deletes mutated self-antigens (Imaturation) in time through surveillance. The false-negative error is reduced. (2) The dynamic growth of Self: The model can extend the depiction scope of self through adding new self-antigens (Inew) into Self. Therefore, the false-positive error is prevented.

### *E. The Antibody Variation*

In order to prevent algorithm from converging prematurely, we take variation operation to the gene set $G_1=\{g_1,g_2,\cdots,g_i,\cdots g_n\}$ after the cross process. Based on the analysis of premature convergence of traditional evolutionary programming, a novel multi-subgroup evolutionary programming algorithm is proposed. This set G consists in accordance with a certain percentage composition of random, variation resulting from the combination of antibody fragments produced from gene sets, in part generated by the inherited characteristics of their parents. Let $i\in I$ be expressed as immature individual detector and $I$ be the space for the individual detectors.

$t=0$;
*initialize* $I(0):\{i_1(0),i_2(0),\cdots,i_\mu(0)\}$;
*evaluate* $I(0):\{\Phi(i_1(0)),\Phi(i_2(0)),\cdots,\Phi(i_\mu(0))\}$;
while $(l(I(t))\neq True)$ do
evaluate: $I''(t):\{\ \Phi(I''(t))\ \}$
crossover: $I'(t):=c(I(t))$;
mutation: $I''(t):=m(I'(t))$

selection: If $(\mu,\lambda)$—selection
then $s(I''(t))$
else $s(I'(t)\cup I''(t))$;
t=t+1;
end

where c is the crossover operator and m is the mutation operation, and s stands for selecting into the next generation of groups. Normally in order to maintain the diversity of individuals, we make the composition of immature detector diversity, reflecting the characteristics of immune diversity.

### *F. The Process of Network Security Surveillance*

The self-adaptability, distributed character and quantization of antibody concentration that biological immune system bears are just the effective method to solve the technological problems of network security situation awareness, thus this article applies the characteristics of biological immune mechanism in the field of network security situation awareness research, and establishes immune network security situation awareness system to make real-time and quantitative analysis on network security condition and its changing trend. Therefore, system evaluates the network security by perceiving the danger around of them. The values of $G_b$ and $M_b$ reflect the intensity of intrusion in current network. The bigger the value G and M are, the more serious the network intrusion degree is. And the bigger

the value $\partial M_{active} / \partial x_{active}$ is, the quicker the change about network situation is. Through distinguishing the type of G and M, we can know different kinds of network intrusion. The values of $\lambda$ and $\theta$ reflect the activity degree of the mature cell. Therefore, with immune cells' status parameters and the parameters weight, we can get network danger situation and evaluate network security at real time. The following contents will elaborate how to establish this model.

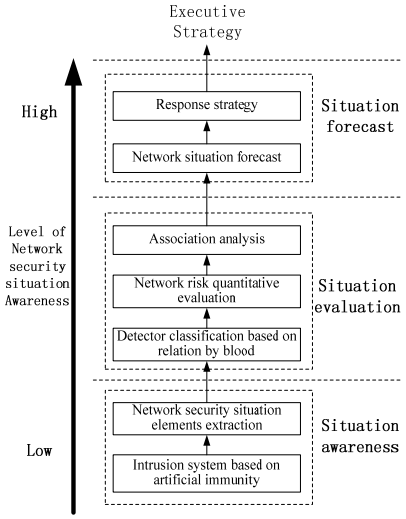



Figure 2. the structure of network security situation Surveillance

## III. EVALUATION OF NETWORK RISK

After we describe the network attacking actions, it is necessary to evaluate the dangerous degree of the network, and judge the severity of the attacking actions. Thus, evaluation is a process involving numerous complicated factors. Owing to the fact that our model relates to enormous factors for evaluation, on purpose of reasonably and entirely measuring the network dangerous status, we classify the involved factors as host dangers, area dangers, detectors dangers, and special dangers. Afterwards, we subdivide and arrange all the factors which influence the network dangers, in order to let them locate on different layers, forming a structure model with identify matrix.

Here we quantify these indicators, associated with the use of multi-level gray scale model. Suppose there are n types of indicators which can impact the Importancce-indicator in the network. And each Importancce-indicator has m kinds of attributes. In other words, we can use m kinds of attributes to measure and influence the values of the Importancce-indicator. To determine the evaluation indicator system based on the evaluation object, we use the following set of indicators to describe.

$$X_i' = \left(x_i'(1), x_i'(2), \cdots, x_i'(m)\right)^T, \quad i = 1, 2, \cdots, n$$

The n types of indicators sequence formed the following matrix:

$$\left(X_1', X_2' \cdots, X_n'\right) = \begin{pmatrix} x_1'(1) & x_2'(1) & \cdots & x_n'(1) \\ x_1'(2) & x_2'(2) & \cdots & x_n'(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_1'(m) & x_2'(m) & \cdots & x_n'(m) \end{pmatrix}$$

The reference data column should be an ideal standard of comparison, which can be composed of optimal value of each index (or the worst value). Depending on the purpose of the evaluation we can also choose another reference data. The reference data column recorded as

$$X_0' = \left(x_0'(1), x_0'(2), \cdots, x_0'(m)\right)$$

The indicators data sequence after dimensionless formed the following matrix.

$$\left(X_0, X_1, \cdots, X_n\right) = \begin{pmatrix} x_0(1) & x_1(1) & \cdots & x_n(1) \\ x_0(2) & x_1(2) & \cdots & x_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ x_0(m) & x_1(m) & \cdots & x_n(m) \end{pmatrix}$$

Here, the dimensionless method we adopted is the mean-quantization way.

$$x_i(k) = \frac{x_i'(k)}{\frac{1}{m}\sum_{k=1}^{m} x_i'(k)}, i = 0, 1, \cdots, n; \quad k = 1, 2, \cdots, m.$$

The absolute difference $|x_0(k) - x_i(k)|$ is individually calculated between each target sequence (sequence comparison) to the references sequences corresponding to the elements. And the values of $\min_{i=1}^{n}\min_{k=1}^{m}|x_0(k) - x_i(k)|$ and $\max_{i=1}^{n}\max_{k=1}^{m}|x_0(k) - x_i(k)|$ are also get out by calculated the correlation coefficient between each comparison sequence to the reference sequence corresponding to the elements.

$$\zeta_i(k) = \frac{\min_i \min_k |x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \cdot \max_i \max_k |x_0(k) - x_i(k)|}$$

$$k = 1, \cdots, m$$

Where the $\rho$ is the distinction coefficient, $\rho \in (0,1)$. The smaller the value of $\rho$ is, the greater the difference between the correlation coefficient will make, and the stronger the ability of distinguish is. Here we let $\rho = 0.5$.

Evaluation of the object (more sequences) were calculated with reference to its target sequence of m elements corresponding to the mean correlation coefficient to reflect the evaluation of the object and the reference sequence of association. Since every indicators play different role in the comprehensive evaluation of this system, we used a weighted average of the correlation coefficient requirements, where the value of $W_k$ is the weight of each index factors.

$$r_{0i}' = \frac{1}{m}\sum_{k=1}^{m} W_k \cdot \zeta_i(k) \qquad k = 1, \cdots, m$$

Eventually based on the observation of objects associated ordinal, the evaluation of the results is obtained.

Let $\rho_i(t)$ be the antibody concentration of jth host detect attacking at time t. Let $u$ be the danger coefficient of the ith kind of attack in the network. Then, we can get the danger level value Ri, j (t) facing the ith kind of attack as follows:

$$R_{i,j}(t) = \tanh(\zeta \cdot u \cdot \sum_{x \in A_i(t)} \rho(t))$$

The results of different abnormity behavior harm to a same host are different. Therefore, the comprehensive danger level value rj (t) is the linear weighted sum of the jth host facing all of the attacks. Let $u_i (0 \le u_i \le 1)$ be the relative weight value of danger of the ith kind of attack in the network. Then, we can define the danger level value $r_j(t)$ of the ith kind of attack as follows:

$$R_j(t) = \tanh(\zeta \cdot \sum_{i=1}^{n} (u_i \cdot \sum_{x \in A_i(t)} \rho_i(t)))$$

The entire network of danger level should fully reflect the value of each of the host facing attacks. As the host of each position is not the same such as running a different system for different users and providing different services, influencing different economic, affecting different social and even political values, they are in possession of different essentiality.

Let $\text{Importance}_i = \sum_{k=1}^{8} (I_k \times W_k)$ be the importance coefficient of jth host in the network. Then, we obtain the network entire danger level value: R (t) =∑ (indicator value × indicator weight). Therefore, we can get network danger R (t) situation and evaluate network security at real time.

$$R(t) = \tanh(\sum_{m=1}^{N} (\sum_{i=1}^{n} (Host_i's \ \ danger \times \text{Importance}_j) \times LCRS\_Weight_m))$$

$$= \tanh(\sum_{m=1}^{N} (\sum_{i=1}^{n} (Host_i's \ \ danger \times \sum_{k=1}^{8} (I_{j,\,k} \times W_k)) \times LCRS\_Weight_m))$$

$$= \tanh(\sum_{m=1}^{N} (\sum_{i=1}^{n} (r_i(t) \times \sum_{k=1}^{8} (I_{j,\,k} \times W_k)) \times LCRS\_Weight_m))$$

By applying the basic principle of the artificial immune system in the domain of network security situation awareness, the architecture is established which includes network security situation detection, network security situation evaluation, network security situation prediction. Through detecting malicious intrusions, in plus with real-time and quantitative analysis, prediction according to the current security situation and the future tendency, so as to make the network information system be self-learning and self-adapting as BIS, thus, to improve immune ability and survivability for web system, as well as alleviate damage made by network attack and enhance the emergency response ability.

## IV. Network Security Situation Forecast

Situation forecast is the highest level of situation awareness, it is based on historical and present network security situation information and makes quantitative prediction of the network security situation some period in the future so that decision-maker can have more complete network security situation and provides accurate grounds for reasonable response strategy to restrain network attack.

As to the fuzziness, randomness and uncertainty of future security situation change, it is put forth that gray theory can be adopted for establishing network security situation forecast model. Meanwhile, considering that network security situation awareness system is non-linear and the data is of high random fluctuation, Markov's state transition matrix is adopted to modify gray model's forecast results and make up for the limitation of gray forecast model. Therefore, gray theory and Markov theory are combined to bring the advantages of both to full play and overcome the defects of both, thus Gray Markov forecast model came into being. During the forecast process, the classical GM (1, 1) model of gray theory is adopted to make prediction of network security situation data and find out its changing trend, and then Markov theory is used to make modifications on model error to improve the forecast accuracy of network security situation changing trend.

According to the theory of time series analysis, we propose a new algorithm for network risk prediction. We divide the non-stationary time series into the determine sequence which represents a trend or cyclical regularity and the random sequence. Network intrusion affected by the combined effects of complex factors such as the social development, individual behavior and equipment and technology updates, which show the network risk situation a clear trend and randomness. This network intrusions behaviors mostly follow certain cycle regular of fluctuation. For example, the day average intrusion behaviors follow every 24 hours for the fluctuations in cycle regularity. We use the ARMA model. The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving-average terms.

$$X(t) = \phi_1 X(t-1) + \phi_2 X(t-2) + \ldots + \phi_p X(t-p) +$$

$$u(t) - \theta_1 u(t-1) - \theta_2 u(t-2) - \ldots - \theta_q u(t-q)$$

If q=0 → pure AR (p) process.

If p=0 → pure MA (q) process.

Introducing the lag operator B in our model, then, we can write:

$$\varphi(B)X(t) = \theta(B)u(t)$$

The predictive value of time series of the entire monitoring network risk equals $\{Y(t)\}$ the predictive of time series value of nonlinear fitting plus $\{X(t)\}$ the predictive value of the residual time series. We can write:

$$\hat{R}(t) = R(t) + X(t)$$

On the foundation of AIS based network risk evaluation, we also make evaluation towards short-term, medium-term, long-term network with different period, discuss the randomness of risk changes in real-time and short-time, as well as the periodicity in mid-term and long-term network risk changes. The model takes an overall overview on risk change tendency on every hierarchy of network from different viewpoints; therefore, it can build a safeguard system.

## V. Experimental Results and Analysis

The following experiments were carried out in the Laboratory of Computer Network Security. Considering the preciseness and efficiency, we use 12 indicators to evaluate the network danger, which include host danger, area danger, cells danger, special danger etc. An antigen was defined as a fixed length binary string composed of the source/destination IP address, port number, protocol type, IP flags, IP overall packet length, TCP/UDP/ICMP fields, and etc. The network was attacked by 25 kinds of attacks, such as Syn Flood, Land, Smurf, and Teardrop. A total of 20 computers in a network were under surveillance. The task aimed to detect network attacks. Figure 3 illustrates the syn attacks. Figure 4 illustrates the land attacks. And Figure 5 depict the evaluation of the network danger in our model. And we developed some series experiments. Here are the coefficients for the model as the Table 2 showing. As is shown in Figure 5, R (t) changes when attack levels changes. The rise in attack levels is accompanied by a corresponding increase in R (t), as implies the bad network security. On the other hand, if attack levels decline, R (t) decreases accordingly after seconds of delay. Therefore, the network can stays on guard even when the attacks occur once again during a very short time.

TABLE II.
COEFFICIENTS FOR THE MODEL

| – Parameter | – Value |
|---|---|
| – r-contiguous bits matching rule | – 8 |
| – The size of initial self set n | – 40 |
| – The Initial Scale of Detectors | – 100 |
| – Match Threshold β | – 40~60 |
| – Activable Threshold λ | – 50~150 |
| – Clone Scale | – 20 |
| – Mutation Scale | – 19 |
| – The Life Cycle of the Mature Detectors | – 120s |

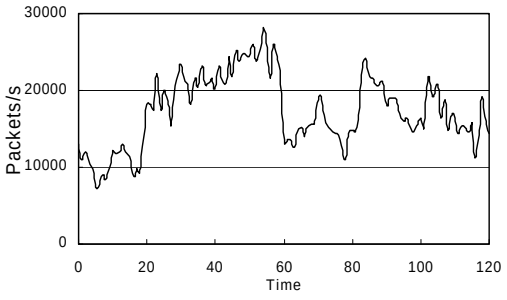



Figure 3. The network suffering from the syn incursions for instance

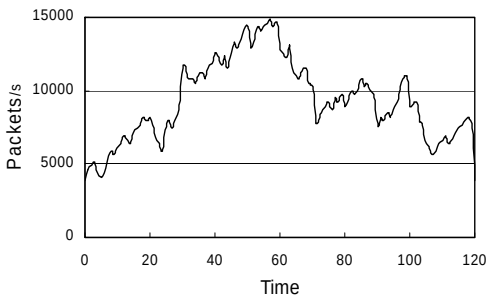



Figure 4. The network suffering from the land incursions for instance

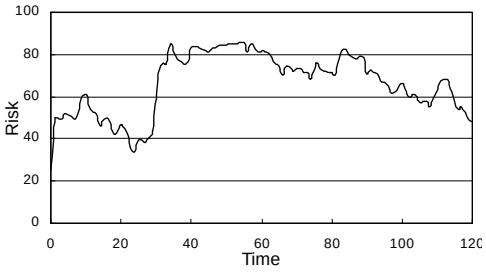



Figure 5. The line of the network dangers obtained by our model at these incursions

## VI. CONCLUSIONS

The self-adaptability, distributed character and quantization of antibody concentration that biological immune system bears are just the effective method to solve the technological problems of network security situation awareness, thus this article applies the characteristics of biological immune mechanism in the field of network security situation awareness research, and establishes immune network security situation awareness system to make real-time and quantitative analysis on network security condition and its changing trend. This paper combines the risk evaluation methods with application security engineering principles, and can change current passive defense situation using traditional network security approaches, and is helpful to establish new generation proactive defense theories and realization techniques. At the same time, the work is of not only theoretic values to design proactive defense systems which have intrusion tolerant ability and survivability in any complex network circumstances, but also very significant to protect network infrastructure. The experimental results show that the proposed model has the features of real-time processing that provide a good solution for network surveillance.

## ACKNOWLEDGMENT

This work was supported by China Postdoctoral Science Foundation (No.2011M501419), and the National Natural Science Foundation of China (No.61003310, No.61103249) and the Scientific Research Fund of Sichuan Provincial Education Department (No. 10ZB005).

## REFERENCES

[1] http: //en.wikipedia.org/wiki/Cloud_Computing.
[2] http: //www.mysql.com/news-and-events/sun-to-acquire-mysql.html
[3] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn, Chalermpol Charnsripinyo. Practical real-time intrusion detection using machine learning approaches, Computer Communications, vol. 34, pp. 2227-2235, 2011
[4] Huwaida Tagelsir Elshoush, Alert correlation in collaborative intelligent intrusion detection systems-A survey, Applied Soft Computing, vol. 11, pp. 4349-4365, 2011.

[5] Fatemeh Amiri, MohammadMahdi Rezaei Yousefi, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications, vol. 34 (4), pp. 1184-1199, 2011
[6] Vincent Toubiana, Houda Labiod, Laurent Reynaud. A global security architecture for operated hybrid WLAN mesh networks. Computer Networks, vol. 54 (2), pp. 218-230, 2010
[7] Kuby J., Immunology. Fifith Edition by Richard A. Goldsby et al.
[8] F.M.Burnet. The Clone Selection Theory of Acquired Immunity. Gambridge, Gambridge University Press, 1959
[9] S A Hofmeyr, and S Forrest. Architecture for an artificial immune system. Evolutionary Computation, vol. 8, pp. 443-473, 2000
[10] S Forrest, A S Perelson, L Allen, and R Cherukuri. Self-Nonself Discrimination in a Computer. Proceedings of IEEE Symposium on Re-search in Security and Privacy, Oakland, 1994
[11] Serap Atay, Marcelo Masera. Challenges for the security analysis of Next Generation Networks, Information Security Technical Report, vol. 16, pp. 3-11, 2011
[12] M. Mezmaz, N. Melab, Y. Kessaci. A parallel bi-objective hybrid metaheuristic for energy-aware scheduling for cloud computing systems, Journal of Parallel and Distributed Computing, vol. 71, pp. 1497-1508, 2011
[13] Md. Tanzim Khorshed. A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing, Future Generation Computer Systems, vol. 28, pp. 833-851, 2012
[14] Manel Bourguiba, Kamel Haddadou, Packet aggregation based network I/O virtualization for cloud computing, Computer Communications, vol. 35 (3), pp. 309-319, 2012
[15] Changbok, Hyokyung Chang, Filtering Technique on Mobile Cloud Computing, Energy Procedia, vol. 16, 1305-1311, 2012
[16] Tao Li. An immunity based network security risk estimation, Science in China Ser. F Information Sciences. vol. 48, no. 2005, pp. 557- 578

**Jin Yang** received his M.S. degree and the Ph.D. degree in computer science from Sichuan University, Sichuan, China. He is an Associate Professor in Department of Computer Science at LeShan normal university. His main research interests include network security, artificial immune, knowledge discovery and expert systems.

**Cilin Wang** received his M.S. degree in computer science from Sichuan University, Sichuan, China. He is an Associate Professor in Department of Computer Science at LeShan normal university. His main research interests include mathematics, knowledge discovery.

**LeYu** received his Bachelor's degree in computer science from Sichuan University, Sichuan, China. His main research interests include computer network, electronic technology.

**Caiming Liu** received his Ph.D. degree in computer science from Sichuan University, Sichuan, China. He is an Associate Professor in Department of Computer Science at LeShan normal university. His main research interests include network security and artificial immune.

# Secure Password-based Remote User Authentication Scheme against Smart Card Security Breach

Ding Wang*†, Chun-Guang Ma*, Qi-Ming Zhang*, Sendong Zhao
* College of Computer Science and Technology, Harbin Engineering University
145 Nantong Street, Harbin City 150001, China
Email: wangdingg@mail.nankai.edu.cn
† Automobile Management Institute of PLA, Bengbu City 233011, China

***Abstract*—It is a challenge for password authentication protocols using non-tamper resistant smart cards to achieve user anonymity, forward secrecy, immunity to various attacks and high performance at the same time. In 2011, Li and Lee showed that both Hsiang-Shih's password-based remote user authentication schemes are vulnerable to various attacks if the smart card is non-tamper resistant. Consequently, an improved scheme was developed to preclude the identified weaknesses and claimed that it is secure against smart card loss attacks. In this paper, however, we will show that Li-Lee's scheme still cannot withstand offline password guessing attack under the non-tamper resistance assumption of the smart card. In addition, their scheme is also vulnerable to denial of service attack and fails to provide user anonymity and forward secrecy. As our main contribution, a robust scheme is presented to cope with the aforementioned defects, while keeping the merits of different password authentication schemes using smart cards. The analysis demonstrates that our scheme meets all the proposed criteria and eliminates several hard security threats that are difficult to be tackled at the same time in previous scholarship.**

***Index Terms*—cryptanalysis, authentication protocol, network security, smart card, non-tamper resistant, user anonymity.**

## I. INTRODUCTION

Password-based authentication is widely used for systems that control remote access to computer networks. In order to address some of the security and management problems that occur in traditional password authentication protocols, research in recent decades has focused on smart card based password authentication. Since Chang and Wu [1] introduced the first remote user authentication scheme using smart cards in 1993, there have been many smart card based authentication schemes proposed [2-7]. In most of the previous authentication schemes, the smart card is assumed to be tamper-resistant, i.e., the secret information stored in the smart card cannot be revealed. However, recent research results have shown that the secret data stored in the smart card could be extracted by some means, such as monitoring the power consumption [8,9] or analyzing the leaked information [10]. Therefore, such schemes based on the tamper resistance assumption of the smart card are vulnerable to some types of attacks, such as user impersonation attacks, server masquerading attacks, and offline password guessing attacks, etc., once an adversary has obtained the secret information stored in a user's smart card and/or just some intermediate computational results in the smart card.

Another common feature of the published schemes is that the user's identity is transmitted in plaintext over insecure networks during the authentication process, which may leak the identity of the logging user once the login messages were eavesdropped, and thus user privacy is not preserved. The leakage of the user identity may also cause an unauthorized entity to track the user's login history and current location [5,7], as well as user's life styles and preferences. In many cases, it is of utmost importance to provide anonymity so that the adversary cannot trace user activity. Therefore, user anonymity is an important feature that a practical authentication scheme should achieve.

As noted by Blake-Wilson et al. [11], forward secrecy is an admired security feature for authentication protocols with session keys establishment. Particularly, forward secrecy is a property concerned with limiting the effects of eventual failure of the entire system. It indicates that, even if the long-term private keys of one or more entities are compromised, the secrecy of previous session keys established by honest entities should not be affected and thus the previous sessions shall remain secure [12,22]. Hence, a sound authentication scheme should achieve this important property.

As mentioned in Refs. [3,7,13-15] and the above description, the following criteria are important for smart card based remote user authentication schemes in terms of friendliness, security and efficiency: (C1) the server needs not to maintain a security-sensitive verification table; (C2) the password is memorable, and can be chosen freely by the user; (C3) the password cannot be derived by the privileged administrator of the server; (C4) the scheme is free from smart card loss attack, i.e., unauthorized users should not be able to easily change the password of the

Manuscript received June 1, 2012; revised July 1, 2012; accepted July 15, 2012. This is a substantially expanded and full version of a paper [29] that has been presented in DBSec 2012.
Wang Ding is the corresponding author


doi: 10.4304/jnw.8.1.148-155

smart card, or guess the password of the user by using password guessing attacks, or impersonate the user to login to the system, even if the smart card is obtained and/or secret data in the smart card is revealed; (C5) the scheme can resist various kinds of sophisticated (conventional) attacks, such as offline password guessing attack, replay attack, parallel session attack, denial of service attack, stolen verifier attack, user/server impersonation attack; (C6) the client and the server can establish a common session key during the authentication process; (C7) the scheme provides the property of timely wrong password detection, i.e. the user will be timely notified if he inputs wrong password by mistake in login phase; (C8) the scheme is not prone to the problems of clock synchronization and time-delay; (C9) the user can change the password locally without any interaction with the authentication server; (C10) the scheme can achieve mutual authentication; (C11) the scheme preserves user anonymity to avoid partial information leakage; (C12) the scheme provides the property of forward secrecy.

In 2004, Yoon et al. [16] proposed an advanced remote user authentication scheme using smart cards, their scheme possesses the merits of providing mutual authentication, no verification table, freely choosing password, involving only a few hashing operations, and so forth. Later on, Hsiang and Shih [17] showed that, in addition to parallel session attack, Yoon et al.'s scheme is vulnerable to offline password guessing attack, user impersonation attack if the smart card is non-tamper resistant. Consequently, an improvement over enhance Yoon et al.'s scheme is presented. In 2011, Li and Lee [18] identified that Hsiang-Shih's scheme still cannot withstand various attacks if the secret data stored in smart is revealed and further proposed an enhanced remote authentication scheme. They claimed their scheme is secure and can overcome all the identified security flaws of Hsiang-Shih's scheme even if the smart card is non-tamper resistant.

In this work, however, we will demonstrate that Li-Lee's scheme cannot withstand denial of service attack and is still vulnerable to offline password guessing attack under their assumption. In addition, their scheme does not provide the feature of forward secrecy and user anonymity. To conquer the identified weaknesses, a robust authentication scheme based on the secure one-way hash function and the well-known discrete logarithm problem is presented.

The remainder of this paper is organized as follows: in Section 2, we briefly review Li-Lee's authentication scheme. Section 3 describes the weaknesses of Li-Lee's scheme. Our proposed scheme is presented in Section 4, and its security analysis is given in Section 5. The comparison of the performance of our scheme with the other related schemes is shown in Section 6. Section 7 concludes the paper.

## II. REVIEW OF LI-LEE'S SCHEME

In this section, we briefly illustrate the remote user authentication scheme proposed by Li and Lee [18] in 2011. Their scheme consists of four phases: the registration phase, the login phase, the verification phase and password update phase. For ease of presentation, we employ some intuitive abbreviations and notations listed in Table 1.

TABLE I.
NOTATIONS

| Symbol | Description |
|---|---|
| $U_i$ | $i^{th}$ user |
| $S$ | remote server |
| $ID_i$ | identity of user $U_i$ |
| $P_i$ | password of user $U_i$ |
| $x$ | the secret key of remote server $S$ |
| $n$ | a large prime number |
| $g$ | a primitive element in Galois field $GF(n)$ |
| $h(\cdot)$ | collision free one-way hash function |
| $\oplus$ | the bitwise XOR operation |
| $\|\|$ | the string concatenation operation |
| $A \Rightarrow B: M$ | message $M$ is transferred through a secure channel from $A$ to $B$ |
| $A \rightarrow B: M$ | message $M$ is transferred through a common channel from $A$ to $B$ |

### A. Registration Phase

The registration phase involves the following operations:

*Step R1.* User $U_i$ chooses his/her identity $ID_i$ and password $P_i$, and then generates a random number $RN_1$.

*Step R2.* $U_i \Rightarrow S$: $\{ID_i, h(h(P_i \oplus RN_1))\}$.

*Step R3.* On receiving the registration message from $U_i$, the server $S$ creates an entry $\{ID_i, N, h(h(P_i \oplus RN_1))\}$ in the verification table, where $N = 0$ if it is $U_i$'s initial registration, otherwise $S$ sets $N = N + 1$. Then, the server $S$ computes $C_1 = h(ID_i \| x \| N) \oplus h(h(P_i \oplus RN_1))$.

*Step R4.* $S \Rightarrow U_i$: A smart card containing security parameters $\{ID_i, C_1, h(\cdot)\}$.

*Step R5.* Upon receiving the smart card, user $U_i$ stores $RN_1$ into the smart card.

### B. Login phase

When $U_i$ wants to login to $S$, the following operations will be performed:

*Step L1.* $U_i$ inserts his/her smart card into the card reader and inputs $ID_i$, $P_i$ and a random number $RN_2$.

*Step L2.* The smart card generates a random number $RC$ and then computes $C_2 = h(P_i \oplus RN_1)$, $C_3 = C_1 \oplus h(C_2)$, $C_4 = C_3 \oplus C_2$, $C_5 = h(h(P_i \oplus RN_2))$ and $C_6 = E_{K_{U_i}}(C_5, RC)$, where $K_{U_i} = h(C_2 \| C_3)$.

*Step L3.* $U_i \rightarrow S$: $\{ID_i, C_4, C_6\}$.

### C. Verification Phase

After receiving the login request message from user $U_i$, the server $S$ performs:

*Step V1.* The server $S$ checks the validity of identity $ID_i$ by checking whether $ID_i$ is already stored in its verification table. If not, the request is rejected. Otherwise, the $S$ computes $C_7 = h(ID_i \| x \| N)$, $C_8 = C_4 \oplus C_7$, $C_9 = h(C_8)$, and compares $C_9$ with the

third field of the entry corresponding to $ID_i$ in its verification table. If it equals, $S$ successfully authenticates $U_i$ and computes symmetric key $K'_{U_i} = h(C_8 \| C_7)$ , and obtains ( $C_5, RC$ ) by decrypting $C_6$ . Then, $S$ replaces the third field $h(h(P_i \oplus RN_1)$ of the entry corresponding to $ID_i$ with $C_5 = h(P_i \oplus RN_2)$ , generates a random $RS$ and computes $K_S = h(C_7 \| C_8)$ .

*Step V2.* $S \to U_i : \{E_{K_S}(RC, RS, C_5)\}$ .

*Step V3.* On receiving the response from server $S$, the smart card computes the symmetric key $K'_S = h(C_3 \| C_2)$ and obtains $(RC', C'_5)$ by decrypting the received message using $K'_S$ . Then, the smart card checks whether $(RC', C'_5)$ equals to $(RC, C_5)$ generated in the login phase. This equivalency authenticates the legitimacy of the server $S$ and replaces original $RN_1$ and $C_1$ with new $RN_2$ and $C_3 \oplus C_5$ , respectively.

*Step V4.* $U_i \to S : \{h(RS)\}$ .

*Step V5.* On receiving $h(RS)'$ , the serve $S$ compares the computed $h(RS)$ with the received value of $h(RS)'$ . If they are not equal, the connection is terminated.

*Step V6.* The user $U_i$ and the server $S$ agree on the session key $SK = h(RC \oplus RS)$ for securing future data communications.

### D. Password Change Phase

The password change phase is provided to allow users to change their passwords freely. Since the password change phase has little to do with our discussion, we omit it here and detailed information is referred to Ref. [18].

## III. Cryptanalysis of Li-Lee's Scheme

In this section we will show that Li-Lee's scheme is vulnerable to offline password guessing attack and denial of service attack. In addition, their scheme fails to preserve user anonymity and forward secrecy. Although tamper resistant smart card is widely assumed in most of the published authentication schemes, such an assumption is difficult in practice. Many researchers have shown that the secret information stored in a smartcard can be breached by analyzing the leaked information or by monitoring the power consumption [8-10]. Be aware of this threat, Li and Lee intentionally based their scheme on the assumption of non-tamper resistance of the smart card. However, Li-Lee's scheme fails to serve its purposes.

### A. Offline Password Guessing Attack

In Li-Lee's scheme, a user is allowed to choose his/her own password at will during the registration and password change phases. The user usually tends to select a password, e.g., his phone number, which is easily remembered for his/her convenience. Hence, these easy-to-remember passwords, called weak passwords [19], have low entropy and thus are potentially vulnerable to password guessing attack. Therefore, one of the most important security requirements for sound password-based authentication protocols is to resist against this threat. Li and Lee showed that Kim and Chung's scheme is vulnerable to offline password guessing attack once the adversary has obtained the secret information stored in the stolen smart card. However, we will show that Li-Lee's scheme still suffers from this threat as follows.

Let us consider the following scenarios. In case a legitimate user $U_i$'s smart card is stolen by an adversary $A$ just before $U_i$'s $j$th login, and the stored secret values such as $C_1$ and $RN_j$ can be revealed. Then, $A$ returns the smart card to $U_i$ and eavesdrops on the insecure channel. Because $U_i$'s identity is transmitted in plaintext within the login request, it is not difficult for $A$ to identify the login request message from $U_i$. Once the $j$th login request message $\{ID_i, C_4^j = h(ID_i \| x \| N) \oplus h(P_i \oplus RN_j), C_6^j\}$ is intercepted by $A$, an offline password guessing attack can be launched in the following steps:

*Step 1.* Guesses the value of $P_i$ to be $P_i^*$ from a uniformly distributed dictionary.

*Step 2.* Computes $T = h(h(P_i^* \oplus RN_j)) \oplus h(P_i^* \oplus RN_j)$ , as $RN_j$ is known.

*Step 3.* Computes $T' = C_1 \oplus C_4^j$ , as $C_1$ has been extracted and $C_4^j$ has been intercepted, where $C_1 = h(ID_i \| x \| N) \oplus h(h(P_i \oplus RN_j))$ , $C_4^j = h(ID_i \| x \| N) \oplus h(P_i \oplus RN_j)$ .

*Step 4.* Verifies the correctness of $P_i^*$ by checking if $T$ is equal to $T'$ .

*Step 5.* Repeats Steps1, 2, 3, and 4 of this phase until the correct value of $P_i$ is found.

After guessing the correct value of $P_i$, the adversary can compute $C_3^j = C_1 \oplus h(h(P_i \oplus RN_j))$ , $C_2^j = h(P_i \oplus RN_j)$ and $K_{U_i}^j = h(C_2^j \| C_3^j)$ . Then the adversary can obtain $RC_j$ by decrypting $C_6^j$ using $K_{U_i}^j$ , and gets $RS_j$ in a similar way. Hence the malicious user can successfully compute the session key $SK_j = h(RC_j \oplus RS_j)$ and renders the $j$th session between $U_i$ and $S$ completely insecure.

Moreover, once the $j$th login request message is intercepted, the adversary may block the communication channel between $U_i$ and $S$ completely until the session key $SK_j = h(RC_j \oplus RS_j)$ was obtained as stated above. Thereafter, he/she can fabricate and send a valid login request to the server $S$ and masquerade as a legitimate user $U_i$, or he/she can fabricate and send a valid password change request to update the entry corresponding to $U_i$ in the verification table on $S$. In either case, from then on $U_i$ will not be able to login to the server $S$. This leads to a strong denial of service attack.

### B. Denial of Service Attack

A denial of service attack is an offensive action whereby the adversary could use some methods to work upon the server so that the login requests issued by the legitimate

user will be denied by the server. In Li et at.' scheme, an adversary can easily launch a denial of service attack in the following steps:

*Step 1.* Eavesdrops over the channel, intercepts a login request $\{ID_i, C_4^j, C_6^j\}$ from $U_i$ and blocks it, supposing it is $U_i$'s *j*th login.

*Step 2.* Replaces $C_6^j$ with an equal-sized random number *R*, while $ID_i$ and $C_4^j$ are left unchanged.

*Step 3.* Sends $\{ID_i, C_4^j, R\}$ instead of $\{ID_i, C_4^j, C_6^j\}$ to the remote server *S*.

After receiving this modified message, *S* will perform Step V1 and V2 of the verification phase without observing any abnormality, as a result, the verifier corresponding to $ID_i$ in the verification table will be updated and the response $E_{K_S}(RC_j^*, RS_j, C_5^{j*})$ will be sent to $U_i$. On receiving the response from *S*, $U_i$ decrypts $E_{K_S}(RC_j^*, RS_j, C_5^{j*})$ and will find $(RC_j^*, C_5^{j*})$ unequal to $(RC, C_5)$, thus the session will be terminated. Thereafter, $U_i$'s succeeding login requests will be denied unless he/she re-registers to *S* again. That is, the adversary can easily lock the account of any legitimate user without using any cryptographic techniques. Thus, Li-Lee's protocol is vulnerable to denial of service attack.

### C. Failure to Achieve Forward Secrecy

Let us consider the following scenarios. Supposing the server *S*'s long time private key *x* is leaked out by accident or intentionally stolen by an adversary *A*. Once the value of *x* is obtained, with previously intercepted $C_4^j$, $C_6^j$ and $E_{K_S}(RC, RS, C_5)$ transmitted in the legitimate user $U_i$'s *j*th authentication process, *A* can compute the session key of *S* and $U_i$'s *j*th encrypted communication through the following method:

*Step 1.* Assumes $N = 0$.

*Step 2.* Computes $C_7^* = h(ID_i \| x \| N)$ and $C_8^* = C_7^* \oplus C_4^j$, where $ID_i$ is previously obtained by eavesdropping on the insecure channel.

*Step 3.* Computes $K_{U_i}^* = h(C_8^* \| C_7^*)$ and $K_S^* = h(C_7^* \| C_8^*)$.

*Step 4.* Decrypts $C_6^j$ to obtain $RC_j^*$ using $K_{U_i}^*$.

*Step 5.* Decrypts $E_{K_S}(RC, RS, C_5)$ to obtain $RC_j^{**}$ using $K_S^*$.

*Step 6.* Verifies the correctness of $N$ by checking if $RC_j^*$ is equal to $RC_j^{**}$. If they are unequal, sets $N = N + 1$ and goes back to Setp2.

*Step 7.* Decrypts $E_{K_S}(RC, RS, C_5)$ to obtain $RS_j$ using $K_S^*$.

*Step 8.* Computes $SK_j = h(RC_j \oplus RS_j)$.

Note that the value of *N* should not be very big, since the re-registration phase is not performed frequently in practice, and thus the above procedure can be completed in polynomial time. Therefore, Li-Lee's scheme fails to provide forward secrecy.

### D. Failure to Preserve User Anonymity

In many e-commerce applications, the violation of user anonymity may leak some personal secret information (e.g., secret online-order placement, transaction records, etc.) about the logging user to the adversary, and thus the provision of user anonymity is very important. What's more, the leakage of the user identity may cause an unauthorized entity to track the user's login history and current location [5]. Therefore, assuring anonymity does not only preserve user privacy but also make remote user authentication protocols more secure.

In Li-Lee's scheme, user's identity *ID* is static and in plaintext form in all the transaction sessions, an adversary can easily obtain the plaintext identity of this communicating client once the login messages were eavesdropped, and hence, different login request messages belonging to the same user can be traced out and may be interlinked to derive some secret information related to the user. Hence, user anonymity is not preserved.

## IV. Our Proposed Scheme

According to our analysis, three principles for designing a sound password-based remote user authentication scheme are presented. First, user anonymity, especially in some application scenarios, (e.g., e-commence), should be preserved, because from the identity $ID_i$, some personal secret information may be leaked about the user. Second, a nonce based mechanism is often a better choice than the timestamp based design to resist replay attacks, since clock synchronization is difficult and expensive in existing network environment, especially in wide area networks, and these schemes employing timestamp may still suffer from replay attacks as the transmission delay is unpredictable in real networks [20]. Finally, the password change process should be performed locally without the hassle of interaction with the remote authentication server for the sake of security, user friendliness and efficiency [3]. In this section, we present a new remote user authentication scheme to satisfy all the twelve criteria listed in section 1.

### A. Registration Phase

Let ($x$, $y = g^x \bmod n$) denote the server *S*'s private key and its corresponding public key, where *x* is kept secret by the server and *y* is stored inside each user's smart card. The registration phase involves the following operations:

*Step R1.* $U_i$ chooses his/her identity $ID_i$, password $P_i$ and a random number *b*.

*Step R2.* $U_i \Rightarrow S$: $\{ID_i, h(b \| P_i)\}$.

*Step R3.* On receiving the registration message from $U_i$, the server *S* computes $N_i = h(b \| P_i) \oplus h(x \| ID_i)$ and $A_i = h(ID_i \| h(b \| P_i))$.

*Step R4.* $S \Rightarrow U_i$: A smart card containing security parameters $\{N_i, A_i, n, g, y, h(\cdot)\}$.

*Step R5.* Upon receiving the smart card, $U_i$ enters *b* into his smart card.

### B. Login Phase

When $U_i$ wants to login the system, the following operations will be performed:

*Step L1.* $U_i$ inserts his/her smart card into the card reader and inputs $ID_i^*$ and $P_i^*$.

*Step L2.* The smart card computes $A_i^* = h(ID_i \| h(b \| P_i^*))$ and verifies the validity of $A_i^*$ by checking whether $A_i^*$ equals to the stored $A_i$. If the verification holds, it implies $ID_i^* = ID_i$ and $P_i^* = P_i$. Otherwise, the session is terminated.

*Step L3.* The smart card chose a random number $u$ and computes $C_1 = g^u \bmod n$ , $Y_1 = y^u \bmod n$ , $h(x \| ID_i) = N_i \oplus h(b \| P_i)$, $CID_i = ID_i \oplus h(C_1 \| Y_1)$ and $M_i = h(CID_i \| C_1 \| h(x \| ID_i))$.

*Step L4.* $U_i \rightarrow S$:$\{C_1, CID_i, M_i\}$.

### C. Verification Phase

After receiving the login request, the server $S$ performs the following operations:

*Step V1.* The server $S$ computes $Y_2 = (C_1)^x \bmod n$ using its private key $x$, and derives $ID_i = CID_i \oplus h(C_1 \| Y_2)$ and $M_i^* = h(CID_i \| C_1 \| h(x \| ID_i))$. $S$ compares $M_i^*$ with the received value of $M_i$. If they are not equal, the request is rejected. Otherwise, server $S$ generates a random number $v$ and computes the session key $SK = (C_1)^v \bmod n$ , $C_2 = g^v \bmod n$ and $C_3 = h(SK \| C_2 \| h(x \| ID_i))$.

*Step V2.* $S \rightarrow U_i$: $\{C_2, C_3\}$.

*Step V3.* On receiving the reply message from the server $S$, $U_i$ computes $SK = (C_2)^u \bmod n$ , $C_3^* = h(SK \| C_2 \| h(x \| ID_i))$, and compares $C_3^*$ with the received $C_3$. This equivalency authenticates the legitimacy of the server $S$, and $U_i$ goes on to compute $C_4 = h(C_3 \| h(x \| ID_i) \| SK)$.

*Step V4.* $U_i \rightarrow S$: $\{C_4\}$.

*Step V5.* Upon receiving $\{C_4\}$ from $U_i$, the server $S$ first computes $C_4^* = h(C_3 \| h(x \| ID_i) \| SK)$ and then checks if $C_4^*$ is equal to the received value of $C_4$ . If this verification holds, the server $S$ authenticates the user $U_i$ and the login request is accepted else the connection is terminated.

*Step V6.* The user $U_i$ and the server $S$ agree on the common session key $SK$ for securing future data communications.

### D. Password Change Phase

In this phase, we argue that the user's smart card must have the ability to detect the failure times. Once the number of login failure exceeds a predefined system value, the smart card must be locked immediately to prevent the exhaustive password guessing behavior. This phase involves the following steps.

*Step P1.* $U_i$ inserts his/her smart card into the card reader and inputs the identity $ID_i$ and the original password $P_i$.

*Step P2.* The smart card computes $A_i^* = h(ID_i \| h(b \| P_i))$ and verifies the validity of $A_i^*$ by checking whether $A_i^*$ equals to the stored $A_i$. If the verification holds, it implies the input $ID_i$ and $P_i$ are valid. Otherwise, the smart card rejects.

*Step P3.* The smart card asks the cardholder to resubmit a new password $P_i^{new}$ and computes $N_i^{new} = N_i \oplus h(b \| P_i) \oplus h(b \| P_i^{new})$, $A_i^{new} = h(ID_i \| h(b \| P_i^{new}))$. Thereafter, smart card updates the values of $N_i$ and $A_i$ stored in its memory with $N_i^{new}$ and $A_i^{new}$.

## V. Security Analysis

Although it is important to user formal methods to provide a formal security proof on any cryptographic protocols, the formal security proof of remote user authentication protocols with smart cards remains a challenging problem in cryptography research domain [21]. As far as we know, an efficient, simple, and convincing formal methodology for security analysis of protocols is still an important subject of research and an open issue. Few schemes [e.g., 6, 13] do provide formal security proof, unfortunately they are shortly found contradictory to their security claims because the formal methods employed all fail to capture some realistic attack scenarios [14]. Due to these reasons, most published user authentication schemes using smart cards [e.g., 1–5, 7, 15-18, 22-24] have been demonstrated with a simple proof. Therefore, we follow the approaches used in [5, 7] for comparison purpose. This opens a prominent future scope of this work to develop a simple and robust formal method for security analysis of user authentication protocols with smart cards.

The security of our proposed authentication scheme is based on the secure hash function and the discrete logarithm problem. In the following, we will analyze the security of the proposed scheme to verify whether the security requirements mentioned in Section 1 have been satisfied under the assumption that the secret information stored in the smart card can be revealed, i.e., the security parameters $N_i$, $A_i$ ,$b$, and $y$ can be obtained by a malicious privileged user.

*1) User anonymity:* Suppose that the attacker has intercepted $U_i$'s authentication messages $\{CID_i, M_i, C_1, C_2, C_3, C_4\}$. Then, the adversary may try to retrieve any static parameter from these messages, but these messages are all session-variant and indeed random strings due to the randomness of $u$ and/or $v$. Accordingly, without knowing the random number $u$, the adversary will face to solve the discrete logarithm problem to retrieve the correct value of $ID_i$ from $CID_i$, while $ID_i$ is the only static element corresponding to $U_i$ in the transmitted messages. Hence, the proposed scheme can preserve user anonymity.

*2) Offline password guessing attack:* Suppose that a malicious privileged user $U_i$ has got $U_k$'s smart card, and the secret information $b$, $N_k$, $A_k$ and $y$ can also be revealed under our assumption of the non-tamper resistant smart card. Even after gathering this information, the attacker has to at least guess both $ID_i$ and $P_i$ correctly at the same time, because it has been demonstrated that our scheme can provide identity protection. It is impossible to guess these two parameters correctly at the same time in polynomial time, and thus the proposed scheme can resist offline password guessing attack with smart card security breach.

*3) Stolen verifier attack and password disclosure to server:* In the proposed protocol, no sensitive verifiers corresponding to users are maintained by $S$. Therefore, the proposed protocol is free from stolen verifier attack. With

$h(b \| P_i)$ instead of plaintext password $P_i$ submitted to server $S$, it is computationally infeasible to derive $P_i$ from $h(b \| P_i)$ without knowing the random number $b$ due to the one-way property of the secure hash function.

*4) User impersonation attack:* As $CID_i$, $M_i$, $C_3$ and $C_4$ are all protected by secure one-way hash function, any modification to these parameters of the legitimate user $U_i$'s authentication messages will be detected by the server $S$ if the attacker cannot fabricate the valid $CID_i^*$, $M_i^*$, $C_3^*$and $C_4^*$. Because the attacker has no way of obtaining the values of $ID_i$, $P_i$ and $N_i$ corresponding to user $U_i$, he/she cannot fabricate the valid $CID_i^*$, $M_i^*$, $C_3^*$and $C_4^*$. Therefore, the proposed protocol is secure against user impersonation attack.

*5) Server masquerading attack:* In the proposed protocol, a malicious server *MS* cannot compute the correct $Y_2 = (C_1)^x \bmod n$ because he/she does not know the value of $S$'s private key $x$, and thus *MS* cannot derive the valid $ID_i = CID_i \oplus h(C_1 \| Y_2)$. Without knowing $U_i$'s valid $ID_i$ and $S$'s private key $x$, *MS* has to break the secure one-way hash function to retrieve $h(x \| ID_i)$. Furthermore, because *MS* cannot obtain $h(x \| ID_i)$, it is impossible to fabricate the proper $C_3 = h(SK \| C_2 \| h(x \| ID_i))$ to pass the verification of $U_i$ in Step V3 of the verification phase. Therefore, the proposed protocol is free from server masquerading attack.

*6) Replay attack and parallel session attack:* Our scheme can withstand replay attack because the authenticity of authentication messages $\{M_i, C_3, C_4\}$ is verified by checking the fresh random number $u$ and/or $v$. On the other hand, the presented scheme resists parallel session attack, in which an adversary may masquerade as legitimate user $U_i$ by replaying a previously intercepted authentication message. The attacker cannot compute valid $C_3$ because he does not know the values of $h(x \| ID_i)$ corresponding to user $U_i$. Therefore, the resistance to replay attack and parallel session attack can be guaranteed in our protocol.

*7) Mutual authentication:* In our dynamic ID-based scheme, the server authenticates the user by checking the validity of $C_4$ in the access request. We have shown that our scheme can preserve user anonymity, so user $ID_i$ is only known to the server $S$ and the user $U_i$ itself. We have proved that our scheme can resist user impersonation attack. Therefore, it is impossible for an adversary to forge messages to masquerade as $U_i$ in our scheme. To pass the authentication of server $S$, the smart card first needs $U_i$'s identity $ID_i$ and password $P_i$ to get through the verification in Step L2 of the login phase. In this Section, we have shown that our scheme can resist offline password guessing attack. Therefore, only the legal user $U_i$ who owns correct $ID_i$ and $P_i$ can pass the authentication of server $S$. On the other hand, the user $U_i$ authenticates server $S$ by explicitly checking whether the other party communicating with can compute the valid $C_3$ or not. Since the malicious server does not know the values of $ID_i$ corresponding to user $U_i$ and $x$ corresponding to server $S$, only the legitimate server can compute the correct $C_3 = h(SK \| C_2 \| h(x \| ID_i))$. From the above analysis, we conclude that our scheme can achieve mutual authentication.

*8) Denial of service attack:* Assume that an adversary has got a legitimate user $U_i$'s smart card. However, in our scheme, the smart card computes $A_i^* = h(ID_i \| h(b \| P_i))$ and compares it with the stored value of $A_i$ in its memory to checks the validity of user identity $ID_i$ and password $P_i$ before the password update procedure. It is not possible for the adversary to guess out $U_i$'s identity $ID_i$ and password $P_i$ correctly at the same time in polynomial time. Moreover, once the number of login failure exceeds a predefined system value, the smart card will be locked immediately. Therefore, our protocol is secure against denial of service attack.

*9) Forward secrecy:* Following our scheme, the client and the server can establish the same session key $S \equiv (C_1)^v \equiv (C_2)^u \equiv g^{uv} \bmod n$. Based on the difficulty of the computational Diffie-Hellman problem, any previously generated session keys cannot be revealed without knowledge of the ephemeral $u$ and $v$. As a result, our scheme provides the property of forward secrecy.

## VI. PERFORMANCE ANALYSIS

To evaluate our scheme, we compare the performance and the satisfaction of the criteria among relevant authentication schemes and our proposed scheme in this section. The reason why the schemes presented in [4,5, 24], instead of other works mentioned earlier in this paper, are selected to compare with is that, these three schemes are the few ones that can withstand offline password guessing attack under the non-tamper resistance assumption of the smart cards. The criteria of a secure and practical remote user authentication scheme are introduced in Section 1, and the comparison results are depicted in Table 2 and 3, respectively.

Since the login phase and verification phase are executed much more frequently than the other two phases, only the computation cost, communication overhead and storage cost during the login phase and verification phase are taken into consideration. Without loss of generality, the identity $ID_i$, password $P_i$, random numbers, timestamp values and output of secure one-way hash function are all recommended to be 128-bit long, while $n$, $y$ and $g$ are all 1024-bit long. Let $T_H$, $T_E$, $T_I$, $T_S$ and $T_X$ denote the time complexity for hash function, exponential operation, inverse operation, symmetric cryptographic operation and XOR operation respectively. Since the time complexity of XOR operation is negligible as compared to the other three operations, we do not take $T_X$ into account. Typically, time complexity associated with these operations can be roughly expressed as $T_E \approx T_I > T_S \geq T_H >> T_X$ [25-27].

In our scheme, the parameters $\{N_i, A_i, y_i, n, g\}$ are stored in the smart card, thus the storage cost is 3456 ($= 3 * 128 + 3 * 1024$) bits. The communication overhead includes the capacity of transmitting message involved in the authenti cation scheme, which is 2560 ($= 4 * 128 + 2 * 1024$) bits. During the login and verification phase, the

total computation cost of the user and server is $6T_E$+$12T_H$. As illustrated in Table 2, the proposed scheme is more efficient than Horng et al.'s scheme, enjoys nearly the same performance with Chen et al.'s scheme and Chung et al.'s scheme.

TABLE II.
PERFORMANCE COMPARISON AMONG RELEVANT AUTHENTICATION SCHEMES

| | **Our scheme** | **Li and Lee[18] (2011)** | **Chung et al.[22] (2009)** | **Chen et al.[4] (2010)** | **Horng et al.[5] (2010)** |
|---|---|---|---|---|---|
| Total computation cost | $6T_E$+$12T_H$ | $12T_H$ | $4T_E$+$2T_I$ +$12T_H$ | $6T_E$+$5T_H$ | $7T_E$+$4T_S$+$8T_H$ |
| Communication cost | 2560 bits | 856 bits | 2560 bits | 2560 bits | 2432 bits |
| Storage overhead | 3456 bits | 384 bits | 3200 bits | 3200 bits | 3328 bits |

TABLE III.
CRITERIA COMPARISON AMONG RELEVANT AUTHENTICATION SCHEMES

| | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** | **C7** | **C8** | **C9** | **C10** | **C11** | **C12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Our scheme | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Li and Lee[18] | No | Yes | Yes | No | No | Yes | No | Yes | No | Yes | No | No |
| Chung et al.[ 22] | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes | No | Yes | No | Yes |
| Chen et al.[4] | Yes | Yes | No | No | No | Yes | No | No | No | Yes | No | Yes |
| Horng et al.[5] | Yes | Yes | Yes | No | Yes | Yes | No | Yes | No | Yes | Yes | Yes |

As compared to Li-Lee's scheme, to withstand offline password guessing attack, public-key techniques are employed, which has been proved necessary by Halevi and Krawczyk in [28], and thus at least two exponentiations are required; to provide the feature of forward secrecy, the generation of the session key based on the Diffie-Hellman key exchange algorithm is common practice, and hence it needs another four exponentiations; to achieve user anonymity and other functionalities simultaneously, some additional costs are necessary. As a word, to conquer all the identified security flaws, the decrease of some performance is unavoidable and reasonable.

Table 3 gives a comparison of the admired features of our proposed scheme with the other relevant authentication schemes. Our proposed scheme provides forward secrecy (C12) and can change password locally (C6), while the schemes presented by Li and Lee fails to achieve these features; Our proposed scheme preserves user anonymity (C11), while the schemes presented by Li and Lee, Chung et al. and Chen et al. do not provide this property; our proposed scheme can resist various kinds of sophisticated attacks (C5), while Li-Lee's scheme is vulnerable to denial of service attack and Chen et al.'s scheme cannot withstand reflection attack. Our scheme and Chung et al.'s scheme is secure against smart card loss attack (C4) while all the other three schemes are prone to offline password guessing attack once the secret data stored in smart card is revealed, thereby these three schemes fail to achieve this security requirement. It is clear that our scheme meets more criteria as compared to other relevant authentication schemes using non-tamper resistant smart cards.

## VII. CONCLUSION

In this paper, we have demonstrated several attacks on Li-Lee's scheme. As to our main contribution, a robust authentication scheme is proposed to remedy these identified flaws, the security and performance analysis demonstrate that our presented scheme achieves all of the twelve independent requirements with high efficiency and thus our scheme is more secure and efficient for practical use. Remarkably, our scheme eliminates several hard security threats that are difficult to be solved at the same time in previous scholarship.

## ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61170241 and No. 61073042, and the open program of State Key Laboratory of Networking and Switching Technology under Grant No. SKLNST-2009-01-10.

## REFERENCES

[1] C.C. Chang and T.C. Wu, "Remote password authentication with smart cards," *IEE Proceedings-E*, vol. 138, no. 3, pp. 165–168, 1993.

[2] W.C. Ku and S.M. Chen, "Weaknesses and improvements of an efficient password based remote user authentication scheme using smart cards," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 1, pp. 204–207, 2004.

[3] I.E. Liao, C.C. Lee, and M.S. Hwang, "A password authentication scheme over insecure networks," *Journal of Computer and System Sciences*, vol. 72, no. 4, pp. 727–740, 2006.

[4] Y. Chen, J.S. Chou, and C.H. Huang, "Improvements on two password-based authentication protocols," Cryptology ePrint archive, Technical report 561, 2010.

[5] W.B. Horng, C.P. Lee, and J. Peng, "A secure remote authentication scheme preserving user anonymity with non-tamper resistant smart cards," *WSEAS Transactions on Information Science and Applications*, vol. 7, no. 5, pp. 619–628, 2010.

[6] J. Xu, W.T. Zhu, and D.G. Feng, "An improved smart card based password authentication scheme with provable security," *Computer Standards & Interfaces*, vol. 31, no. 4, pp. 723–728, 2009.

[7] S.K. Sood, "Secure Dynamic Identity-Based Authentication Scheme Using Smart Cards," *Information Security Journal: A Global Perspective*, vol. 20, no. 2, pp. 67–77, 2011.

[8] P. Kocher, J. Jaffe, and B. Jun, "Differential Power Analysis," in proceedings of CRYPTO'99, LNCS, vol. 1666, 1999, pp. 388–397.

[9] F.X. Standaert, T.G. Malkin, and M. Yung, "A unified framework for the analysis of side-channel key recovery attacks," in proceedings of Advances in Cryptology-EUROCRYPT 2009, LNCS, vol 5479, 2009, pp. 443–461.
[10] T.S. Messerges, E.A. Dabbish, and R.H. Sloan, "Examining Smart-Card Security under the Threat of Power Analysis Attacks," *IEEE Transactions on Computers*, vol. 51, no. 5, pp. 541–552, 2002.
[11] S.B. Wilson, D. Johnson, and A. Menezes, "Key agreement protocols and their security analysis," in proceedings of 6th IMA International Conference on Cryptography and Coding, Cirencester, LNCS, vol. 1355, 1997, pp.30–45.
[12] H. Krawczyk, "HMQV: A High-Performance Secure Diffie-Hellman Protocol," in proceedings of CRYPTO'05, LNCS, vol. 3621, 2005, pp. 546–566.
[13] R.C. Wang, W.S. Juang, and C.L. Lei, "Robust authentication and key agreement scheme preserving the privacy of secret key," *Computer Communications*, vol. 34, no. 3, pp. 274–280, 2011.
[14] S.H. Wu, Y.F. Zhu, and Q. Pu, "Robust smart-cards-based user authentication scheme with user anonymity," *Security and Communication Networks*, vol. 5, no. 2, pp. 236-248, 2012.
[15] C.G. Ma, D. Wang and Q. M Zhang, "Cryptanalysis and improvement of Sood et al.'s dynamic id-based authentication scheme," in proceedings of 8th International Conference on Distributed Computing and Internet Technology, Lecture Notes in Computer Science, Vol. 7154, 2012, pp. 141–152.
[16] E.J. Yoon, E.K. Ryu, K.Y. Yoo, "Further improvement of an efficient password based remote user authentication scheme using smart cards", IEEE Transactions on Consumer Electronics, vol. 50, no. 2, pp. 612–614, 2004.
[17] H. C. Hsiang and W. K. Shih, "Weaknesses and improvements of the Yoon-Ryu-Yoo remote user authentication scheme using smart cards," Computer Communications, vol. 32, no. 4, pp. 649–652, 2009.
[18] C.T. Li and C.C. Lee, "A Robust Remote User Authentication Scheme using Smart Card," Information Technology and Control, vol. 40, no. 3, pp. 231–238, 2011.
[19] D.V. Klein, "Foiling the Cracker: A Survey of, and Improvements to, Password Security," in proceedings of 2nd USENIX Security Workshop, 1990, pp. 5–14.
[20] L. Gong, "A security risk of depending on synchronized clocks," *ACM Operating System Review*, vol. 26, no. 1, pp. 49–53, 1992.
[21] D. He, M. Ma, Y. Zhang, and C. Chen, "A strong user authentication scheme with smart cards for wireless communications," *Computer Communications*, vol. 34, no. 3, pp. 367–374, 2011.
[22] C.G. Ma, D. Wang, P. Zhao and Y.H. Wang, "A new dynamic ID-based remote user authentication scheme with forward secrecy," in proceedings of the 14th Asia-Pacific Web Conference (APWeb Workshops 2012), Lecture Notes in Computer Science, Vol. 7234, pp. 199–211, Springer-Verlag, 2012.
[23] D. Wang and C.G. Ma, "Cryptanalysis and security enhancement of a remote user authentication scheme", The Journal of China Universities of Posts and Telecommunications, vol. 19, no. 5, pp.104-114, 2012
[24] H.R. Chung, W.C. Ku, and M.J. Tsaur, "Weaknesses and improvement of Wang et al.'s remote user password authentication scheme for resource-limited environments," *Computer Standards & Interfaces*, vol. 31, no. 4, pp. 863–868, 2009.
[25] W.B. Mao, *Modern Cryptography: Theory and Practice*, Prentice Hall PTR, New Jersey (2004)
[26] D.S. Wong, H.H. Fuentes, and A.H. Chan, "The Performance Measurement of Cryptographic Primitives on Palm Devices," in Proceedings of ACSAC'01, pp. 92–101, 2001.
[27] N.R. Potlapally, S. Ravi, A. Raghunathan, and N.K. Jha, "A study of the energy consumption characteristics of cryptographic algorithms and security protocols," *IEEE Transactions on Mobile Computing*, vol.5, no. 2, pp. 128–143, 2006.
[28] S. Halevi, H. Krawczyk, "Public-key cryptography and password protocols," ACM Transactions on Information and System Security, vol. 2, no. 3, pp. 230–268, 1999.
[29] D. Wang, C.G. Ma, and W. P., "Secure password-based remote user authentication scheme with non-tamper resistant smart cards," in proceedings of 26th Annual IFIP Conference on Data and Applications Security and Privacy (DBSec 2012), Lecture Notes in Computer Science, Vol. 7371, pp. 114–121, Springer Berlin /Heidelberg, 2012.

**Ding Wang** received his B.S. Degree in Information Security from Nankai University, China, in 2008. And then he went to Information Engineering University of PLA to work toward Information Security Engineering. Currently, he is under the supervision of Prof. Chunguang Ma. He has published more than 20 research papers in international journals and conferences. His research interests include cryptographic protocols and wireless network security.

**Chunguang Ma** is currently a Professor and Ph.D. candidate supervisor in the Department of Computer Science and Technology, Harbin Engineering University. He got his Ph.D. degree in cryptography from Beijing University of Posts and Telecommunications. His current research interests include cryptography, information security and wireless sensor networks.

**Qiming Zhang** got his Bachelor degree from Henan Polytechnic University in 2010. Currently, he is under the supervision of Prof. Chunguang Ma. His research interests include cryptography, wireless network security and trusted computing.

**Sendong Zhao** is currently one postgraduate student of Harbin Engineering University. His current research interests include cryptography, network security, software security, network massive data processing.

# Mutihop-enabled Trusted Handoff Algorithm in Heterogeneous Wireless Networks

Dan Feng
Computer School of Wuhan University, Wuhan, China
Email: Smart_titan@126.com

Huang Chuanhe, Wang Bo, Zhu Junyu, Xu Liya
Computer School of Wuhan University, Wuhan, China
Email: huangch@whu.edu.cn

***Abstract*—Ubiquitous and heterogeneous wireless network is an important form of network, and vertical handoff is one of the key issues of the mobility management in this type of network. In Ad hoc network, multihop-enabled forwarding is able to expand network coverage and reduce single-hop propagation distance. Based on the above, TMVHA (trusted and multihop-enabled Vertical Handoff Algorithms) is designed and multihop trust management mechanism is established, thus enhancing the handoff performance of the mobile nodes and the network performance after handoff. The experimental result shows that: the algorithm proposed in this paper outperforms those without considering multihop and node trust in handoff times and throughput, and at the same time being able to reduce the influence of the attack on network by malicious nodes.**

***Index Terms*—Heterogeneous wireless network, vertical handoff, multihop forwarding, trust**

## I. INTRODUCTION

Traditional cellular network is the typical single-hop wireless network, and only in the last hop the transmission between mobile terminal and base station is wireless connection, offering seamless connections to users by handoff management. In the past years, wireless Ad hoc network and multi-interface mobile devices have greatly developed, and smart phones, multihop cellular networks (MCNS), the integration of cellular network and Ad hoc mode have emerged, and this type of network stems from an idea: new cellular network should provide an all-IP platform, and on this platform, various types of wireless networks (such as WiFi, WiMAX and traditional cellular network, etc) can complete a variety of services by seamless interconnection. 3GPP has begun to be engaged in the research of the tendency, and the norms of the interconnection between WCDMA and WLAN have already been standardized [1].

Meanwhile, WLAN based on 802.11 has become a widely used technology, with many advantages, such as radio spectrum without registration, low-cost devices, relatively inexpensive flow resources, and end-to-end IP connectivity. However, the biggest shortcoming of WLAN is the small coverage area of the network unit. The adopting of Ad hoc technology can expand the coverage area of WLAN and enhance the network performance. As is shown in Fig.1, mobile node A can directly communicate with the access point AP1 and achieve the maximum throughput; node B can directly communicate with AP1 but only with the help of node A can it achieve the maximum throughput; while without the relay of node B, node C cannot connect with AP1.

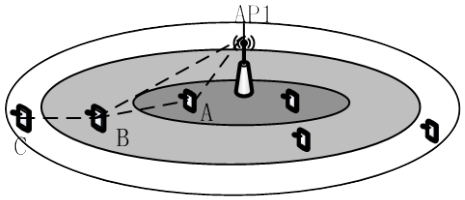



Fig.1 Application scenario

Ad hoc network is a multihop temporary autonomous system possessing time-varying topology, which is the collection of mobile nodes connected by wireless links. It can be applied in the construction of military battlefield information system, the situation of civil emergency and law enforcement, commercial and residential environment, thus provoking widespread attention on the issue of its safety. The features of dynamic topology, the limitation of resource and wireless communication make the safety assurance of routing protocol more complicated than that of traditional wired network. Therefore, how to prevent various attacks and improve network safety is an active research area about Ad hoc network when designing the handoff algorithms.

In this paper, a handoff algorithm is proposed by introducing trust and Ad hoc mode into network handoff management, combined with the current research on trust model in Ad hoc network. This algorithm can improve the handoff performance by Ad hoc mode and enhance the safety of the algorithm by trust similarity model, and also it carries out the relative theoretical analysis and simulation. The result shows that, the algorithm is able to alleviate the demand for resource reservation, retain the high spectral efficiency while reduce the access blocking probability. It could also prevent malicious nodes from joining the forwarding table, prevent malicious links from joining the establishment of links in the process of finding forwarding nodes, and improve a lot in throughput in the aspect of end-to-end delay.


doi: 10.4304/jnw.8.1.156-164

## II. RELATED WORK

### A. Handoff Algorithm

The process of vertical handoff in heterogeneous wireless network can be divided into network discovery, handoff judgment and handoff implement. Vertical handoff algorithm is widely used in cellular networks like GSM, and the handoff is mainly triggered by setting threshold of one or more concrete parameter. In [2], heterogeneous wireless network of Ad hoc, cellular network and WLAN is proved to outperform other networks when adopting the minimum transmission time and transmission cost strategy. A mobile model proposed in [3] is to predict the next target access point of the mobile node, and the handoff strategy based on the above can efficiently reduce the handoff time. An analysis model proposed in [4] is used to plan the service area of mobile gateway, thus enhancing the network performance. In [5] the authors divide the users by a kind of neural network to predict the location where users often go, and design a vertical handoff algorithm to reduce the average handoff times and the average handoff dropping probability. In [6], on the basis of PMIPv6, a vertical handoff algorithm based on reply/mobility ratio is designed. In [7] a mobile prediction model is established, predicting the movement tendency of the users by the movement historical records, improving handoff performance as well as the prediction accuracy.

### B. Trust Management Mechanism

The current research on Ad hoc trust model mainly stems from e-business and peer-to-peer computing. There are plenty of related researches that apply trust into Ad hoc network. In [8] the authors use entropy to quantify the uncertainty of trust, and by combining the characteristics the trust model is built. As the model is of no universal significance, and at the same time the probability of the false recommendation information is not eliminated, the calculated trust value is incorrect. In [9] the half-ring path and half-ring model based on half-ring theory is proposed and finally the trust evaluation model is established, however, it is not suitable to be promoted in large-scale network. In [10], according to probabilistic methods of Bayesian network, the authors build model for trust based on First-hand and Second-hand information. A trust model based on subjective logic is given in [11], considering that the measure for trust has subjective logic. The measure of trust value is described and defined in [12] combined with fuzzy mathematics, and trust model is established.

### C. Multihop-enabled cellular networks

The assumption integrating multihop into cellular network system began at 1990s. The thought that integrates relay system into cellular network system is proposed in [13], and in this paper, the proposed multi-hop cellular networks and the multihop mode between mobile terminal and the base station can reduce transmission consumption and provide more users with connection service. In [14] the authors propose to arrange some relay nodes in some area beforehand in case the signal is crowded. These relay nodes are called ARSs (Ad hoc relaying stations), which forwards the communication in the signal crowded units of the cellular network to the uncrowded units. In [15] the authors have conducted in-depth research on the idea that the integration of Ad hoc network and cellular network can improve the performance of the system, and they divide a unit in the cellular network into two concentric circles. When in inner circle users communicate with the base station directly, while when mobile users staying in the farther circular regions Ad hoc mode is adopted.

## III. VERTICAL HANDOFF ALGORITHM

### A. Network Model

Take WiFi as high-bandwidth, low coverage while UMTS as low-bandwidth, high coverage Wireless Access Technology as examples to introduce the handoff algorithms proposed in the paper. The algorithm also adapts to other vertical handoff between WLAN (Wireless Local Area Networks, WLAN) and WWAN (Wireless Wide Area Networks, WWAN).

$A=\{a_1,a_2,...,a_N,a_{N+1},...,a_{N+M}\}$ is used to represent the access points in wireless network, where $\{a_1,a_2,...,a_N\}$ is the access point AP of WLAN, and $\{a_{N+1},...,a_{N+M}\}$ is the base station BS. M is far smaller than N, because normally several APs will be equipped in the coverage area of one base station. $U=\{u_1,u_2,...,u_K\}$ is used to represent all mobile users in the network.

TABLE 1
DEFINED VARIABLE TABLE

| | |
|---|---|
| $a_i$ | The access point in the wireless network |
| $u_1$ | All the mobile users in the network |
| $T_{a_i}$ | Historical average value of handoff time of all mobile nodes handoff to $a_i$. |
| HoTime | The time spending for mobile node from $u_j$ sending handoff request to completing handoff and establishing new links |
| $V_{a_i}$ | The set of all the mobile nodes that connect to $a_i$ |
| $b_{ij}$ | The bandwidth demand for mobile node $u_j$ connecting to $a_i$. |
| $B_{a_i}$ | The biggest bandwidth that access point $a_i$ can provide |
| $L_{a_i}$ | The load of the access point $a_i$ |
| $RSS_{ij}$ | received signal strength that mobile node $u_j$ received from access node $a_i$ |
| $T_{cur}^{d}(i,j)$ | The direct trust level between node i and node j |
| $T^{ind}(i,j)$ | The indirect trust level between node i and node j |
| $N_j$ | All the neighbor nodes of $u_j$ |
| $TN_j$ | The trust neighbor table of node $u_j$ |
| $TR_j$ | The set of nodes that can become relay nodes in the trust neighbor node of mobile node $u_j$ |
| $T_{threshold}$ | Threshold of trust |

### B. The Statement of the Algorithm

The idea of the handoff algorithm is as follows: firstly, dividing the type of nodes (high-speed mobile nodes and low-speed mobile nodes) by their moving speed, when it is high-speed mobile node, only choosing base station as the candidate access point, and when the received signal

strength of current access point is lower than the preset threshold, handoff is triggered; when it is low-speed mobile node, handoff is triggered on the following two conditions: (a)when finding the signal strength of new WLAN access point is greater than the current signal strength of the access point, (b)when the signal strength of current access point is lower than the pre-set threshold, handoff is triggered, and low-speed mobile nodes are allowed to access by multihop; next, establish the decision attribute set by collecting the network information(including the load of access point, the received signal strength, handoff time, the remaining bandwidth), and when the mobile nodes access by multihop, malicious nodes can be prevented from joining the network by calculating node credibility; finally, make handoff decision by TOPSIS algorithm based on fuzzy logic. Unlike the existing algorithm based on fuzzy logic or multi-attribute decision theory, this paper proposes to improve the coverage area and the performance of WLAN in hotspot of wireless network by multihop, and at the same time to introduce trust management mechanism based on trust similarity to improve the safety when multihop. The flow chart is shown in fig. 2, and the specific algorithm and the mathematical analysis are to be described in section D.

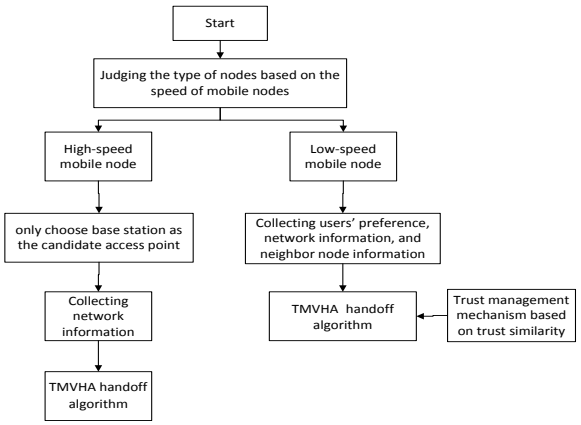



Fig.2 vertical handoff flow chart based on TMVHA

*C. The Calculation of Handoff Decision Attributes*

Definition 1: average handoff time. Average handoff time $T_{a_i}$ refers to the historical average value of handoff delay of mobile nodes handoff to $a_i$ (i.e. HisHoTime).

Each AP or BS maintains a candidate access point table, which contains the latest received handoff time (HoTime) and historical handoff time (HisHoTime).

After each mobile node (MN) succeeds in handoff, it will send HoTime to new access point, and new access point sends HoTime to the original access point by wired network.

The access point which has received HoTime updates HisHoTime based on formula (1).

$$HisHoTime = \alpha \times HisHoTime + \beta \times HoTime \quad (1)$$

Where $\alpha + \beta = 1$, and compromise $\alpha$ and $\beta$ based on the reality. If the current network situation focuses more on current network environment, set $0 < \alpha < \beta < 1$.

Definition 2: the load of access point. The load of access point refers to the sum of all the required bandwidth of mobile nodes connected to the access point AP (or base station BS).

$$L_{a_i} = \sum_{u_j \in V_{a_i}} b_{ij} \qquad a_i \in A, 1 \le i \le N+M \quad (2)$$

Where $V_{a_i}$ is the set of all the mobile nodes connected to $a_i$, $b_{ij}$ is the required bandwidth when mobile node $u_j$ connects $a_i$, and $L_{a_i} \le B_{a_i}$. $B_{a_i}$ is the maximum bandwidth that access point can provide.

*D. Trusted and Multihop-enabled Vertical Handoff Algorithms(TMVHA)*

*1.The choose of access point when single-hop*

When mobile node $u_j$ triggers handoff, it requires the current access node to handoff, and the current access point obtains load $L_{a_i}$ of all the access node in the candidate access point table through wired network, where $a_i \in C \subset A$ and C is the set of all the access point in the candidate access point table. Then through wireless links, send candidate access point tables, their average handoff time and load information to mobile node $u_j$. $RSS_{ij}$ is the received signal strength between mobile node $u_j$ and access point $a_i$, and the value of $RSS_{ij}$ can be easily obtained by monitoring.

In order to manage the issue of the multi-attribute decision, we transform the multi-attribute decision of the four attributes (handoff time, the load of access point, received signal strength, remaining bandwidth) of m candidate access points into the geometry system composed by m nodes in four-dimensional space. At the same time, all schemes can be regarded as the points in the space of the system. We adopt TOPSIS to measure the relative degree of closeness with the ideal solution, and judge the quality of the scheme by the distance between ideal solution and the negative ideal solution.

Suppose the judgment matrix is X,

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ \vdots & & & \\ x_{m1} & x_{m2} & x_{m3} & x_{m4} \end{bmatrix} \quad (3)$$

where $x_{i1} = T_{a_i}$, $x_{i2} = \frac{L_{a_i}}{B_{a_i}}$, $x_{i3} = RSS_{ij}$, $x_{i4} = B_{a_i} - L_{a_i}$.

Step 1: As unit, order of magnitude and dimension of the four attributes of candidate access points are different, all the judgment indexes need normalization.

$$y_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^{\ 2}}} \qquad \text{i=1,2…m} \quad \text{j=1,2,3,4} \quad (4)$$

Standard matrix $Y = (y_{ij})_{m\times 4}$ can be expressed as

$$Y = (y_{ij})_{m\times 4} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} \\ \vdots & & & \\ y_{m1} & y_{m2} & y_{m3} & y_{m4} \end{bmatrix} \quad (5)$$

Step 2: Weighted standardization decision matrix is built based on users' preference to the four attributes (handoff time, the load of access point, received signal strength, remaining bandwidth). The weight vector is

$$w=(w_1,w_2,w_3,w_4),\ \sum_{j=1}^{4}w_j=1 \tag{6}$$

In reality, the definition of users' preference for network properties possesses ambiguity and vagueness, and the introduction of fuzzy set theory can implement the conversion between linguistic variables and fuzzy numbers through membership function. However, traditional fuzzy multi-attribute decision-making has the problem of complete calculation when using fuzzy number in fuzzy logic operation, which brings adverse impact on handoff performance and has higher demands on the computing power of mobile nodes.

Chen and Hwang proposed a multi-attribute decision-making method which is able to effectively solve the problem [16]. In the method proposed by Chen and Hwang, we synthesize and revise the research of several scholars, propose eight semantic scales, and represent semantic items by triangular and trapezoidal fuzzy number, which is suitable for two to eleven semantic items representatively. In this paper, we adopt five linguistic variables to characterize the user's preferences: very low, low, medium, high, very high. According to the formula $\mu_T=\left[\frac{\mu_R(M)+1-\mu_L(M)}{2}\right]$ (where $\mu_R(M)$ and $\mu_L(M)$ are the boundary values around fuzzy number M), the fuzzy number is converted into the corresponding exact value:" 0.091, 0.283, 0.5, 0.717, 0.909." Bring them to formula (6), and normalize them.

Weighted standardized decision-making matrix V is the product of each column in matrix Y and its corresponding weights, therefore, weighted standardized decision-making matrix V is:

$$V=\begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ \vdots & & & \\ v_{m1} & v_{m2} & v_{m3} & v_{m4} \end{bmatrix}=\begin{bmatrix} w_1y_{11} & w_2y_{12} & w_3y_{13} & w_4y_{14} \\ \vdots & & & \\ w_1y_{m1} & w_2y_{m2} & w_3y_{m3} & w_4y_{m4} \end{bmatrix} \tag{7}$$

Step 3: determine ideal solution and negative ideal solution, set $X^+$ and $X^-$ as the ideal solution and negative ideal solution respectively:

$$X^+=\{(\max_i v_{ij}\,|\,j\in J),(\min_i v_{ij}\,|\,j\in J')\}=\{v_1^+,v_2^+,v_3^+,v_4^+\} \tag{8}$$

$$X^-=\{(\min_i v_{ij}\,|\,j\in J),(\max_i v_{ij}\,|\,j\in J')\}=\{v_1^-,v_2^-,v_3^-,v_4^-\} \tag{9}$$

Step 4: calculating the distance. The distance between alternative solutions and the ideal solution is measured by four-dimensional Euclidean distance:

$$D_i^+=\sqrt{\sum_{j=1}^{4}(v_{ij}-v_j^+)^2},i\in M \tag{10}$$

Similarly, the distance to negative ideal solution is:

$$D_i^-=\sqrt{\sum_{j=1}^{4}(v_{ij}-v_j^-)^2},i\in M \tag{11}$$

Step 5: calculating the sequence relative nearness degree with the ideal solution:

$$C_i^+=\frac{1}{\frac{D_i^+}{D_i^-}+1},0<C_i^+<1,i\in M \tag{12}$$

When $C_i^+$ and $C_i$ approaches 1, $X_i$ approaches $X^+$.

Then the target function of the vertical handoff algorithm without considering multihop is:

$$\underset{\forall a_i}{Max}(C_i^+)\qquad a_i\in C\subset A \tag{13}$$

The information of the candidate access point, which is needed by mobile nodes in handoff process, is provided by current connected base station. The above handoff strategy can effectively reduce the communication of the mobile nodes in handoff process, thus reducing the handoff time and lessening the interference.

*2. The choose of access point and relay node when multihop*

(1)The calculation of neighbor nodes' credibility

Based on the research foundation of the current trust model [17, 18, 19, 20], we ensure the accuracy and generality of the description of trust, thus meeting the soft security requirements of network. However in Ad hoc network, as there is uncertainty in the concept of trust between nodes, we can estimate and judge the credibility of nodes only by observing the forwarding behaviors of each other in periodic time. Also as the subjectivity, asymmetry and timeliness of the trust, watchdog mechanism [21] is adopted to monitor the forwarding behaviors of nodes and the concept of credibility is defined based on actual forwarding situation, so as to reflect the authenticity and accuracy of the node's current forwarding behavior.

Combined with the characteristic of Ad hoc network, for the sake of formalized modeling and analysis, we define the network as follows:

Weighted directed graph $G=(V,E)$ is used to indicate Ad hoc network, where V is the nonempty node set of networks, E is the communication link set of connected node pairs (neighbor nodes are within the communication ranges of each other), $|V|$ and $|E|$ represent the numbers of nodes and links in the network, respectively.

Definition 3: direct credibility degree. Direct credibility degree refers to the direct trust level that is calculated by the experience history of direct interaction (the status of sending and receiving data packets) between node i and its neighbor node j in a periodic time.

$T_{cur}^{d}(i,j)$ is used to express direct trust level. Combined with the timeliness of node i's direct trust level, as well as the corresponding incentives and penalties based on the impacts on the network security of its own forwarding behavior, we have the following formula:

$$T_{cur}^{d}(i,j)=\begin{cases}1-TF\times(RF\times S-PF\times F)\times T_{last}^{d}(i,j)+TF\times(RF\times S-PF\times F) & (s>0 \text{ or } f>0 \quad T_{last}^{d}(i,j)>0)\\ 1-TF\times T_{last}^{d}(i,j) & (s=0,\ f=0,\ T_{last}^{d}(i,j)>0\ )\\ 0 & \text{others}\end{cases} \tag{14}$$

Where $\Delta t$ is the time interval between present and the last interaction, and TF is the time factor of the direct credibility degree in $\Delta t$, i.e. $TF=\Delta t/(\Delta t+1)$. $T_{cur}^{d}(i,j)$ is the direct credibility degree of the two interaction nodes last time, and $T_{cur}^{d}(i,j)$ is the direct credibility degree of the current time. RF and PF are reward factor and penalty factor respectively$(1\geq RF>PF\geq 0, RF+PF=1)$, and we set parameters experientially. ST and FT represent the probability of successful and failed forwarding of node i respectively in $\Delta t$,i.e. S=s/(s+1) , F=f/(f+1) , and s and f represent the actual times of successful and failed forwarding from node i to node j. Obviously,$0\leq T_{cur}^{d}(i,j)\leq 1$.

Malicious nodes intentionally deceive goodwill nodes to enhance their own credibility degree and frame the case that goodwill nodes have hostile attack, in order to avoid false declaration and malicious forwarding behavior to cover up the malicious nodes themselves. In this way, we consider adopting the recommended mechanism of third-party node to calculate the credibility degree indirectly.

Definition 4: Similarity. Similarity describes the similar level of the judgment ability (credibility degree) of node i and node k. The high similarity between node i and node k indicates that they have the same view to the other node w, i.e. these two nodes have the same recommendation level $S(i,k)$. The specific formula is as follow:

$$S(i,k)=\frac{\sum_{w\in CN(i,k)}(T^{d}(i,w)-\overline{T}i)\times(T^{d}(k,w)-\overline{T}k)}{\sqrt{\sum_{w\in CN(i,k)}(T^{d}(i,w)-\overline{T}i)^{2}}\times\sqrt{\sum_{w\in CN(i,k)}(T^{d}(k,w)-\overline{T}k)^{2}}} \tag{15}$$

Obviously, $0\leq S(i,k)\leq 1$. Where $CN(i,k)$ is the mutual neighbor nodes of node i and node k, in addition, $T^{d}(k,w)$ is the direct credibility degree between node k and w, and $T^{d}(i,w)$ is the direct credibility degree between node i and w, respectively. Combined with the interaction status between node k, i and $CN(i,k)$, we can work out the average direct credibility degree of node k and node I, i.e. $\overline{T}_k$ and $\overline{T}_i$ respectively.

By formula (15), the similarity between node i and its every neighbor node can be calculated. Through sequencing these similarities, node i obtains the neighbor node set (the number of nodes is m) whose similarity to it achieves a certain threshold $\tau$ ( $\tau\geq 0.6$ ). Therefore, calculating the trust level between node i and node j can be based on recommendation trust level between i and m most similar nodes, and thus the indirect credibility degree can be obtained.

Defination5: Indirect credibility degree. Indirect credibility degree is the direct credibility degree recommended by the neighbor node which has high similarity with node i. Synthesizing the direct credibility degree of each neighbor node with similarity can reflect the trust level of recommendation more reliably, authentically and actually.

The calculation formula of indirect credibility degree $T^{ind}(i,j)$ is defined as:

$$T^{ind}(i,j)=\frac{\sum_{k\in m}T^{d}(k,j)\times s(i,k)}{\sum_{k\in m}s(i,k)} \tag{16}$$

Obviously, $0\leq T^{ind}(i,j)\leq 1$.

Defination6: Credibility degree. Credibility degree is the sum of direct credibility degree and indirect credibility degree between nodes.

Total credibility degree between node i and node j can be calculated based on formula (14) and formula (16). We have:

$$T(i,j)=\alpha\times T^{d}(i,j)+\beta\times T^{ind}(i,j) \tag{17}$$

Where $0\leq T(i,j)\leq 1$. $\alpha+\beta=1$, we select the compromising value of $\alpha$ and $\beta$. If the current network condition prefers the estimate of the direct credibility degree, set $1>\alpha>\beta>0$.

By preventing the access of nodes with low credibility degree to the network and the participation in the process of forwarding packets, network safety can be enhanced, and the implementation efficiency and robustness of network's normal data packets can be improved, in addition, the disruption of malicious nodes' attacks on the network can be reduced.

(2) The broadcast of heartbeat packets

All mobile nodes in WLAN can periodically broadcast heartbeat packets. The heartbeat packets include: (a)ID of mobile nodes $u_j$; (b) ID of the current access point of mobile nodes $a_i$; (c)the maximum bandwidth $B_{ij}$ of links between mobile node and access point $L_{a_i u_j}$; (d)load of the current access point $a_i$ and the provided maximum bandwidth $B_{a_i}$

For the sake of better calculation, we suppose mobile nodes can estimate the possible maximum bandwidth communicating with AP by candidate relay node $\rho_j$. Mobile nodes can obtain candidate relay node set $R(j)$ by receiving the heartbeat packets broadcast by its neighbor nodes.

(3)The choose of relay node

Definition **7:** trusted neighbor node list. Trusted neighbor node list is the set of all the trusted nodes in the neighbor nodes of node i.

$N_j$ represents all the neighbor nodes of , and $TN_j$ represents the trusted neighbor node list of node $u_j$($TN_j\subset N_j$), where any node $u_k$($u_k\in TN_j$), i.e.$T(u_k)\geq T_{threshold}$ is ensured to be the trusted node. $T_{threshold}$ is the trusted threshold of network nodes.

Definition 8: Trusted neighbor relay node list. Trusted neighbor relay node list is the set $TR_j$ of nodes able to

become relay nodes in the trusted neighbor node of mobile node $u_k$. Therefore,

$$TR_j = TN_j \cap R_j \qquad 0 < j < K \tag{18}$$

Similar to the algorithm calculating the vertical handoff without considering multihop, with TOPSIS algorithm based on fuzzy logic, judgment matrix X is established, and the attribute set of each candidate node contains four attributes (handoff time, the load of access point, signal strength of the access point, bandwidth preference). For $X_i = \{x_{i1}, x_{i2}, x_{i3}, x_{i4}\}$, where $x_{i1} = T_{a_i}$, $x_{i2} = L_{a_i} / B_{a_i}$, $x_{i3} = RSS_{jk}$, $x_{i4} = \min(\rho_i, (B_{a_i} - L_{a_i}))$.

$RSS_{jk}$ is the received signal strength between mobile node $u_j$ and neighbor node $u_j$.

Then the target function of relay node chosen by mobile node $u_j$ when multihop is:

$$\underset{u_k \in TR_j}{Max}(C_i^+) \tag{19}$$

After choosing the optimal relay nodes, target access point is the AP which the relay node connected to.

*3.TMVHA pseudocode*

Before designing TMVHA , Filtering Neighbor MNodes(G, i) algorithm (see Algorithm1) is used to filter out the malicious nodes contained in neighbor node of node i, forming trusted neighbor node list, and get rid of the links between malicious node and node i. Mobile node choose whether to find relay node to access by multihop based on the algorithm proposed in this paper. CLOSED-DEGREE function refers to the closed degree of the target solution and the optimal solution (according to formula (13) and (19)). Based on the algorithm proposed in this paper, mobile nodes choose the optimal access point (relay node) to access, and the algorithm pseudocode is shown in Algorithm2.

**Algorithm1**: Filtering NeighborMNodes (*G*,*i*)

```
1 /* ensure all the neighbor nodes of i are trusted nodes */
2 for each edge (i,j) in E
3   if T_j < T_threshold then
4     E ← E − edge(i, j)      V ← V − {j}
5   end if
6 end for
```

**Algorithm2** Generating the Recommended Access Point

```
1:  //L_CAP: the Candidate Access Points list
2:  //CAP_i: the ith access point in L_CAP
3:  //N: the number of access points in L_CAP
4:  //RAP: Recommended Access Point
5:  Copy L_CAP to L_RAP
6:  m =1;
7:  RAP= CAP_0;
8:  CD=CLOSED-DEGREE(CAP_0)
9:  while m< N-1 do
10: j ← CLOSED-DEGREE (CAP_m)
11:   if j> CD then
12:     RAP= CAP_m
13:     CD=j
14:   end if
15:   m = m +1;
16: end while
```

(4) Analysis on the algorithm effectiveness

By analyzing the moving speed of the nodes, TMVHA adaptively adjust handoff trigger condition, and prevent high-speed mobile nodes from accessing WLAN whose coverage area is relatively small, thus reducing the handoff times of the high-speed mobile nodes and signaling overhead due to handoff. For low-speed mobile node, network performance can be enhanced by multihop, trust model can be established, the nodes with low credibility degree are prevented accessing the network and the participation in the process of opportunity forwarding data packets are refused, thus the network security is enhanced and the execution efficiency and robustness of the normal data packets of network is improved, in addition, the influence of the malicious nodes' attack on network is reduced. Compared with classic algorithms such as SAW, adopting TOPSIS algorithm based on fuzzy logic when multi-attribute decision can improve the scientificalness, accuracy and operability of multi-objective decision analysis, and at the same time, can better work out the problem of uncertainty and ambiguity in the definition of network attribute preference.

**Theorem 1: The time complexity of the algorithm is $O(N^2)$**

Prove: the time complexity of the algorithm mainly depends on the execution efficiency of while loop from step 9 to step 16.

The time complexity of while loop is O(N). Time cost in step 10 mainly concentrates on the execution of the function CLOSED-DEGREE (CAPm), while the time complexity of the closed degree of target solution and the optimal solution in function CLOSED-DEGREE is related to the established Judgment matrix, being O(2*m*n), and in the algorithm of this paper, m=N, n=4, then the time cost of step 10 is O(8N), i.e. O(N). For the whole algorithm, the time complexity is $O(N^2)$, QED.

## IV. SIMULATION AND ANALYSIS

*A.The Setting of Simulation Environment*

To evaluate the effectiveness of the TMVHA , the expansion packet nsclick and Madwifi is applied to the real simulation tests. The simulation packet nsclick is the expansion packet that combines the framework of ns2 [29] and click modular router [27] together, in order to be used to transplant when designing routing protocol in the real wireless network.

The simulation environment is based on DCF using IEEE 802.11b [30] in layer MAC, and the mobile model adopts random way-point model. The assumption of the attack types of malicious nodes is: malicious nodes selectively lose received data packets randomly in the probit range of [0.4-0.8], prominently showing its selfishness. The setting of other parameters is shown in table 2:

TABLE 2
THE SETTING OF SIMULATION PARAMETERS

| parameters | meanings | values |
|---|---|---|
| Area | topology area | 1000m*1000m |
| K | the number of network nodes | 20-200 |

| | | |
|---|---|---|
| N | the number of access points of WLAN | 15 |
| $R_{WLAN}$ | the coverage radius of WLAN | 50m |
| M | the number of base stations of WiMAX | 3 |
| $R_{WiMAX}$ | the coverage radius of WiMAX | 1000m |
| R | signal transmission radius of mobile nodes | 30 |
| V | the maximum moving speed of high-speed mobile nodes | 15m/s |
| V' | the maximum moving speed of low-speed mobile nodes | 2m/s |
| $\alpha$ | weight coefficient of $T^d(i,j)$ | 0.6 |
| $\beta$ | weight coefficient of $T^r(i,j)$ | 0.4 |
| $\Delta t$ | the interval of the updating of credibility degree | 0.5s |
| T | the whole simulation time | 200s |
| M | the number of malicious nodes in network | 1-60 |
| $T_{threshold}$ | threshold of credibility degree in network | 0.5 |

*B. Analysis on the Simulation Results*

*1. The comparison of performance parameters*

In order to test and verify the performance of TMVHA and compare with the ordinary multiple attribute decision algorithm (T-VHA) based on TOPSIS by experiment, in this paper we focus on analyzing the following parameters:

Handoff times: the sum of handoff times of all the mobile nodes.

Average handoff dropping probability: the failure probability for mobile node handoff operation.

Average handoff time: the average needed time for the establishment of new wireless link of all the nodes from handoff trigger to handoff completion, which is an important indicator reflecting handoff algorithm performance.

Average throughput: the average number of bytes sent by all mobile nodes per second.

*2.Performance comparison*

Fig.3 to Fig.6 reflects the performance comparison between T-VHA and TMVHA regarding the above parameters under different conditions.

Fig.3 shows the comparison of the changes of total handoff times when the total number of mobile nodes increases: overall, the total handoff times of both these two algorithms and the average handoff times of nodes are increasing, while the total handoff times of T-VHA algorithm (from 25 to 343 times ) is higher than TMVHA (from 21 to 230 times); the reason is that as the increase of the number of nodes, based on vertical handoff strategy proposed in this paper ,unnecessary handoff can be avoided if considering various handoff decision attribute comprehensively and finding suitable access point, besides, distinguishing types of nodes and preventing high-speed mobile node from accessing WLAN.

Fig.4 shows the comparison of the average handoff dropping probability of mobile nodes: the average handoff dropping probability of T-VHA algorithm increases to 0.265, while that of TMVHA increases to 0.147. The causes are as follows: like Fig.3, TMVHA is able to find more suitable access point and distinguish node type to prevent high-speed mobile node from accessing WLAN.

Fig.5 presents the comparison of average handoff time of mobile nodes: overall, the average handoff time of these two algorithms are increasing, however, when the number of mobile nodes is 40, the average handoff time of TMVHA is less than that of T-VHA algorithm; when the number of mobile nodes increases from 40 to 120, average handoff time of TMVHA is less than that of T-VHA (TMVHA from 16ms to 41ms, T-VHA from 26ms to 43ms); when the number of mobile nodes exceeds 120, the average handoff time of TMVHA is more than that of T-VHA, while as the number of nodes increases, the increase speed of both algorithms tend to be the same(TMVHA from 41ms to 130ms, T-VHA from 43ms to 112ms).The reasons are as follows: when the number of mobile nodes is small, network topology is relatively sparse, and it is not easy for mobile nodes to find relay nodes. TMVHA proposed in this paper can better choose freer and less delayed access point through multi-attribute decision, while as the number of mobile nodes increases, mobile nodes can access the network by multihop to acquire better network bandwidth, and handoff time is bound to be increased when mobile nodes calculate trust and choose relay nodes.

Fig.6 shows the comparison of average throughput of mobile nodes on condition that there does not exist malicious nodes: as the increase of mobile nodes, the average throughput of these two algorithms are decreasing, while the average throughput of TMVHA (from 210K/sec down to 120KB/sec) is higher than that of T-VHA (from 205KB/sec down to 63KB /sec). The cause is: like Fig.5, TMVHA can choose freer access points, and access to WLAN by multihop to improve the average throughput of nodes.

In order to verify the influence of trust management mechanism proposed in this paper on handoff and network performance, we remove trust management mechanism from T TMVHA, i.e. only consider network performance (MVHDA) when choosing relay nodes, and then compare with TMVHA proposed in this paper. Assume the number of mobile nodes is 200, the simulation result is shown in Fig.7: as the increase of malicious nodes, the throughput of these two methods are decreasing , while the throughput of TMVHA (from 121KB/s down to 78KB/s) is slightly higher than that of MVHDA(from 123KB/s down to 56KB/s). The reason is: through the monitor and judgment of trust management mechanism, malicious nodes and edges in network are removed, and the influence of malicious nodes on network throughput is decreased.

Overall, compared with the performance of T-VHA, that of TMVHA increases by 32.9% on handoff times (the number of mobile nodes is 200), on average handoff dropping probability it increases by 44.5% (the number of mobile nodes is 200), on average handoff time it decreases by 25.3% (the number of mobile nodes is 172 reaching the peak), on average throughput it increases by 47.5% (the number of mobile nodes is 200). The performance of TMVHA increases by 30.7% on average

throughput compared with that of the algorithm which does not consider node credibility (the number of malicious nodes is 43 reaching the peak).

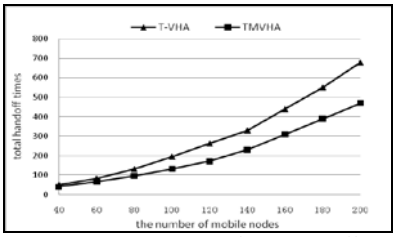


Fig.3 the comparison of total handoff times of mobile nodes

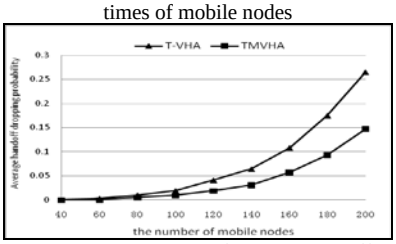


Fig.4 the comparison of average handoff dropping of mobile nodes

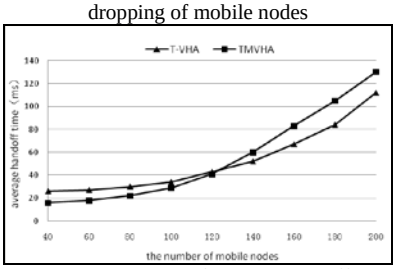


Fig.5 the comparison of average handoff time of mobile nodes

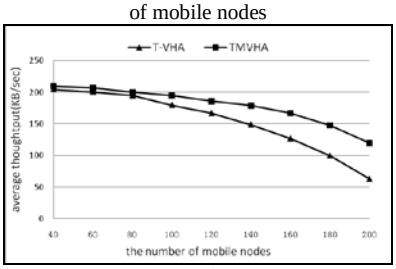


Fig.6 the comparison of average throughput of mobile nodes

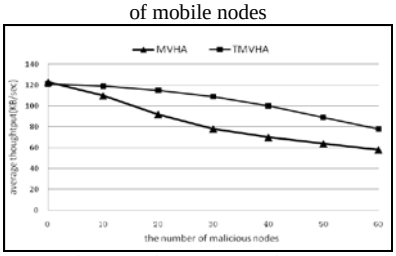


Fig.7 the influence of the change of malicious nodes on average throughput

## V. CONCLUSION

On the basis of the study on the handoff mechanism of heterogeneous network, we focus on how to improve handoff performance and network performance. Combined with the characteristic of Ad hoc network, a vertical handoff algorithm of multi-attribute decision which permits mobile nodes to find relay nodes to access (or forward part of the data packets) is designed, the concept of trust is introduced, the trust management model is proposed, handoff performance of mobile nodes and network performance after handoff is improved, and finally through the implementation and test by nsclick simulation package, the simulation result proves the effectiveness of TMVHA and the significant improvement in reducing the influence on network performance by preventing the malicious nodes' attack. In the following work, we will conduct more in-depth performance test on this algorithm, and further study on the handoff management mechanism under high-speed environment.

## VI. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China(60970117), National Natural Science Foundation of China(61173137), Hubei Province Natural Science Fund Key Issue(2010CDA004), The Central University Basic Scientific Research Business Expenses Special Funds Projects(3104002).

## REFERENCES

[1] Luo Jun-zhou, Wu Wen-jia, Yang Ming, "Mobile Internet: Terminal Devices, Networs and Services", Chinese Journal of Computer, Vol 34, NO.11, pp.2029-2051, NOV.2011.

[2] Kaveh Shafiee, Alireza Attar, Victor C, M. Leung, "Optimal Distributed Vertical Handoff Strategies in Vehicular Heterogeneous Networks", IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 29, NO. 3, pp.534-544, MARCH.2011.

[3] Zhenxia Zhang, Richard W, Pazzi, Azzedine Boukerche, Bjorn Landfeldt, "Reducing Handoff Latency for WiMAX Networks using Mobility Patterns", 2010 IEEE Wireless Communications and Networking Conference (WCNC), pp.18-21, April 2010 .

[4] Yen-Cheng Lai,, Phone Lin, and Shin-Ming Cheng, "Performance Modeling for Application-Level Integration of Heterogeneous Wireless Networks", IEEE Transactions onVehicular Technology, VOL.58 , NO.5, pp.2426 – 2434, Jun.2009,

[5] Salahshuori. Y, Azemi, G, "A pattern recognition based handoff algorithm for micro-cellular systems", 2011 19th Iranian Conference on Electrical Engineering, May.2011 .

[6] Jeon.S, Younghan Kim, "Adaptive Handoff Management in the Proxy Mobile IPv6 Domain", 2011 IEEE 73rd Vehicular Technology Conference, May.2011.

[7] Meetei K.P, George, "A Handoff management in wireless networks using predictive modeling", 2011 National Conference on Communications, JUN.2011.

[8] Y. Sun, W. Yu, Z. Han, and K. J. R. Liu, "Information Theoretic Framework of Trust Modeling and Evaluation for Ad Hoc Networks" , IEEE Journal on Selected Areas in Communications, Vol.24, pp. 305-317, 2006.

[9] G. Theodorakopoulos and J. S. Baras, "On Trust Models and Trust Evaluation Metrics for Ad Hoc Networks", IEEE Journal on Selected Areas in Communications, Vol. 24, Issue 2, pp. 318-328, Feb.2006.

[10] J. Li, R. Li, J. Kato, "Future Trust Management Framework for Mobile Ad Hoc Networks", IEEE Communications Magazine, Vol. 46,Issue 4, pp.108-114, April 2008.

[11] PIRZADA A A, MCDONALD C, "Establishing trust in pure ad hoc networks", Proceedings of the 27th Australasian Conference on Computer Science", pp.47-54, 2004.

[12] J. Luo, et al., "Fuzzy Trust Recommendation Based on Collaborative Filtering for Mobile Ad-hoc Networks",

Proc. of the 33rd IEEE Conference on Local Computer Networks (LCN 2008), pp.305-311, Oct. 2008.
[13] Luo, H., Ramjee, R., Sinha, P., Li, L., & Lu, S., "UCAN: Aunified cellular and ad-hoc network architecture", Mobicom'03, Sep.2003.
[14] Wu, H., Qiao, C., De, S., & Tonguz, O., "Integrated cellular and ad hoc relaying systems: iCAR", IEEE Journal on Selected Areas in Communications (JSAC), pp.2105–2115.
[15] Law, L., Krishnamurthy, S., & Faloutsos, M., "Capacity of hybrid cellular-ad hoc data networks", Infocom'08, Apr.2008.
[16] Shu-Jen Chen Ching-Lai Hwang and Frank P. Hwang, "Fuzzy Multiple Attribute Decision Making: Methods and Applications", Springer-Verlag 1992.
[17] G. Theodorakopoulos and J. S. Baras, "On Trust Models and Trust Evaluation Metrics for Ad Hoc Networks", IEEE Journal on Selected Areas in Communications, Vol.24, Issue 2, pp.318-328, Feb.2006.
[18] J. Li, R. Li, J. Kato, "Future Trust Management Framework for Mobile Ad Hoc Networks", IEEE Communications Magazine, Vol. 46,Issue 4, pp.108-114, April.2008.
[19] PIRZADA A A, MCDONALD C. "Establishing trust in pure ad hoc networks". Proceedings of the 27th Australasian Conference on Computer Science, pp.47-54, 2004.
[20] J. Luo, et al., "Fuzzy Trust Recommendation Based on Collaborative Filtering for Mobile Ad-hoc Networks, Proc", The 33rd IEEE Conference on Local Computer Networks (LCN 2008), pp.305-311, Oct. 2008.
[21] Marti S,et al, "Mitigating routing misbehavior in mobile ad hoc networks", Proc of MobiCom'00. New York : ACM , pp. 255-265, 2000 .

**Dan Feng** was born in 1981. He is a Ph.D. degree candidate in school of computer of Wuhan University, China. His research interests include, Heterogeneous Wireless Networks, mobile Ad hoc networks, and computer networks.

**Huang Chuanhe** was born in 1963. He is currently a professor and Ph.D. supervisor of school of computer of Wuhan University, China. He is also a senior member of China Computer Federation. His research interests are in the areas of wireless network, computer networks.

**Wang Bo** was born in 1982. He is a Ph.D. degree candidate in school of computer of Wuhan University, China. His research interests include mobile Ad hoc networks, Wireless Mesh Networks, computer networks

# LCCWS: Lightweight Copyfree Cross-layer Web Server

Haipeng Qu
Department of Computer Science, Ocean University of China, Qingdao, China
Email: haipeng.qu@gmail.com

Lili Wen, Yanfei Xu and Ning Wang
Department of Computer Science, Ocean University of China, Qingdao, China
Email: wll0920@126.com, kloseangell@gmail.com,wnmanman@hotmail.com

***Abstract*—For the purpose of improving the performance of web server, this paper implements a high-performance web server prototype system, which is named LCCWS. Adopting PF_RING technology, which is similar to zero-copy technology, this system achieved to copy data between network interface device and kernel ring buffer in DMA mode and access data between application program and kernel ring buffer in MMAP way, so that the CPU participation and memory copies are reduced, saving much CPU overhead. When data packets splitting and encapsulating, using the lightweight TCP/IP protocol suite, the improved web server passed up the packets directly from the data-link layer to application layer, so that the time of copies is reduced and the packet processing is accelerated. LCCWS reduces the CPU overhead effectively, decreases the transferred data copying between memories, and improves transferred efficiency, laying foundation for further research to improve strong practical, feature-rich and high-performance web server.**

***Index Terms*—PF_RING, zero-copy, cross-layer design, Web Server**

## I. INTRODUCTION

With the development of WWW (World Wide Web), the scope and number of the dissemination of information on the Internet, including the vast majority of web services and information, is growing exponentially [1] As the key node of network information processing and promulgating, Web Server (Web server, also known as WWW server) needs to carry more traffic load than ever before. This requires the web server to have a higher packet processing rate and lower transfer delay. However, there are performance bottlenecks in CPU, memory, network bandwidth, storage, and applications and so on for traditional web server. There are two main factors affect the performance: firstly, the traditional working method of network cards relies on capturing packets from the NIC(Network Interface Card), and copying them to the upper layers, which both packet capturing and copying consumes much resources of CPU, resulting in a drop in overall performance; secondly, the packet content copied multiple times when being split and encapsulated for the traditional protocol suite, which is not only complex and cumbersome, but also takes significantly much time and occupies large amount of memory space.

The current methods of solving server performance bottlenecks [2] are to add network bandwidth, expand the memory capacity, use the SCSI (Small Computer System Interface) or RAID (Redundant Array of Inexpensive Disk) array storage disk, use multi-processor, increase the cache, optimize disk I/O, etc. Nevertheless, those methods above are not the fundamental solutions to the web server performance bottlenecks but cost high.

Speeding up the capture rate and reducing the transfer data copy between the memories, can effectively improve the efficiency of communication, thereby improving the performance of web server. Copyfree idea is also refer to zero-copy. Zero-copy technology [3] reduces the operating system and the protocol overhead of data transfer and increases bit rate of communication, then makes high-speed communications, by reducing or eliminating the operations of critical communication routines that affect the communication rate. However, there is no existing uniform standard for the zero-copy technology, and modifications to the network card driver are required [4], which affects the normal use of the most of web servers, and it has poor portability and generality. In this paper, CPU processing overhead is reduced and data packet capture performance is improved, based on zero-copy technology by loading PF_RING module in the Linux kernel, capturing packets using the NAPI based on device polling, storing packets in the ring buffer through the DMA (Direct Memory Access) mode; upper layers access the ring cache directly in MMAP (memory-mapped) way, reduces the time of copying data in kernel space, and saves the system time and space resources [5]; PF_RING is an independent modules without having to modify the network card driver, therefore it improves the generality and portability. For packets splitting and encapsulating, the traditional system protocol stack is improved in our paper, and a simplified protocol stack, which is only with the necessary functions such as packing, unpacking, checking of the packet header, fixed address communications, reduces copies between each protocol layer, passes up packets directly from the data-link layer to the application layer to speed up the packet


doi: 10.4304/jnw.8.1.165-173

analyzing. Based on the above two techniques, a lightweight high-performance web server prototype system LCCWS is derived from a small open source Web server Mattows [6].

## II. Related Work

This paper presents several technologies to solve the web server performance bottlenecks: PF_RING technology based on zero-copy [7], improved TCP/IP protocol suite, the mechanism of static web cache and multi-queue priority control mechanism.

### A. PF_RING Technology based on Zero-copy

Zero-copy [8] is a fast network packet processing mechanism [9], without any memory copy in the packet separation and encapsulation process, and only when packet is sent out by the NIC driver or read by applications will the data be copied. Zero-copy technology uses DMA and memory mapping.

PF_RING [10] technology is using the Device Polling, also known as NAPI in Linux, for NIC interrupt mode, then the captured packets will be copied to the circular queue in the kernel buffer through DMA. Interface of reading data packets is implemented by MMAP. If the DMA transfer interrupts during Device polling, the packets reach the DMA buffer through DMA [11] channels, and then the network card interrupt mode can be shut down. All the subsequent data should be received in polling mode. Once the kernel polls the device, network cards will generate an interrupt. If no packet is waiting for being received or some of packets have been received while polling, network card interrupt mode will be open.

PF_RING provides a modified version of network access interface, and the improved network access interface uses the PF_RING protocol suite instead of the PF_PACKET protocol suite to create a socket and receive packets to be built on top of the PF_RING network access interface. Working process is shown in Fig. 1 [5].

### B. Improved TCP/IP Protocol Suite

TCP/IP is a protocol suite [12], which is usually divided into four layers, corresponding to the OSI reference model of the transport layer, internet layer, data-link layer and physics layer. Each layer implements a set of protocols. Its core is the transport layer protocol (TCP and UDP) internet layer protocol (IP) and physical interface layer, and the protocol implementations of these three layers usually are completed in the operating system kernel. The data-link layer is the basis of the TCP/IP. The IP layer provides transport service to the transport layer (TCP and UDP). The IP layer's main functions are addressing, routing, segmentation and reload. TCP provides functions of dividing/merging, joining/splitting, multiplexing/demultiplexing, confluence/diversion, and retransmission, etc.

Traditional TCP/IP protocols mainly focus much attention on ensuring the reliability of data transmission and dataflow control, however its real-time performance is bad, and implementation is complex. It takes up a lot of system resources [13]. The improved TCP/IP protocol suite, on the premise that it is not contrary to the protocol standards, reduces the gap between service requirements of high-level (session layer) and services of the low-level network protocol stack can provide, and its features only include such as the header encapsulation and splitting, the header checking, and message acknowledgement of the data-link layer, transport layer and internet layer.

During data packets is splitting and encapsulating, the improved web server passes up the data directly from data-link layer to application layer, reducing the times of copy of the data packets between the other layer protocols. It is the pointer of data that passed between the layers of protocol suite, and only when the data is sent out by the driver or read by the application program will the data really be copied. The code size and essential memory space are smaller than the general-purpose TCP/IP protocol suite to speed up packets processing.

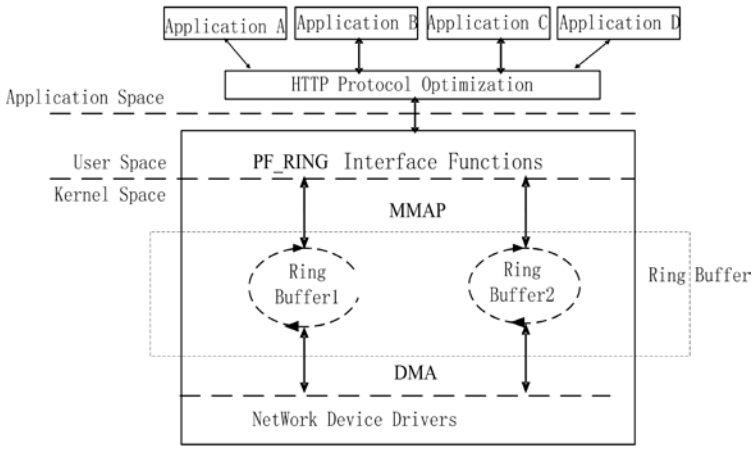



Figure 1. PF_RING Operating Principle.

There are a variety of TCP/IP gateways programming interfaces, and currently the most popular one is the socket programming interface. Socket is a kind of socket specification built in the transport layer protocol [12]. In our paper, using the network access interface that PF_RING provides, we improve TCP / IP protocol stack, then encapsulate the improved TCP / IP protocol stack, and rewrite Web server socket communication functions.

TABLE I.
SOCKET STRUCTURE BODY DESCRIPTION

| Socket Structure Body Field | Contents |
|---|---|
| addrs | source MAC, destination MAC ,source IP ,destination IP, source port and destination port information |
| sendc | serial number of data bytes sent |
| recvc | serial number of data bytes expected |
| windows | Send window size |
| windowr | Receive window size |
| packet[SOCKET_MTU+14] | Storage of receiving / sending packets |
| TTL | Time to live |
| MSS | Maximum Segment Size |
| family | Protocol family |
| type | Type identification of SOCK_DGRAM, SOCK_RAW, etc. |
| protocol | Protocol type: TCP,UDP etc |
| errbuf[PCAP_ERRBUF_SIZE] | the error messages |
| status | Socket status identification including listening, connected, default, etc. |
| *device | NIC device descriptor |
| *fp | Capturing instance descriptor to open socket |
| filter | Filter structure |

### C. The Mechanism of Static Web Cache

The mechanism of static web cache has been proved to be extremely effective and it is one of the widely used techniques for web browsing acceleration. It reduces the waste of network bandwidth by reducing duplicated transmission on the network of the requested data. And also it decreases computing of application server to lighten the load of web server and application server by duplicated computing to dynamic data, ultimately to achieve purpose of shortening waiting time of user, improving processing efficiency of current system.

Frequently accessed web pages are stored in cache servers in the form of static pages whose index is stored in web server. When web server receives the request from client, it will first check whether indexes contain static page the request needs. If yes, the static page is fetched from cache server and is sent to the client. Otherwise, a new static page will be generated, and it will be added to the cache server and the web server cache with its indexes according to cache replacement algorithms. The web cache system operating principle is shown in Fig. 2

### D. Multi-queue Priority Control Mechanism

Even though the server implementation is fast enough, when faced up with the large number of network requests, the processing capability of the server may still be in inefficient state, which is due to the large number of the requests exceeding the maximum number of the requests the server can deal with. In order to ensure the service capabilities of the server under the extremely terrible environment, we propose the multi-queue priority control mechanism.

We separate the requests into three queues: black queue, white queue and gray queue, the priority order is as follows: white>gray>black. All new requests from the client enter gray queue by default. If a certain request's frequency exceeds a specified threshold for some time, then the request enters the white queue and its request is processed preferentially. When the request is identified as dangerous request or untrusted request by server intrusion detection mechanism, then the request enters the black queue.

When the server handles the requests of the client, requests in the white queue will be processed by the server preferentially. Not until the white queue is empty, the server processes requests in the gray queue. When the white queue and gray queue are both empty, the server turns to process the requests in the black queue.

## III. LCCWS SYSTEM DESIGN

The lightweight web servers usually have features such as simple functions and high speed. Generally it is open source, and easy for analyzing and improving. This paper develops a high-performance web server prototype - LCCWS, which is exactly improved on the base of an open source web server Mattows.

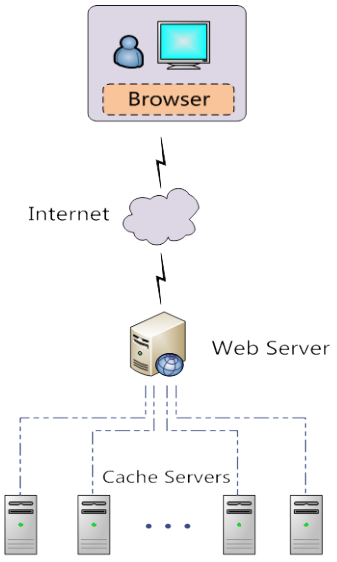



Figure 2. Web cache system operating principle

### A. The Overall Design of LCCWS

LCCWS is a lightweight Web server running on the Linux, supporting CGI (Common Gateway Interface). At the kernel of the operating system, PF_RING with cache

loaded as a module combines with NAPI to reduce the interrupt response frequency of NIC. In user space, applications access the network packets in the ring buffer by MMAP. The splitting and encapsulation of the packets uses a simplified TCP/IP protocol suite. It is data pointers that passed between the layers of protocol stack, and thus the copy times between each layer are reduced. The design of the system is shown in Fig. 3. In order to be faster, we also adopt the mechanism of static web cache and multi-queue priority control mechanism, its working process is shown in Fig. 5.

*B. LCCWS System Implementation*

The implementation of LCCWS mainly includes four modules: loading PF_RING module into the Linux kernel implementing simplified TCP/IP protocol suite, Web caching mechanism and multi-queue priority control mechanism.

**PF_RING module load:** PF_RING is running in the kernel space as a module. It can be dynamically loaded and unloaded, which becomes a part of kernel after being loaded. Loading process is divided into three steps: the kernel upgrade, configuration loading and test [10]. First, we can check the version of the kernel through the "uname" command; download the source code and kernel patch according to version, and upgrade the kernel by the command: "zcat linux-2.6.25-1-686-smp-PF_RING. patch.gz | patch -p0"; then, we compile the kernel configuration, load PF_RING module and configure it to support devices polling mechanism; finally, we test the operation results of loading process by “cat info” command. If it is successfully loaded, PF_RING information as Fig. 4 will be shown.

**Lightweight TCP/IP protocol suite:** LCCWS improves the TCP/IP protocol suite. At the same time, it wraps the improved TCP/IP protocol suite using the network access interface provided by PF_RING. It also rewrite socket communication functions: fss_socket (), fss_bind(), fss_listen(), fss_accept(), fss_connect(), fss_send(), fss_recv(), fss_close() etc. The implementation in detail is as follows:

```
root@Linux-desktop:/proc/net/pf_ring# cat info
PF_RING Version       :  3.9.7
Ring slots            :  4096
Slot version          :  10
Capture TX            :  Yes [RX+TX]
IP Defragment         :  NO
Transparent mode      :  Yes
Total rings           :  0
Total plugins         :  0
```

Figure 3. PF_RING information

In data-link layer, data frames are built and split. The IP protocol and TCP protocol are simplified in the internet layer and in the transport layer separately. Each layer is described as follows: the data-link layer realizes the functions like building and splitting the link layer frame; IP layer protocol does not realize the function of routing, but tests the IP address and check sums of the packets received, what’s more, it adds IP headers for packets which are waiting to be sent off etc. The TCP protocol is to ensure that the data transmission is correct and orderly. It uses the technologies such as checksums, ACK and sequence number. In the design process of protocol stack, we shall try to reduce the gap between service requirements from upper-layer protocol stack (application layer) and the service that lower-level network can provide, such as when we package packets, the maximum of TCP layer packets, the maximum of IP packets and the maximum of the link layer frames are set to be the same value, which avoids the trouble when TCP layer’s packets divide/merge and IP layer’s data recombine /split [14]. Simplified protocol suite structure is shown below in Fig. 6.

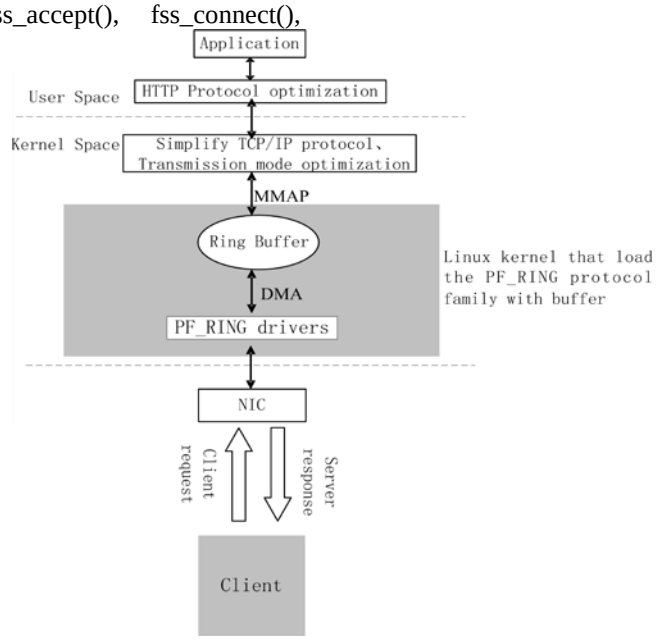



Figure 4. LCCWS system Design

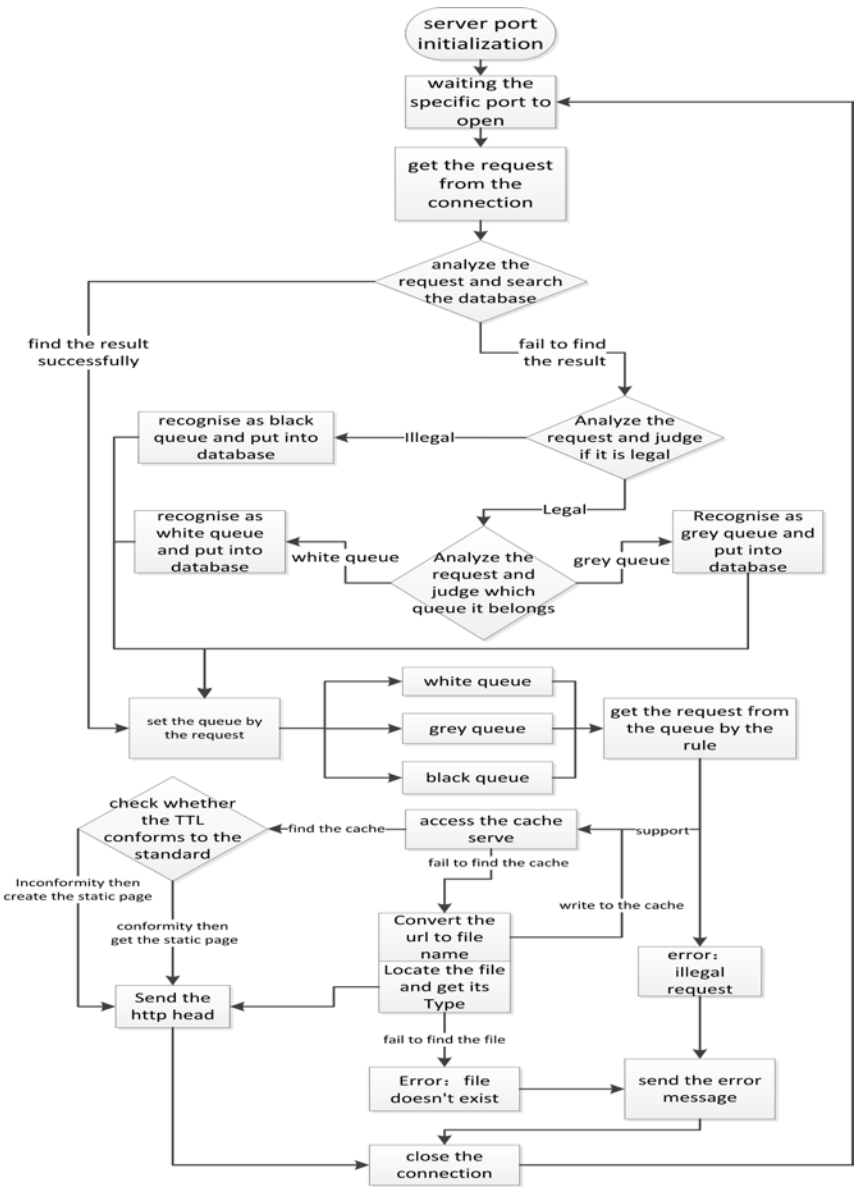



Figure 5. Simplified TCP/IP stack

Using the network access interface provided by PF_RING, we rewrite web socket communication functions and redefine socket structure through packaging the improved TCP/IP protocol suite. Socket structure includes addrs (storing communication information: source MAC, destination MAC, source IP, destination IP, source port and destination port), sendc(storing the serial number of the sent bytes), recvc (storing the serial number of the byte that wish to received), windows (size of send window), windowr (size of receive window), packet[SOCKET_MTU+14](storing packets received and sent), TTL(time to live), MSS(maximum segment size), family(protocol family), type(marking SOCK_DGRAM, SOCK_RAW type etc), protocol (protocol type TCP, UDP), errbuf[PCAP_ERRBUF_SIZE](Storing error messages), status(marking the socket's current status: listening, connected, default, etc.), *device(the network card descriptor), *fp(the capture instance descriptor of opening this socket), filter(filter structure), etc [14].

**Web caching mechanism and cache replacement algorithms** Static content caching improves the capacity and response speed of web server handling client requests. Caching mechanism is concretely realized as follows:

Static content index is cached in web server. Every index is a data block and every data block contains address information of the requested content, the attribute of access time and access frequency. The cache server is

used to store content of static page pointed by data block which is indexed by web server.

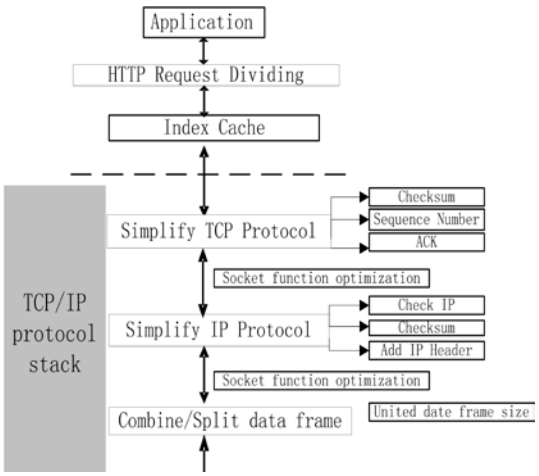



Figure 6. Simplified TCP/IP stack

Web server pre-checks user requirement every time. When this requirement hits index buffer, static page in cache server pointed by data block is immediately returned to user. But when the request misses, according to cache replacement algorithms, the resource this request pointing to will generate static content, then it is added to cache server and the index pointed to the static content is added to web server cache.

This cache replacement algorithm is achieved as follows:

When the buffer is full, demarcate indexes in web server according to the attribute of access frequency. That is, partition requests whose access frequency is greater than *I* to *domain High* and those access frequency lower than *I* to *domain Low*. The value of *I* is that the ratio of *domain High* and *domain Low* is 2:1. That is, the number of index blocks in *domain High* accounted for 2/3 of its total number.

Sort indexes in *Low* by the attribute of access time. Swap the location of least recently accessed index block and current block. And then swap synchronously static content in cache server.

In order to guarantee cache data to be in accordance with real request cache, the algorithm adopted is like this. Each data block object (content) in web cache is assigned a valid time TTL (Time To Live). When TTL is over, the object (content) contains this TTL is invalid. Next request to this object (content) will cause web server to flush the cache.

**Implementation of multi-queue priority control mechanism** Multi- queue priority control mechanism can guarantee that, when faced up with a large number of instantly user requests, In order to ensure the capability of the service, the web server can be able to make the appropriate response. Encapsulate each request, and set the last request time, the number of the request, credible or not, reference count and other attributes.

The implementation of its priority queue is as follows:

Initialize three request queues, as the black queue, gray queue, white queue respectively, its processing priority order: white>gray>black.

For each user request, firstly the server checks whether it exists in the current queues or not. If it exists, then the reference count of the corresponding request object is incremented by 1.

If the current request is not in the respective queue, detect whether the request is secured (trusted).

If it is a trusted request, check historical request of the request frequency for a period of time. If the number of the request is greater than *K* times specified by the system, the request will be added to white queue; otherwise, it will still in the gray queue.

If the request is untrusted, the request will join the black queue.

When the server handles the requests, requests in the white queue will be processed by the server preferentially. Not until the white queue is empty, does the server process requests in the gray queue. When the white queue, and gray queue are both empty, the server turns to process requests in the black queue. The multi-queue system work flow is as shown in Fig. 7.

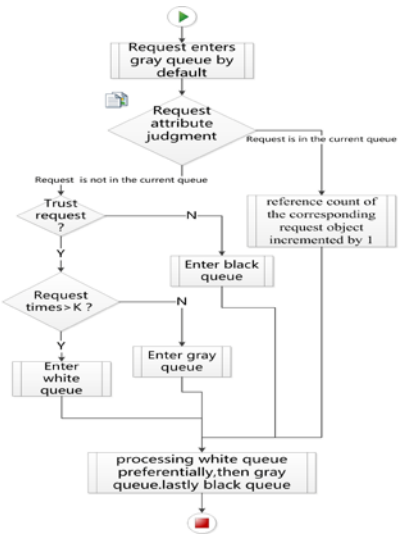



Figure 7. Multi-Queue System Work Flow

## IV. Performance Analysis

User data should to be split and encapsulated many times during being transmitted from local host to the remote host. Data copies are passed for data transmission between every protocol layer. This process increases the system cost, and reduces the system performance. LCCWS performance improvement mainly includes the following respects：

- Using the NAPI mechanism based on devices polling improves the speed of capturing data packets [5]. In big network traffic environment, the standard network card interrupts combining with layer-by-layer data copies and system calls will take up vast of CPU resources, while resources actually dealing with the data are very little. For big network traffic environment, once a DMA transmission interrupt is found, the first thing is to close NIC interrupt mode. All the subsequent data should be received in polling mode. Once the kernel polls the device, the network card generates an interruption, which would greatly reduce the interruption times of the network card while polling. If no packet is waiting to be received or some of packets have been received, then the NIC interrupt mode can be open. This means accelerate capturing and processing of packets greatly.
- Data copying in DMA mode does not require CPU to participate in and reduces the CPU overhead.
- User layer adopts the MMAP way to access the cache of the network interface directly, avoids copies between the memories, shortens the path that packets need to move and saves the CPU overhead.
- When data packets splitting and encapsulating, the improved web server passes up the packets directly from the data-link layer to application layer, reducing the times of copies of the data packets of the middle layer protocols. It is data pointer that passed between the layers of protocol suite, and only when the data is sent out by the driver or read by the application program will the data really move.
- Under the multi-queue priority control mechanism, the requests are put into white, gray, black three priority, web server will decide priority to the request of higher priority, so that the lower priority requests will occupy less server resources and the response time of web server in the overall will be shorten. Thus, the service performance of the server will increase under the situation that network access requests exceeds server's capability
- Under the static page cache mechanism, the frequently visited pages are stored in the cache of the server in static page, and the page indexes are saved in web server. If the client requests exist in the index package, the corresponding static page will be extracted from the cache server to return to the client. This mechanism shortens the response time of the server, and increases the amount of concurrent access and system processing efficiency of the system.

LCCWS reduces the CPU overhead effectively, decreases the communication data copying between memories, and improves communication efficiency effectively. Consequently the performance of the web servers is greatly enhanced. The comparison of traditional access mode and DMA mode is as shown in Fig. 8, and the comparison of page not cached and page cached is as shown in Fig. 9

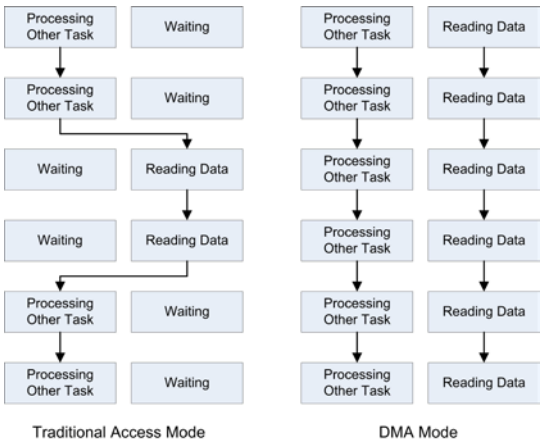



Figure 8. Comparison of Traditional Access Mode and DMA Mode

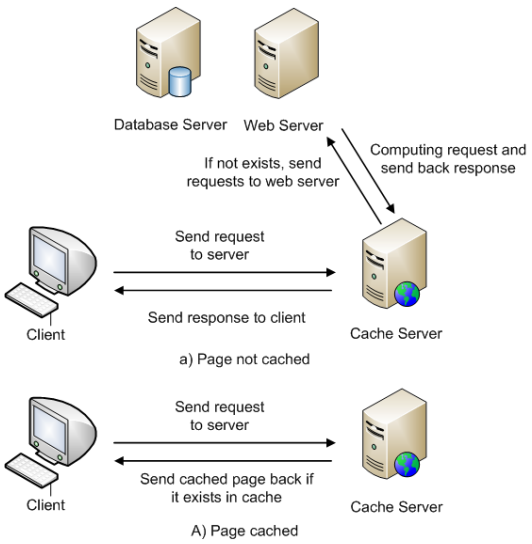



Figure 9. Comparison of page not paged and page cached

## V. Simulation Experiment

Because of resource constraints, there's no computer supporting DMA mode. Its performance cannot compare with the Tomcat server's, so we carried out a simulation experiment.

The simulation experiment used the sending side of the Distributed DoS Pressure Measurement System to generate data traffic. The experiment is divided into four parts: firstly, we compared the ordinary libpcap packet capture rate with that installed the PF_RING kernel

modules. The libpcap with PF_RING modules uses the NAPI mechanisms and MMAP mechanism, and its packet size is between 512 bytes to 1500 bytes. Packet capture rate raised up 19.86%. After, LCCWS system reduced once data copy and the packet size keeps between 512 bytes and 1500 bytes. Compared with the original system, the response time was improved by 2.35%.Then, static page cache mechanism enabled LCCWS system and a certain number of pages index was established. For the same set of network access requests, compared with the original system without the mechanism, the average response time improved by 21.96%. Finally, the LCCWS system enabled multi-queue priority control mechanism and improved each queue. For the same group network access requests, which the number of exceeded the upper limit of the server's capability, compared its corresponding rate with the system without the mechanism, the average response time was up 10.73%.

From the experiments above we can see, loading the PF_RING module and combining the NAPI and a MMAP mechanism, packet capture performance is improved; Reducing data copying between the protocol stack and the upper application and adopting static page cache mechanism and multi-queue priority control mechanism, also improves the web server response speed. By using aforementioned four measures comprehensively the performance of the improved web server LCCWS will be improved.

## VI. CONCLUSION

The LCCWS prototype system designed in this paper improves the web server performance bottleneck through the zero-copy technologies, optimization of the TCP/IP protocol stack etc. This prototype system provides research foundation for the development of rapid and efficient web server system. But the protocol suite of the system still needs to improve, some functions such as packets retransmission, flow and congestion control also need further study.

The later work is mainly to improve the lightweight TCP/IP protocol stack and Socket communication function to implement the strongly practical, feature-rich, fast and efficient web server prototype system step by step.

## ACKNOWLEDGMENT

This work was supported in part by a grant from the National Natural Science Funds No.60970129.

## REFERENCES

[1] N. Yao, M. Zheng and J. Ju, "A High Performance Web Server based on Pipeline," *Journal of Software,* vol. 14, no. 6, pp. 1127-1133, 2003.

[2] "How to Improve Web Servers Performance," [Online]. Available: http://www.macrounion.com. [Accessed 12 11 2012].

[3] X. Ke, Z. Gong and J. Xia, "Research of the Zero-copy Technique and Its Implementation," *Computer Engineering & Science,* vol. 22, no. 5, pp. 17-20,24, 2000.

[4] L. Wang, M. Wang and X. Wang, "Research and Implementation on Linux Paekets Capturing of Gigabit Network," Jinan, 2007.

[5] "Luca Deri.Improving Passive Packet Capture:Beyond Device Polling," [Online]. Available: http://luca.ntop.org/Ring.pdf. [Accessed 11 11 2012].

[6] "Web Servers under the Linux Platform," [Online]. Available: http://www.hackchina.com. [Accessed 10 11 2012].

[7] H. Zhu, C. Zhou and G. Chang, "Research and Implementation of Zero-Copy Technology Based on Device Driver in Linux," in *In Proceeding of International Multi-Symposiums on Computer and Computational Sciences*, 2006.

[8] B. Wang, B. Fang and X. Yun, "The Study and Implementation of Zero-Copy Packet Capture Platform," *Chinese journal of computers,* vol. 28, no. 1, pp. 46-52, 2005.

[9] D. Stancevic, "Zero copy I: User-mode Perspective," *Linux Journal,* vol. 2003, no. 105, pp. 1-3, 2003.

[10] "Luca Deri,PF_RING User Guide," [Online]. Available: http://luca.ntop.org/User Guide .pdf. [Accessed 20 11 2012].

[11] L. Wang, "Working Principles of DMA and its Application to Improve hard Disk Speed," *Heilongjiang Science and Technology Information,* no. 18, p. 64, 2010.

[12] J. Song, X. Xie and S. Ran, "Simplified and Zero-copy TCP/IP Protocol based on Data Captrure System," *China measurement technology,* vol. 33, no. 1, pp. 114-117, 2007.

[13] H. Xu, J. Liu and Y. Wang, "Simplified Realization of Embedded TCP/IP Protocol Stack Based on ARMCore," *Application Research of Computers,* no. 10, pp. 251-253, 2006.

[14] G. Cao, "Analysis of the Linux kernel network stack source code," Beijing, Posts & Telecom Press, 2010, pp. 146-181,235-612.

[15] A. K. Amer Mohammed Al-Canaan, "Multimedia Web Services Performance: Analysis and Quantification of Binary Data Compression," *Journal of Multimedia,* vol. 6, no. 5, pp. 447-457, 2011.

[16] D. B. Lorenzo Sommaruga, "Towards a Semantic Web Based Model for the Tonal System in Standard IEEE 1599," *Journal of Multimedia,* vol. 4, no. 1, pp. 40-45, 2009.

[17] "Analysis of Web Server Performance Bottlenecks," [Online]. Available: http://www.yesky.com.

[18] J. Hu, J. Li and Z. Zeng, "SWSCF: A Semantic-based Web Service Composition Framework," *Journal of Networks,* vol. 4, no. 4, pp. 290-297, 2009.

[19] J. Zhu, H. Wu and G. Gao, "An Efficient Method of Web Sequential Pattern Mining Based on Session Filter and Transaction Identification," *Journal of Networks,* vol. 5, no. 9, pp. 1017-1025, 2010.

[20] Y. Zhou and L. Li, "Network Programming Technique and Implementationin TCP/1P Protoco," *Aeronautical Computer Technique,* vol. 32, no. 3, pp. 123-124,128, 2002.

**Haipeng Qu,** was born in Qingdao of China on Dec. 6, 1972. He got Master degree of Computer Applications Technology in Ocean University of China, Qingdao, China, in Jul. 2002 and PhD degree of Information Security, in State Key Laboratory of Information Security, Institute of Software of Chinese Academy of Sciences, Beijing, China in Mar. 2006. His major field of study is information security.

He has been working in the department of Computer Science and Technology since Mar. 2006, published articles: Fast Ethernet Instant Monitor Customizable Memory Access Model, Proceedings of NetSec 2007, and An IP Trace-back Scheme with Packet Marking in Blocks, Journal of Computer Research and Development, 2005. Current and previous research interests include information security and network security, delay-tolerant networks, underwater acoustic sensor networks, etc.

**Lili Wen** was born in Liaocheng, Shandong Province. She got her Master degree of Computer Applications Technology in Ocean University of China, Qingdao, China in Jun, 2011. Her major field of study is information security and computer network.

**Yanfei Xu** was born in Nangning, Guangxi Province. He got his Bachelor degree of Computer Science and Technology in Ocean University of China, Qingdao, China in Jun. 2011. His major field of study is information security and computer network.

**Ning Wang** was born in Laiwu, Shandong Province. She got her Bachelor degree of Computer Science and Technology in Ocean University of China, Qingdao, China in Jun. 2011. His major field of study is information security and computer network.

# Self-Adaptive and Energy-Efficient MAC Protocol Based on Event-Driven

Xin Hou
College of Computer and Information Engineering Zhejiang Gongshang University, Hangzhou, China
Email: houxinemail@163.com

Xingfeng Wei, Ertian Hua and Yujing Kong
College of Computer and Information Engineering Zhejiang Gongshang University, Hangzhou, China
Email: miss_wxf310@163.com, huaertian@mail.zjgsu.edu.cn, kongyujing2007@163.com

***Abstract*—Combined with WSN MAC layer protocol characteristics and design requirements, according to the characteristic of WSN monitoring application requirements, this paper puts forward a method based on event driven MAC protocol. The agreement algorithm is to solve the problem of network congestion and node energy unnecessary consumption cause by a large number of redundant monitoring data transceiver. It is a kind of adaptive low power consumption of the MAC layer protocol, which is pointed out based on theoretical foundation of S_MAC protocol, made use of the event driven mechanism system theory, combined with event driven mechanism and the characteristics of the WSN. It has the periodic dormancy mechanism of S_MAC protocol, in the premise of the reliability data, to reduce data redundancy and communication delay time, improve the overall network throughput, to ensure the safety and reliability of the network, which can greatly extends the node of working time.**

***Index Terms*—Wireless sensor network; Event driven; Low power consumption; MAC protocol**

## I. INTRODUCTION

Wireless sensor network (WSN) [1], composed by a large number of small sensor nodes, deployed in the monitoring area, formed a multi-hop self-organizing network system via wireless communication, aims to collaborate awareness by collecting and processing network coverage area in the perceived object information, and then send the consequence to observers. Information technology can achieve a mass of information storage, high-speed transmission, and rapid processing. The development in Micro-sensor technology, microelectronics, wireless communications technology and computing technology has greatly promoted the development of WSN technology. With the development of WSN technology, WSN have a wide range of applications in the military, environmental monitoring, industrial control, medical and civilian fields [2].

However, due to the applications, price and volume limitations, sensor network nodes are powered by battery which is limited and difficult to update, computing power and storage space is also very limited [3, 4]. In order to ensure the WSN long-term effective work, to take an effective MAC protocol to reduce energy consumption and maximize the network lifetime is very important. Therefore, low power, low complexity is a key technology of WSN.

Combined with WSN MAC layer protocol characteristics and design requirements, this paper designs a low-power and adaptive MAC layer protocol based on event-driven on basis of S_MAC protocol for monitoring the environment in the high-voltage transmission grid [7,8,9], to avoid a large number of cumbersome repeat monitoring data caused by network congestion and unnecessary node energy consumption. Through the analysis of experimental data, this algorithm can reduce the cumbersome data transceiver, reduce the data redundancy of the cluster head node, and to some extent, improve the practicality and reliability of the data. The contents of this article is organized as follows: the first part describes the current situation of the WSN; the second part describes the related work, including the energy consumption of wireless sensor nodes, MAC protocol and event-driven evaluation system; the third part describes the model of the entire WSN; the agreement algorithm is described in detail in the fourth part; the fifth part is about the experiment and its results; finally, conclusions and further research are described.

## II. RELATED WORK

### A. Energy Consumption Analysis of Wireless Sensor Nodes

Hardware of typical wireless sensor network terminal node is mainly composed of four basic modules, which are the processor module, wireless communication module, sensor module and the energy supply module. Block diagram of node function is shown in Fig.1. Sensor module is usually constituted by two parts of the sensor and A/D converter. Sensor node generates an analog signal through observation and sensing the surrounding environment information, and then converts into a digital signal through the A/D converter. Processor module comprises a microprocessor and a memory two subsystems, and the microprocessor is used to process the internal information data of the node, while memory is used to store data of management information. Wireless communication module is responsible for the communication and exchange information with the


doi: 10.4304/jnw.8.1.174-181

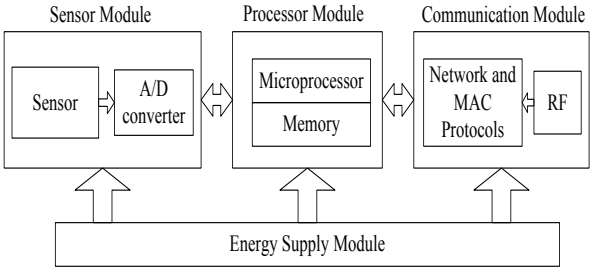



Figure1. Block diagram of node function

neighboring nodes; energy module is a very important part of a sensor node, which provides energy for the other part of the sensor node. Furthermore, according to requirement of the network applications, different sensor nodes may include some other component parts, such as the positioning system, the mobility management system, the energy regeneration system, etc.

As the power provided by the power supply module is limited, the node of every module needs to adopt original device in order to meet the characteristics of low-power characteristics. In four basic modules of the Wireless sensor nodes, processor modules, wireless communication module and the sensor module are mainly power module.

*1) The energy consumption analysis of the sensor module*

The energy consumption of the sensing module includes environmental signals sampled, physical signal into a digital signal, and the signal modulation, etc. As the sensor module work is determined by the specific application, its working mode is decided by the specific application environment, which may be bust or cycle-type, so its energy consumption is substantially fixed. Its consumption of total energy consumption is equal to the product of the single sampling energy and the sampling frequency.

*2) The energy consumption of the processor module*

Microprocessor power is mainly decided by the production process of the microprocessor itself, including its operating voltage, the operating clock and the internal logic. The higher the voltage, the faster the running speeds, the more complex internal logic, so its power consumption is greater. In order to reduce the energy consumption of the processor module, it can be considered from the operating mode of the processor itself and data processing algorithms. The working mode of the processor includes the "run", "free" and "sleep", then the processor is considered as much as possible so that it is not in the idle state. The complexity of the process or internal data processing algorithms should try to simplify.

*3) The energy consumption of the wireless communication module*

The tasks of the wireless communication module is responsible for communication between nodes of information data, and its energy consumption is mainly derived from the two parts of the RF signal generation and signal analysis. The energy consumption of the RF signal is determined by the antenna transmission power and is related to the modulation mode of the RF generator and the target node distance. Energy consumption of the signal analysis is mainly the components of the frequency synthesizer, frequency conversion and filtering. However, the energy consumption of this section is essentially fixed. In order to save energy, it should be possible to make the wireless communication module in sleep state.

The experimental data shows that the energy consumed by the node to send the information data of the 1 bit sufficient processor to perform 3000 computer instructions. Therefore, the major energy-consuming module of node is the wireless communication module. Shown in Fig.2, the energy consumption of the wireless communication module to send/receive occupies the major part of the communication energy consumption, and idle state also has a higher power consumption is almost the same as the receiving state, while the dormant state energy consumption is far lower than in work mode and idle state energy consumption. It is a highly efficient energy-saving is to ensure that WSN can guarantee to complete the task (i.e. the data is not distorted) under the premise of effective low complexity of data processing programs to control the transceiver to reduce redundant data, as far as possible, so that RF module in a dormant state.

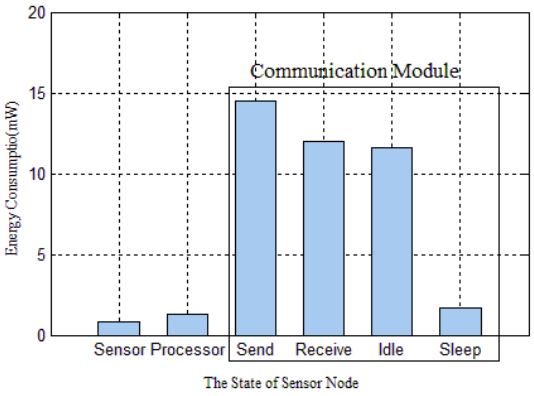



Figure2. Energy consumption of sensor node component unit

## B. MAC Protocol

In WSN, the node energy is limited and difficult to supplement. To ensure the WSN long-term effective, the WSN network protocol stack infrastructure of media access control (MAC) protocol, which determines the use of the radio channel and allocates for node the resources of the wireless communication, is a direct impact on overall network performance and has become the top priority of the WSN network protocol. To reduce energy consumption and maximize network lifetime is primary design goals of MAC protocol. Secondly, in order to adapt the node distribution and topology changes, MAC protocol needs to have good scalability; and concerning of traditional wireless networks real-time, throughput and bandwidth utilization and other performance indicators as a secondary goal. Currently, there are a large number of

different characteristics and specific applications of the WSN MAC protocols have been proposed. One of the typical agreements is S_MAC protocol.

S_MAC protocol's main objective is to reduce energy consumption. In order to reduce the energy consumption of idle listening, the protocol uses periodic work/sleep mechanism [10], letting those not participate in the sending or receiving nodes into sleeping, thereby reducing energy consumption. Shown in Fig. 3, S_MAC protocol divides time period into multiple frames. Each frame consists of two parts, the active time and sleep time. Active time is divided into two parts, the synchronization time and data time. Synchronization time maintain synchronization scheduling by sending SYNC packets. To communicate with the adjacent nodes in data time, and send the message queue of the dormant period. In order to reduce the energy consumption caused by the data conflicts, protocol adopts backoff mechanism to compete for the channel to complete to receive or send data. The protocol adopts the adaptive circular listener mechanism and RTS/CTS mechanism to reduce the crosstalk listener to bring the energy consumption. Each node in the transmission of data goes through the communication process RTS/CTS/DATA/ACK (except for broadcast packets), to avoid cross-talk listener.

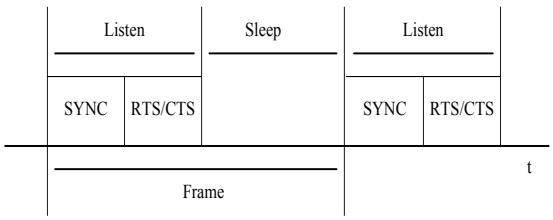



Figure3. The work/sleep mechanism of S_MAC.

### *C. The Event-driven Evaluation System*

The basic view of the event-driven mechanism is a system can dynamically identify the event status of its internal or external, based on the performance of the external, and then make the corresponding strategies to meet certain characteristics of the entire system. Shown in Fig.4, in the monitoring system, the continually collected data of sensor nodes is the input, then the event receiver identifies the event status, and finally the event processor makes the appropriate strategies and measures to mobilize available resources to perform tasks. The final processing results are the output. Some systems also need a feedback mechanism.

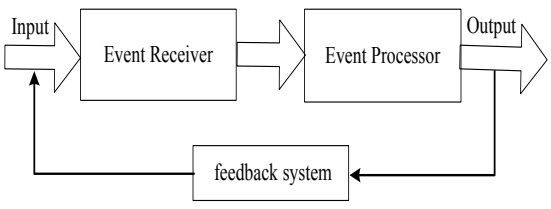



Figure4. The diagram of event driven mechanism system

In the power grid monitoring system often can't predict what event will happen or when the event will happen, so it is not possible in advance to do the appropriate measures. Therefore, the paper put forward a theory of event-driven mechanism system to refer to identify the event status. Event-driven mechanism is firstly to do some process environment parameters (such as temperature, acceleration, etc.) collected by sensor node, to judge whether there are any changes by system rules. Finally, to a corresponding strategy depend on the event status to meet the low-power adaptive system.

## III. THE DESCRIBE OF SYSTEM MODEL

The geographical environment of the high-voltage transmission grid is very complex. The paper simplifies the system model according to its architecture, shown in Fig.5. In such a network model, most of the time of the surrounding environment is relatively stable, and occasionally sudden incident, such as strong winds from time to time, wire local shaking strongly. Nodes can communicate with each other, and the node has ability to process data. Data communication mode as follow: Each node firstly sends data to the cluster node; Cluster node goes through a series of processing of the data analysis and integration, sends the processed data to base stations which communicate with the server directly by the serial port.

The monitoring geographical environment of Wireless monitoring system may be quit complex, but we can simplify a local system model as shown in Fig.5. In this monitoring environment, the monitoring environment is uncontrollable and unpredictable, and the monitoring system is also unable to predict the future of the object being monitored forthcoming of what happens, or when the occurrence of any event, so can't do the appropriate measures in advance. In such a network model, most of the time of the surrounding environment is relatively stable and occasionally sudden incident. Such like the high-voltage transmission grid environment, where sometimes strong winds lead part wire shaking or produce the high temperature by part sudden short circuit.

As shown in Fig.5, the nodes in its own monitoring area can communicate with each other. And the data communication mode of the wireless monitoring system: when the monitoring system is run, each sensor node sampling various types of data (such as temperature, acceleration, etc.) from environmental by sensors, and then send them by wireless communication to head node; head node processes data received according to the system design goals, and then sent to the base station; base station can be used as high-level controller and external interfaces, through the aggregation node communication, coordination, optimization, management, and directly with the server through the serial port communications.

Network quality of service is measured through the data packet loss rate, packet collision sampling rate, packet delays. For practical WSN's bandwidth is mainly determined by the packet size and sampling frequency. The packet size is determined by the protocol stack. So

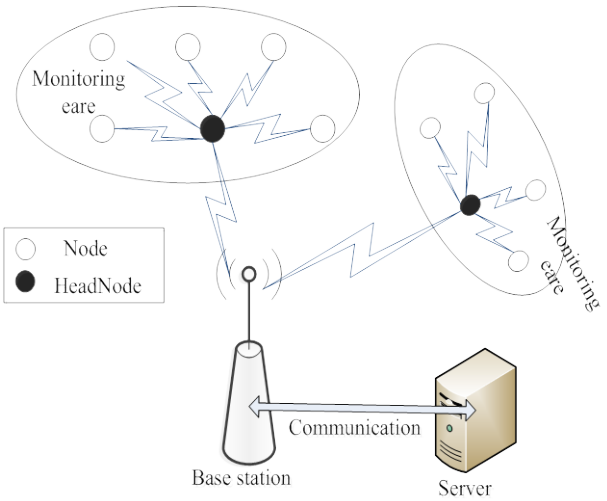



Figure5. System model of Wireless sensor data transmission

the bandwidth is mainly affected by the sampling frequency. It can be known by Shannon sampling theorem that the sampling frequency should be determined according to the characteristics of monitored objects, the higher the sampling frequency, the better the system performance; low sampling frequency can improve the network quality of service, but also resulted in network utilization and network bandwidth waste.

Known from above, the monitoring environment of WSN is dynamic changes. In order to ensure the network quality of service, it needs to design an event-driven mechanism system to dynamically change the sampling frequency of the sensor nodes and the node between the transceiver frequencies, which can be adapted to reliable dynamic quality of service in the entire network.

## IV. THE DESCRIPTION OF PROTOCOL ALGORITHM

### A. The Describe of Event-Driven Mechanism

The event driven mechanism has its own core. This paper designed four kinds of event status, which are idle event state, ordinary event state, intense event state and warning event state. The core is according on the correlation process the historical data to obtain an estimate of an event state and the event is about to occur in the future of the current environment, that is whether the event can cause danger and disaster. The core method is to collect data and calculate data volatility exceeds the Corresponding threshold, whose reflection whether is the current system of early warning.

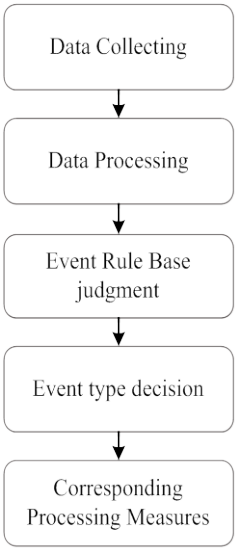



Figure6. The event driven mechanism framework

The event driven mechanism framework is shown in Fig.6. When system started running, the node begins to collect monitored parameters data by the sensor module. Then the node deals with the parameters data to evaluate the state of the current event according to the judgment result. Finally, depending on the event status, the node makes the appropriate event handler, which enhances or reduces the frequency of data collection and transmit.

### B. Rules of Algorithm and Processing

This algorithm is mainly to realize real-time monitoring and detect events' state. Then taking a series of processing consider on the different events, to achieve a low-power adaptive MAC layer protocol. So the algorithms should be regulated by the following rules.

*Rule1:* The definition of events.

Suppose a node needs to monitor n environment variables, which's warning event data characterized threshold respectively $[Y_1, Y_2, ..., Y_k, ..., Y_n]$, the normal event data fluctuation characteristics threshold respectively $[S_1, S_2, ..., S_k, ..., S_n]$, the intense event data fluctuation characteristics threshold respectively $[T_1, T_2, ..., T_k, ..., T_n]$. Such as $Y_1$ can be expressed as the node warning event of temperature fluctuations characterized threshold, $Y_2$ can be represented as the node warning event humidity fluctuations characterized threshold; $S_1$ can be expressed as the node temperature fluctuation of the ordinary event characteristics threshold, $S_2$ can be represented as the node humidity fluctuation characteristics of the ordinary event threshold; $T_1$ can be represented as the node ordinary temperature fluctuation characteristics of the event threshold, $T_2$ can be represented as the node ordinary event humidity fluctuations characterized the threshold value, and so on. If a node monitors the k variable values of $F(k) > Y_k$ $(1 \le k \le n)$, then we can consider that the node's k variable warning event occurred. If a node monitors the k variable valuesof F (k), which's fluctuation value $W(k) > S_k$ $(1 \le k \le n)$, then we can consider that the node's k variable normal event occurred. If a node monitors values of F (k), which's fluctuation value $W(k) > T_k$ $(1 \le k \le n)$, then we can consider that the node's intense incident of k variable occurred.

*Rule2:* The method of calculation data fluctuation.

Node energy and computing power constraints is limitation, can't retain too much historical data, yet cannot be computationally excessive handling. And know from the analysis of the energy consumption of the wireless sensor network node, in order to reduce energy consumption, the complexity of data processing algorithms in processor should be try to simplify. Therefore, in the present algorithm, the data is just only simple processing.

Fluctuations of the environment variable data can be calculated according to the (1), and in accordance with the definition of events from Rule1, to determine whether the event happens.

$$W = \frac{|A - A_0|}{T_M} \quad (1)$$

Where A is the current data collected, $A_0$ is the initial value, the $T_M$ is the actual data collection cycle, W is the curvature of the data changes.

*Rule3:* The method of processing data.

Considering that the latest data influence greatly, the sooner data the smaller influence, and reducing the influence of abnormal or irregular data from sensors, this paper points out that the average iteration algorithm to handle dynamic real-time collection data of sensor. The average iteration algorithm can ensure the validity and accuracy of the data, and meet the characteristic of the latest data of the most influential, the smaller the sooner the data influence.

Assuming that $A_1$, $A_2$, … , $A_n$, totally N historical data, $A_0$ is the initial value, and $A_0 = A_1$.

$$A_i^{'} = \frac{A_i + A_0}{2} \quad (2)$$

Where $A_i^{'}$ is the $i^{th}$ data processed data.

$$A_0 = A_i^{'} \quad (3)$$

First according to (2) takes the average value of the initial value $A_0$ and the latest data, to obtain the latest after the processed data $A_i^{'}$, and then according to the (3), to set the initial values as the $A_i^{'}$, for the next the new data used to obtain the next latest processed data.

According to the above, to respectively put i = 1,2, ..., n into (2) and (3), then the (4) can be obtained.

$$A_n^{'} = \frac{\frac{\frac{A_0 + A_1}{2} + A_2}{2} + \cdots + A_n}{2} \quad (4)$$

Equation (5) is got by Simplifying (4). It is clear see that the algorithm can meet the characteristic of rule3 by (5).

$$A_n^{'} = \frac{A_0}{2^n} + \sum_{i=1}^{n} \frac{A_i}{2^{n-i+1}} \quad (5)$$

*Rule4:* The sustainability of the event.

If the recent events, the event will continue for some time. So in this paper, when the monitored event occurs, the monitoring data will be continuously sent several data packets.

If the monitoring system monitors an event occurs, then the event will continue for some time. And when monitoring the incident, monitoring analysis platform and display platforms require more accurate historical data to monitor data. The algorithm processing program, when the monitored event occurs, the node needs to continuously transmit several monitoring data.

*Rule5:* The processing measures of different event

When event level is different, it needs to adopt different frequency to collect data. The events are more intense, it needs more precise the data to describe. In paper, when sharp events occur, it needs to increase the frequency to gather sensor data, to improve the accuracy of the data. While when sharp event end or event-free status, it needs to decrease the frequency to reduce energy consumption.

### *C. The Treatment Methods of Different Events*

In this paper, it adopts the different method to process data for three different events state. When it is idle event state, then not to send the current data; when the ordinary event state happens, then it sends the current data; when the intense event happens, it will double frequency of sensors to collect data, and then send data; when the warning event state happens, it will sends the current data and the warning flag.

When it is idle event state, the algorithm does not send the data, but it needs to process historical data for getting a more reasonable current data according to Rule3. It uses the (4) to calculate the reasonable the value of current data, and sets the common frequency for node sensor to collect data. In order to distinguish the event of lost node (the cluster head cannot receive a node data for a long time that means the node has been lost), so even if the non-event occurs, the algorithm also needs to send data in fixing time. When the $T_{send}$ is coming ($T_{send}$ is the max delay send data cycle), then it needs to send the current data.

When ordinary event state, it needs to send the current data, and set the common frequency for node sensor to collect data. According to the rule4, it is required continuously send and receive several collected current data. And then initial and recalculate the value of W, which is to determine the current event state.

When intense event occurs, it needs to send the current value and set the collecting frequency of node as strengthen frequency according to the rule5. Also it needs to constantly calculate the value of W and send some collected current data, until the end of the dramatic events, and turn into general event state or non-event state.

When warning event state, it needs to send the current data and the warning flag, and set the common frequency for node sensor to collect data. According to the rule4, it is required continuously send and receive several collected current data. And then initial and recalculate the value of W, which is to determine the current event state.

### *D. Protocol Algorithm Concrete Steps*

As shown in Fig.7, the concrete steps of protocol algorithm as follow:

Step1: The initial value should be set as follow: $T_{send}=N*T_{collect}$, $T_{real}=T_{collect}$, $A_0=A'=A$, where $T_{send}$ is the collecting frequency, $T_{collect}$ is the collecting frequency, $T_{real}$ is the real collecting frequency, and the N is 20 in general, $A_0$ is the initial data, A' is the processed data, A is the collecting data. Then the flag of $T_{send}$ should be set as 0. Finally, each type of data's threshold value should be set as follow: the warning event data characteristic threshold respectively as [$Y_1$, $Y_2$, ..., $Y_k$, ...,

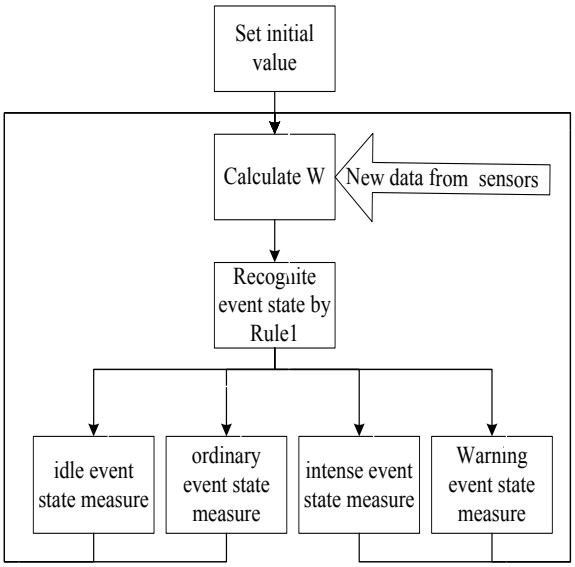



Figure7. The flow chart of protocol algorithm

$Y_n$], the normal event data fluctuation characteristics threshold respectively as [$S_1$, $S_2$, ..., $S_k$ , ..., $S_n$], the intense event data fluctuation characteristics threshold respectively as [$T_1$, $T_2$, ..., $T_k$ , ..., $T_n$].

Step2: The warning flag is set 0, then according to the (2) and (3), the A' and $A_0$ can be calculated by the new collecting data A.

Step3: According the (1), the value of W can be calculated. Then according the rule1, the event state can be obtained.

If it is idle event state, $T_{real}$ is set as $T_{collect}$, then it goes into step4;

If it is ordinary event state, $T_{real}$ is set as $T_{collect}$, then it goes into step5;

If intense event state, $T_{real}$ is set as half of $T_{collect}$, then it goes into step6;

If it is warning event state, the warning flag is set 1, $T_{real}$ is set as $T_{collect}$, and then it goes into step5;

Step4: The flag of $T_{send}$ will be checked, which means whether the time slot of $T_{send}$ is coming. If the flag of $T_{send}$ equals 1, then the A' needs be sent, the flag of $T_{send}$ need be set as 0. At last, it goes into the step2.

Step5: The A' needs be sent. Then it needs to constantly collect 20 data and send them. Then set the $A_0$ and A' as the last collecting data A. At the meanwhile, the flag of $T_{send}$ should be set as 0. At last, it goes into the step2.

Step6: The A' needs be sent. Then it needs to constantly collect 40 data and send them. Then set the $A_0$ and A' as the last collecting data A. At the meanwhile, the flag of $T_{send}$ should be set as 0. At last, it goes into the step2.

## V. EXPERIMENTS AND DATA ANALYSIS

To testify the feasibility of this algorithm, we simulated the algorithm with the physical node and base station in a lab environment. At first, we fixed two nodes together, and maintained them in the same state, to guarantee the two nodes collect monitoring data, such as temperature, humidity, acceleration, basically the same. Secondly, we set the node1 and node2 collect the data per 10 seconds. Thirdly, we set the node1 send the data immediately when collecting the new data, while node2 send data using this algorithm. In experiment, we set the threshold of node2 as follow: [$Y_1$, $Y_2$, $Y_3$, $Y_4$, $Y_5$]=[60,60,0.5,0.5,1.5], [$S_1$, $S_2$, $S_3$, $S_4$, $S_5$]=[0.05, 0.135, 0.005, 0.005, 0.005], [$T_1$, $T_2$, $T_3$, $T_4$, $T_5$]=[0.2, 0.3,0.015, 0.015, 0.015]. Here the "Y" representative warning event data characteristic threshold, "S" representative normal event data fluctuation characteristics threshold, "T" representative intense event data fluctuation characteristics threshold, and the number "1", "2", "3", "4", "5" respectively representative temperature, humidity, X acceleration, Y acceleration, Z acceleration. For example, the "$Y_1$" means the temperature warning event data characteristic threshold.

After nearly four hours of data measurement, we got the some data shown in Fig.8, Fig.9, and Fig.10. We can see that the various types of data of the node1 and node2 curves are basically similar trend, and can be roughly reflect the same environmental characteristics. Fig.11 expresses that the number of data the base station get form node1 and node2 in this period. We can see when the node1 sent 1000 each type of data, the node2 sent the number of data is significantly far less than the node1, that the data number of the temperature type is 140, the humidity type is 154, the X-axis acceleration type is 262, the Y-axis acceleration type is 231, the Z-axis acceleration type is 210. Through the above, we can see the algorithm, under the premise without data distortion, can reduce the data redundancy, reduce the energy consumption of nodes, and to a certain extent, improve the authenticity and reliability of data.

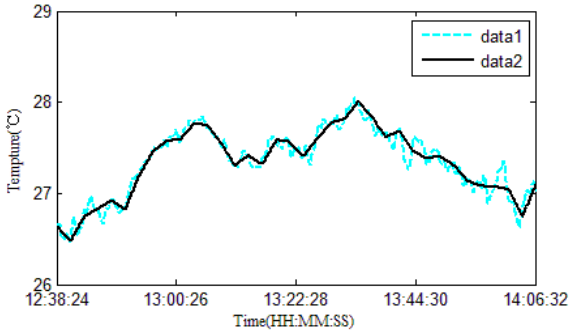



Figure8. The nodes temperature data curve

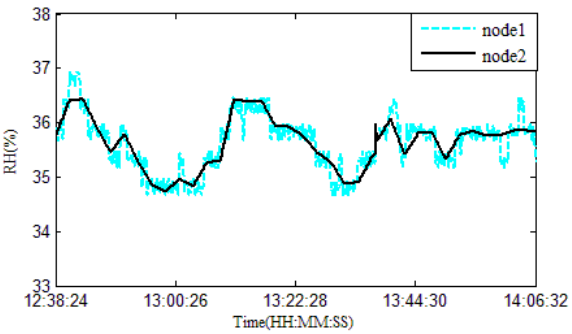



Figure9. The nodes humidity data curve

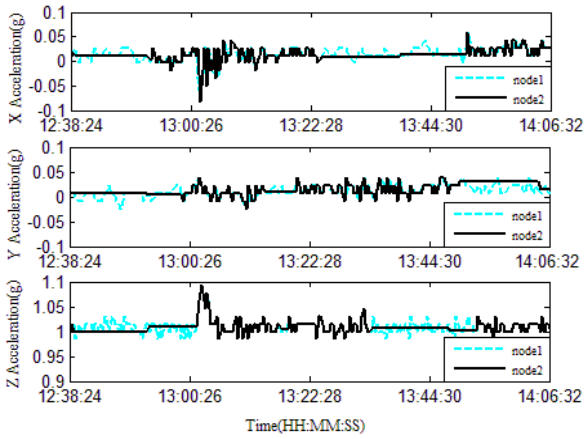



Figure10. The nodes acceleration data curve

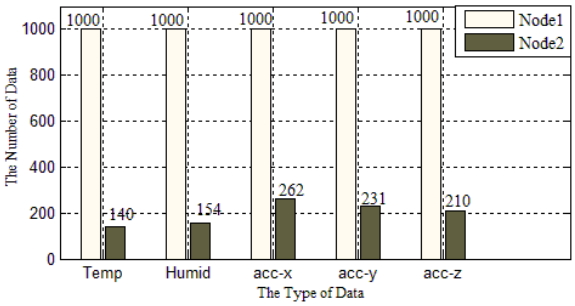



Figure11. The number of data each node send chart.

## VI. Conclusion and Future Work

This Paper designs a low-power and adaptive MAC layer protocol based on event-driven on basis of S_MAC protocol. By experiments, this algorithm can be seen greatly reducing the cumbersome and repeat data, to avoid network congestion caused due to a large number of cumbersome data, and reduce the energy of the node, basing on ensure the data is effective to describe the current state of the monitoring environment.

Cluster head node energy consumption is much larger than the other nodes, so future work is to study how adaptive cluster within the cluster head rotation mechanism to extend the lifetime of the overall network, and further study of the event to determine the mechanism and internal data of node fusion technology.

## Acknowledgment

This work was supported in part by the Zhejiang Major Science and Technology and Priority Themes Projects (No. 2010C11051) and Foundation of Zhejiang Educational Committee (No. Y201225796)

## References

[1] SUN Li-min, LI Jian-zhong, CHEN Yu, etc. *The wireless sensor network*. Beijing: Tsinghua university press, 2005.

[2] Akyildiz I F, Weilian S, et al. *A survey on sensor networks. IEEE Communications Magazine*, vol.40, pp.102-114, 2002.

[3] SU Jun, HU Fang-yu. Energy-saving Improvement for SMAC Protocol in WSN. Computer Engineering, vol.35, pp.106-107, 2009.

[4] Xiang-yu, FENG Dong-qin. Design and Implementation of Low Power Consumption Industrial Wireless Sensor Networking. Control and Instruments in Chemical Industry, vol.35, pp.47-50, 2008.

[5] WANG Da-hai, LI Jie, XIE Qiang. Dynamic tension model for wind-induced vibration of long spanned transmission line. Proceedings of the CSEE, vol.29, pp.122-128, 2009.

[6] LU Jia-zheng, ZHANG Hong-xian, FANG Zhen, etc. Result and its analysis of ice disaster monitoring of Hunan power system. Power System Protection and Control, vol.37, pp.99-105, 2009.

[7] K Jamieson, H Balakrishnan, Y C Tay. Sift: A MAC protocol for event-driven wireless sensor networks. *MIT, Tech Rep: MIT-LCS-TR-894,* 2003.

[8] XIN Ying, XIE Guang-zhong, JIANG Ya-dong. Wireless temperature and humidity sensor network based on Zig Bee protocol. Transducer and Microsystem Technologies, vol.25, pp.82-85, 2007.

[9] ZHAO Zeng-hua, SHI Gao-tao, HAN Shuang-li, SHU Yan-tai, ZHOU Wen-tao, CHEN Jian-min. A Heterogeneous Wireless Sensor Network Based Remote District High-voltage Transmission Line On-line Monitoring System. Automation of Electric Power Systems, vol.33, pp.80-84, 2009.

[10] ZHU Yu-lin, YAN Jun, SHEN Ming-hua, ZHANG Xue-fan. Improvement and Implementation of MAC Protocol for WSN Based on Wake. Computer Engineering, vol.36, pp.108-110, 2010.

[11] ZHAO Guohong. Multiviews on power administration. China Power Enterprise Management, vol.2, 2004.

[12] Fan Li, Guizhong Liu, Lijun He. Cross-Layer Approach to Multiuser H.264 Video Transmission over Wireless Networks. Journal of Multimedia, vol.5, pp.110-117, 2010.

[13] Guoqiang Zheng, Shengyu Tang. Spatial Correlation-Based MAC Protocol for Event-Driven Wireless Sensor Networks. Journal of Networks, vol.6, pp.121-128, 2011.

[14] Qian Hu, Zhenzhou Tang. ATPM: An Energy Efficient MAC Protocol with Adaptive Transmit Power Scheme for Wireless Sensor Networks. Journal of Multimedia, vol.6, pp.122-128, 2011.

[15] FU Zhengcai , WU Bin , HUANG Xiandong, et al. Online monitoring system based on GSM network for transmission line fault. High Voltage Engineering, vol.33, pp.69-72, 2007.

[16] Ren FY, Huang HN, Lin C. Wireless sensor networks. Journal of Software, vol.14, pp.1282-1291, 2003.

[17] Bharghavan V, Demers A, Shenker S, Zhang LX. Macaw: A media access protocol for wireless Lans. *In: Proc. of the ACM SIGCOMM Conf. London: ACM Press,* 1994, pp.212-225.

[18] Ye W, Heidemann J, Estrin D. Medium access control with coordinated, adaptive sleeping for wireless sensor networks. *IEEE/ACM Trans. on Networking,* vol.12, pp.493-506, 2004.

[19] Jamieson K, Balakrishnan H, Tay YC. Sift: A MAC protocol for event-driver wireless sensor networks. Technical Report, *MIT-LCS-TR-894, MIT*, 2003.

[20] Zheng T, Radhakrishnan S, Sarangan V. P-MAC: An adaptive energy-efficient MAC protocol for wireless sensor networks. *In: Proc. of the Parallel and Distributed Processing Symp. Piscataway: IEEE Computer Society,* pp.237-247, 2005.

[21] SI Haifei, YANG Zhong; WANG Jun. Review on research status and application of wireless sensor networks. Journal

of Mechanical & Electrical Engineering, vol.28, pp.16-20, 2011.
[22] CHEN Huihui. Wireless Sensor Network(WSN)Node Energy Loss Analysis and Visualization. Logistics Engineering and Management, vol.34, pp.108-110, 2012.

**Xin Hou**, 1981, male, associate researcher of Zhejiang Gongshang University. His main research direction is wireless sensor network, intelligent control and embedded system.
Email: houxinemail@ 163.com

**Xingfeng Wei**, 1988, male, the master student of Computer Applications Technology of Zhejiang Gongshang University. His main research direction is intelligent control and embedded system.
Email: miss_wxf310@163.com

**Ertian Hua**, 1963, male, Ph.D and professor. His main research field is the product design and theory, Mechanical and Electrical Integration technology.
Email: huaertian@mail.zjgsu.edu.cn

**Yujing Kong**, 1986, female, the master student of Computer Applications Technology of Zhejiang Gongshang University. Her main research direction is intelligent control and embedded system.

# An Improved Retransmission-based Network Steganography: Design and Detection

Jiangtao Zhai
School of Automation, Nanjing University of Science and Technology, Nanjing, China
Email: jiangtao_zhai@yahoo.com.cn

Guangjie Liu and Yuewei Dai
School of Automation, Nanjing University of Science and Technology, Nanjing, China
Email: gjieliu@gmail.com, daiywei@163.com

***Abstract*—Network steganography is a covert communication technique that uses redundancies in network protocols to transfer secret information. The retransmission-based steganography (RSTEG) embeds covert messages into the payload field of the intentionally retransmission packets. So its capacity is higher than most of the existing methods. Because TCP checksum field of the original packet is different from that of the retransmitted packet, RSTEG is not stealthy in fact. An improved method named IRSTEG is presented to resolve the flaw by introducing the payload compensation. Further, a method is proposed to detect IRSTEG based the payload segment comparison. Experiments show that the method can detect IRSTEG well.**

***Index Terms*—payload segment comparison, retransmission-based network steganography, covert communication**

## I. INTRODUCTION

With the rapid development of network communication technology, network steganography which uses redundancies in protocols to transfer secret information has attracted lots of attentions of researchers. Network steganography is also named network covert channel which can be divided into two kinds—covert storage channel and covert timing one [1]. The first is to embed the message bits into the packet header field or the payload. The second is to encode the information by changing the packet transmission rate or the packet time delays. The early network steganography [2, 3, 4, 5] utilized the unused field in TCP/IP header to transmit covert message. The kind of methods is simple but cannot avoid the pattern of the normal packets being modification. It is easy to make the reliable detection by checking the pattern such as the progressive increase of IPID filed, all zeros in the default extended field of TCP etc. Hence, the researchers have to develop new methods.

For example, Ji [6] et al. proposed a method which used different packet lengths to transmit covert messages and made the distribution of packet lengths near to the normal one as close as possible. Szczypiorski [7] et al. proposed a method to use VoIP (voice over internet protocol) packets as carriers to transmit covert messages. In their method, the sender delays some packets intentionally and the receiver distinguishes these packets from network traffic and picks up covert messages. Yao [8] et al. proposed an ON/OFF covert timing channel which could make the distribution of the covert time delays close to that of the overt ones. Gianvecchio [9] et al. proposed a covert timing channel based on the model of normal time intervals. Liu [10] et al. proposed a simple but efficient covert timing channel with distribution matching. Those methods mentioned above

Those methods mentioned above represent the main direction to design more secure and stealthy network steganography through preventing the statistics of the covert network packets not or less being changed. However, the covert communication rate of those methods is still very low. In [11], Szczypiorski et al. proposed a retransmission-based steganography (RSTEG). The authors argued that retransmissions caused by network overload, excessive delays or reordering packets accounted for about 7% in total Internet packets [12, 13, 14]. By replacing the payload field of those retransmitted packets, RSTEG can achieve higher capacity than the traditional methods.

Because the checksum of the retransmitted packet generated by RSTEG is different from that of the original packet, one can make the detection just by comparing the captured retransmitted checksum with that of the recoded one. So, we make an improvement on RSTEG based on payload compensation, which can counter the checksum-comparison-based detection method. Further, a more efficient detection method based on payload segment comparison is proposed.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 introduces the improved RSTEG (IRSTEG) algorithm and its detection method. Section 4 gives experimental results and corresponding analysis. Section 5 concludes the whole paper.

## II. RELATED WORK


doi: 10.4304/jnw.8.1.182-188

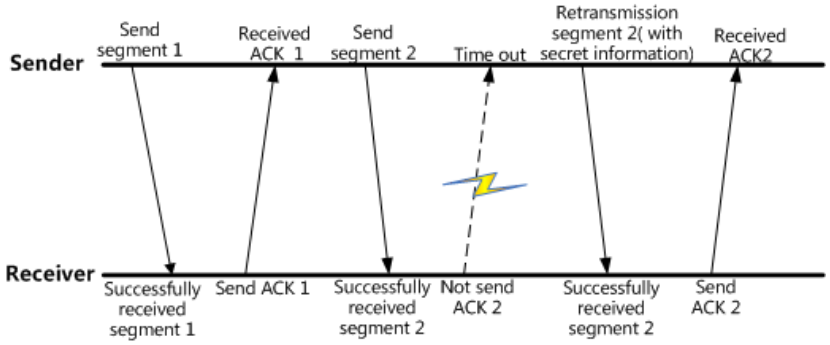



Figure 1 The communication process illustration of RSTEG based on TCP timed-out

### *A. Principle of RSTEG*

The communication process of RSTEG based on TCP time-out is illustrated in Fig.1. In the process, despite that the receiver gets a segment successfully, the ACK packet is not sent intentionally. If the sender doesn't receive the ACK packet within a fixed time, the last packet will be retransmitted. In the retransmission step, the payload field of the retransmitted packet is replaced by the covert message bits. When the receiver gets the retransmitted packet, he can obtain the covert messages and send ACK packet back. In [11], the authors also give some other retransmission-based methods which are similar to this one, and the main difference is just the retransmission condition. So here we only discuss RSTEG based on TCP time-out.

### *B. Detection Methods against RSTEG*

Based on the characteristics of RSTEG, we think that there are three potential methods that can be used to make detection. The first method is based on retransmission probability check which was also mentioned in [11]. The basic philosophy is that if the retransmission probability is larger than a fixed threshold such as 7%, the traffic will be judged abnormal. It may be efficient for RSTEG using a larger retransmission probability, but will be invalid when the retransmission probability is set blow the detection threshold.

The second method is based on the direct payload comparison. The detection process is that when a TCP retransmission is observed, the payload fields of the original and the retransmission packet are compared. If they are same, the retransmission will be considered abnormal, else it will be considered normal. The method needs large cache space because all packets without being acknowledged should be stored beforehand. For example, suppose that the speed of one TCP link is 50 packets per second with each packet length sized 1500B, the retransmission probability is 5% and the number of TCP links in the gateway is 200. As the TCP window size is 16 bit, the maximum receive number of one TCP link is 65535B. It is to say that an ACK packet will be returned every 43 packets. If the payload comparison method is used, there are totally about 2GB ( $(50-43)\times 1500\times 200$ ) memory is needed per second. It must bring about the burden of store and computation.

The last method is based on the comparison of TCP checksum values. It is known that after the original packet payload is replaced, the TCP checksum will be recalculated to keep the legality of the new packet. Generally, the new checksum is different from the original one with a very high probability. So if we just store the original checksum fields and make the comparison when retransmission occurs, the store cost will be reduced to a more allowable level than the second method. Taking the case mentioned in the second method for example, there are only about 2.73MB cache per second.

According to the above analysis, we can see that RSTEG is not stealthy enough. To make RSTEG be able to resist the checksum comparison detection method, the improvement should be made. It will be discussed in the following section.

## III. IMPROVED RSTEG AND ITS DETECTION

### *A. Improved RSTEG*

The communication framework of IRSTEG is shown in Fig.2. IRSTEG mainly contains three parts: Embedder, Compenastor and Filter. Embedder is used to embed covert messages into the normal traffic, and Compensator is responsible for compensating the TCP checksum field. Filter is used by the receiver to identify the packets that carry covert messages.

The communication process of IRSTEG is mainly consisted of four steps:

Step1: Embedder replaces the packet payload by covert message bits.

Step2: Compenastor to make the checksum field of the modified packet same with that of the original un-modified packet by the compensation algorithm described in Section 3.2.

Step3: The modified and compensated packet is sent to the receiver.

Step4: The receiver uses Filter to identify the covert packets and extract the covert message from the payload field.

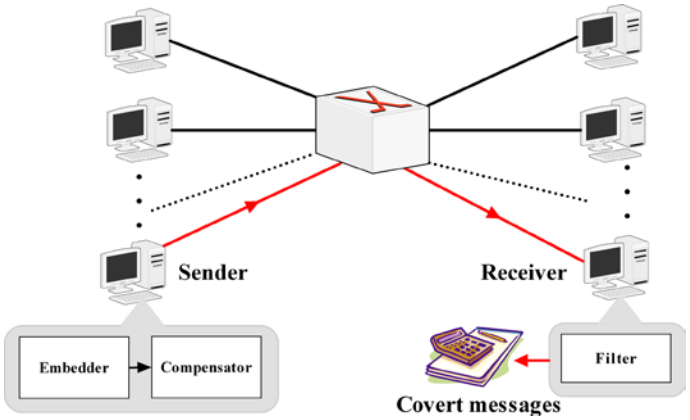



Figure 2 The framework of IRSTEG

| 4B | 4B | 1B | 1B | 2B |
|---|---|---|---|---|
| Source IP address | Destination IP address | 0 | Type | Length |

Figure 3 The format of TCP pseudo-header

*B.Compensation Algorithm*

To begin with the discussion of the compensation algorithm, a brief introduction about TCP checksum calculation rule is given which is defined by RFC793 document in detailed. The checksum field is a 16 bit one's complement of the one's complement sum of all 16 bit words in the header and text. If a segment contains an odd number of header and text octets to be checksummed, the last octet is padded on the right with zeros to form a 16 bit word for checksum purposes. The padding is not transmitted as part of the segment. The pseudo-header is only used in calculating checksum value, whose format is shown in Fig.3.

Theorem1: For two TCP packets with the same header except the checksum field, despite their payload fields are different, the TCP checksum also can be made the same by at most 16 bit word compensation in payload field.

Prove: Suppose $\Sigma f$ indicates the sum of all elements in the set $f$ of 16 bit words. As the sum may exceed the capacity of a 16 bit word, we will express this sum as in terms of two 16 bit words, $a$ and $b$. First, the sum of all 16 bit words in $f$ is expressed in the form

$$\Sigma f = 2^{16}a + b \tag{1}$$

the notation ~f means to get the compliment of the integer value f with trimming to a maximum width of 16 bits. As the TCP checksum mentioned above, it can be expressed as:

$$cks = \sim (a + b) \tag{2}$$

Assume the payload field of one TCP packet $p$ is $x$, and the payload field of packet q is $y$. $q$ is the retransmitted packet of $p$. $T_p$ and $W_p$ represent the TCP header and its pseudo-header of packet $p$ respectively. As packet $q$ is the retransmitted packet of $p$, the TCP header and pseudo-header of $q$ are the same with packet $p$. Suppose that the checksum value of the original packet is cks and the compensation 16-bit word is u. There exists an integer value v which satisfies $v = \Sigma u$ ( $0 \le v \le 65535$ ). The original and retransmitted packet are called $p$ and $q$ respectively. And suppose that the covert message is m, there is a 32 bit word w which satisfies $w = \Sigma m$. The remaining payload field of the retransmission packet is $y$. According to Eq. (1),

$$w + v + \Sigma y + \Sigma T_p + \Sigma w_p = 2^{16}a + b \tag{3}$$

According to Eq.(3),

$$v = 2^{16}a + b - (w + \Sigma y + \Sigma T_p + \Sigma w_p) \tag{4}$$

As the parameters in Eq. (4) are all known, the integer compensation value can be got. When $2^{16}a + b$ is less than $w + \Sigma y + \Sigma T_p + \Sigma w_p$, the value of $a, b$ should be recalculated to satisfy $2^{16}a + b > (w + \Sigma y + \Sigma T_p + \Sigma w_p)$ and Eq.(2).

An example is given to illustrate the above compensation algorithm. Frame 0030(Fig.4.a) is a normal packet in network, and Frame 0031(Fig.4.b) is its retransmission packet. The bold italic characteristics in Frame 0030 are replaced by covert messages which are shown in Frame 0031. For getting the compensation 16 bit word, the sum of all elements in Frame 0030 of 16 bit words is calculated first, which is shown in Eq.(5)

$$\begin{aligned} \Sigma T_{0030} + \Sigma w_{0030} &= 39F78 + 1DD28 \\ &= 57CA0 \end{aligned} \tag{5}$$

The sum of covert messages shown in Frame 0031 of 16 bit words is shown in Eq.(6)

$$\begin{aligned} w &= \Sigma m \\ &= 104+608+12+5C9+46E+20 \\ &= 1175 \end{aligned} \tag{6}$$

The sum of Frame0031 of 16 bit words can be expressed by

$$\Sigma f_{0031} = 3A730 + 1DD28 + 1175 = 595CD \quad (7)$$

Thus, according to Eq.(4), the compensation number is

$$v=65535*6+36202-361560-4469 = 63383 \quad (8)$$

The hex format of 63383 is F797, which is the compensation 16 bit word. The blue characteristics in Frame 0030 and 0031 are the checksum value. It can be seen that the checksum values of the original and retransmission packet are same after the compensation

Frame 0030 (original packet)

0000|| 01 00 01 00 00 00 3A CE 20 00 01 00 08 00 45 00 ||
0010|| 00 34 00 00 40 00 35 06 68 C1 3D 93 7A 60 CA 77||
0020|| 5A 98 00 50 04 22 D5 F1 C6 95 FD 14 6A 8A 80 12||
0030|| 16 D0 72 95 00 00 02 04 05 B4 ***01 01 04 02 01 03*** ||
0040|| 03 07

a. The original packet

Frame 0031 (retransmission packet)
0000|| 01 00 01 00 00 00 3A CE 20 00 01 00 08 00 45 00||
0010|| 00 34 00 00 40 00 35 06 68 C1 3D 93 7A 60 CA 77||
0020|| 5A 98 00 50 04 22 D5 F1 C6 95 FD 14 6A 8A 80 12||
0030|| 16 D0 72 95 00 00 02 04 05 B4 ***01 04 06 08 00 12***||
0040||*F7 97*

b. The retransmitted packet

Figure 4 An illustration of compensation algorithm

### *C. Detecting IRSTEG by Payload Segment Comparison*

It is discussed above that there are three detection algorithms against RSTEG. Through the simple analysis, it is not hard to find that the first and third methods are not valid for IRSTEG. The second method does still work without considering the huge storing demand of the detector. To reduce the storing demand, we propose to make the detection by comparing some randomly-selected payload segments. For example, for a TCP packet, only ten Bytes from its payload field are selected from the pseudo-random position which is determined by a pseudo-random number generator(PRNG). According to the instinctive consideration, the chosen ten Bytes will be different with a very high probability. Despite that the proposed method can't guarantee 100% correction rate, it also can give a reliable detection results with the demand of cache storing greatly less than that of the original second method.

## IV. Experimental Results and Analysis

### *A. Environmental Environment*

The experiments are carried out in two different network conditions. One is China Education and Research Network (CERNET) and the other is China Telecom Network (CTN). CERNET is constructed to connect most of universities and institutes in P. R. China and CTN provides the public Internet access service as the largest commercial ISP in P. R. China. The IRSTEG receiver is laid in CERNET of Nanjing University of Science and Technology with an independent IPv4 address. The IRSTEG sender is laid in CERNET and CTN network respectively. For shorting and distinguishing, the two network conditions are called CE-CE and CT-CE respectively. The hardware condition of the sender and receiver are same, which is shown in Table I. The operating system of the sender and receiver are Windows XP sp3 and Windows Server 2003 respectively. The sending and receiving software are both realized by Visual C++.

TABLE I
HARDWARE CONDITIONS

| Hardware device | Configuration information |
|---|---|
| CPU | P4 2.8GHz |
| Memory | 512M |
| Network Interface Card | 10/100M adapter card |

The overt flows are generated by uploading files to a FTP server which is installed in the receiver computer. The uploading speed is set to 200KB/s and the whole payload length is set to 1460B (reserve 2 Bytes for making compensation). In our experiments, the retransmission probability is set to 5% artificially. In Table II, five time experiments are listed and the average transfer speeds are shown similar between CE-CE and CT-CE. It is because the uploading speed of the two network conditions is same.

### *B. Detection Results*

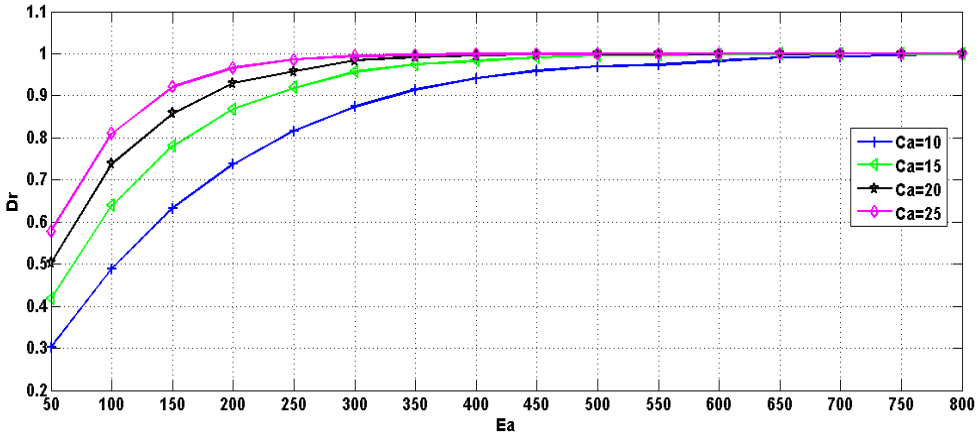
Assume that the correct detection rate as DR. It depends on the comparison amount (CA) and embedding amount (EA) of IRSTEG. Here, CA is the selected byte amount. EA is the byte amount that the sender embeds. The detection algorithm base on the payload segment comparison is tested by total 10,000 times with different EA and CA. The average DR is calculated. Fig.5 shows the detection results under different CA and EA. From Fig.5, it can be seen that with the increase of EA and CA, the detection result becomes more and more reliable.

From Fig.5, it can be seen that when CA is kept invariable and EA increases, the DR will also increase

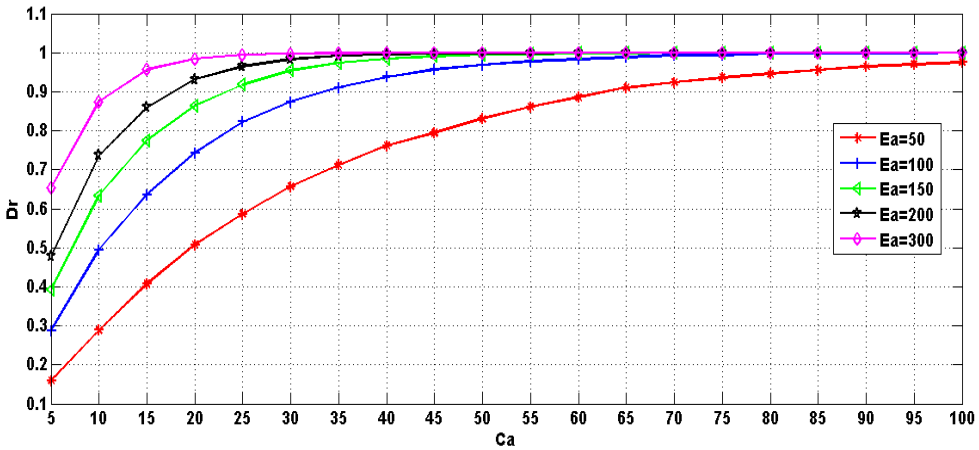
along with EA. When CA equals to 25B and EA equals to 200B, DR is as high as 98.3%. The more detailed experimental results are listed in Table III. It can be seen that when EA is larger than 150B, the detection method can achieve pretty results with less CA. At the situation, IRSTEG is not stealthy so far. Thus, if we want to keep stealthy of IRSTEG, EA should keep in a suitable low level. According to the results in Table III, when CA is set to 40B, EA should be set to less than 35B to guarantee that DR is less than 60%. It explains that despite the proposed method can detect IRSTEG, the sender can still avoid being detected by lowering down the covert communication speed.

TABLE II
EXPERIMENTAL RESULTS OF COVERT MESSAGES TRANSFER

| Experimental Times | Network Condition | Speed(B/s) |
|---|---|---|
| 1st | CE-CE | 564.5 |
| | CT-CE | 554.67 |
| 2nd | CE-CE | 581.7 |
| | CT-CE | 567.1 |
| 3rd | CE-CE | 562.9 |
| | CT-CE | 563.6 |
| 4th | CE-CE | 573.1 |
| | CT-CE | 572.6 |
| 5th | CE-CE | 579.3 |
| | CT-CE | 578.4 |




(a) The detection rate when EA keeps invariable and CA increases




(b) The detection rate when CA keeps invariable and EA increase

Figure 5 The detection rate with different CA and EA

TABLE III
EXPERIMENTAL RESULTS OF THE PAYLOAD-SEGMENT-COMPARISON-BASED METHOD

| EA | CA | DR |
|---|---|---|
| | 20 | 13.2% |
| 10 | 40 | 24.6% |
| | 60 | 34.4% |
| | 20 | 23.5% |
| | 40 | 39.3% |
| 20 | 60 | 53% |
| | 20 | 31.8% |
| 30 | 40 | 54.8% |
| | 60 | 69.3% |
| | 20 | 31.8% |
| 40 | 30 | 54.8% |
| | 40 | 67.8% |
| | 20 | 50% |
| 50 | 30 | 64.9% |
| | 40 | 75.5% |
| | 10 | 48.8% |
| 100 | 20 | 73.8% |
| | 30 | 86.7% |
| | 10 | 63.3% |
| 150 | 20 | 85.8% |
| | 30 | 95.2% |
| | 10 | 73.8% |
| 200 | 20 | 93% |
| | 30 | 98.6% |

## V. CONCLUSIONS

Different from the network steganographic methods which use the packet fields, RSTEG make the covert communication by replacing the retransmitted packet payload. It make RSTEG has very high covert communication rate. In this paper, we give three methods to make detection of RSTEG. In them, the detection method based on the checksum comparison is must efficient. To resist it, an improved RSTEG is presented by the payload compensation. Further, we give a detection scheme based on payload segment comparison which is reformed from the direct payload comparison. The experimental results show that the proposed method can make reliable detection against IRSTEG and the sender have to lower the covert communication speed to enhance its stealthy.

## ACKNOWLEDGMENT

This study was supported by the NSF of China (Grantno.61170250, 61103201), and NSF of Jiangsu province (Grantno.BK2010484).

## REFERENCES

[1] S. Zandel, G. Armitage, P. Branch. A survey of covert channels and countermeasures in computer network. IEEE Communications. Surveys and Tutorial. 2007, 9(3), pp.44-57.

[2] G. Girling. Covert channels in Lan's. IEEE Transactions on Software Engineering. 1987, 13(2), pp.292-296.

[3] C. H. Rowland. Covert channels in the TCP/IP protocol suite. First Monday, Peer Reviewed Journal on the Internet, July 1997.

[4] K. Ahsan, D. Kundur. Practical data hiding in TCP/IP. In proceedings of Texas workshop on security of information systems, 2003, pp. 18-22.

[5] A. Hintz. Covert channels in TCP and IP headers. URL: http://www.defcon.org/images//dc10-hintz-covert.ppt. 2003.

[6] L. Ji, W. Jiang, B. Dai, X. Niu，A novel covert channel based on length of message. In proceedings of 2009 International symposium on information engineering and electronic commerce. 2009, pp. 445-450.

[7] K. Szczypiorski. A performance analysis of HICCUPS-s steganographic system for WLAN. Telecommunication systems modeling, analysis, design and management. 2009, vol.49, pp. 3-4.

[8] L. Yao, X. Zi, L. Pan and J. Li. A study of on/off timing channel based on packet delay distribution. Computers & Security, 2009,28(8), pp.785-794.

[9] S. Gianvecchio, H. Wang, D. Wijesekera. Model-based covert timing channels: automated modeling and evasion. Lecture Notes In Computer Science, 2008, vol.5230, pp.211-230.

[10] G. Liu, J. Zhai, Y. Dai, Z. Wang, Network Covert Timing Channel with Distribution Matching. Telecommunication System, 2012, 49(2), pp.199-205.

[11] W. Mazurczyk, M. Smolarczyk, K. Szczypiorski. Retransmission steganography and its detection. Soft Computing. 2009, 15(3), pp. 505-515.
[12] S. Rewaskar, J. Kaur, F. Smith. A Performance Study of Loss Detection/Recovery in Real-world TCP Implementation. In Proceedings of the IEEE International Conference on Network Protocols, ICNP 2007, 2007, pp. 256-265.
[13] Internet Traffic Report. URL:http://www.internettrafficreport.com/30day.htm.
[14] Chen,C., Mangrulkar, M.,Ramos, N,. and Sarkar, M. Trends in TCP/IP Retransmissions and Resets, Technical Report, URL:http://www-cse.ucse.edu/classes/wi01/cse222/projects/reports/tcp-flags-13.pdf.

**Mr. Jiangtao Zhai** received the M.E. from Nanjing University of Science and Technology in 2008. He now is candidate for PHD. with Nanjing University of Science and Technology. His research field is the network security including network steganography, covert channel detection and elimination.

**Dr. Guangjie Liu** received the B.E. and M.E. from Nanjing University of Science and Technology in 1998, 2002 respectively. He now is an associate professor with Automation School of Nanjing University of Science and Technology. He has contributed more than 40 refereed papers covering topics of network and information security.

**Dr. Yuewei Dai** received the M. E., and Dr. Eng. degrees from Nanjing University of Science and Technology in 1987 and 2002 respectively. He now is a professor with Automation School of Nanjing University of Science and Technology. He has contributed more than 100 refereed papers including automation control, information technology and information security. He is also the director of System Engineering Academy of Jiangsu Province in China.

# Cross-Layer Dual Domain Scheduler for 3GPP-Long Term Evolution

Wei Kuang Lai
National Sun Yat-sen University, Kaohsiung, Taiwan
Email: wklai@cse.nsysu.edu.tw

Kai-Ting Yang
National Sun Yat-sen University, Kaohsiung, Taiwan
Email: terence.kaiting@gmail.com

***Abstract*—The bandwidths of wireless communication technologies have grown fast in recent years. Thus, more users access wireless networks and experience various network services in their daily life. As plenty users and various services coexist in networks, many strategies will deeply affect system performances. Scheduling is absolutely one of these important strategies. A good scheduling strategy can significantly improve network system performances in terms of throughput, delay time and so on. From users' point of view, an elaborate scheduling strategy can also greatly enhance the quality of network services so that users can enjoy better network services. At the critical moment that wireless technologies will stride toward next-generation, to design good scheduling schemes for LTE is necessary and urgent. Therefore, in this paper, aiming at LTE, we propose an efficient scheduling scheme to improve the performance of LTE. We take into account the real-time multimedia transmission, which has stricter QoS requirements to design our scheduling schemes. Overall system throughput and fairness are also considered in this paper. The scheduler will be devised according to the characteristics of downlink/uplink transmissions in LTE. The goals of the proposed scheduling scheme are (1) avoiding service disruptions (2) keeping fairness (3) improving system throughput (4) reducing delay time.**

***Index Terms*—long term evolution, scheduler, system throughput, delay, real time, fairness.**

## I. II. INTRODUCTION

With the development of wireless technologies, more and more users are used to access Internet via wireless communications no matter where they are. Adopting their personal devices such as smart phones and laptop computers, users can share information or demand data without the tangle of cables. The convenience of wireless networks motivates the rise of network services and then increases the requirement of network capacities. Improving network capacities in terms of data rate, transmission delay, is always one of major challenges for researchers in the fields of telecommunication and networking. For this reason, 3GPP has devoted to design a novel and efficient protocol, named Long Term Evolution (LTE), which is regarded as one of the probable candidates for 4G. The main targets of LTE include peak data rate higher than 100Mbps, latencies less than 10 ms in user plane and 100 ms in control plane, etc. Table I summarizes the main performance requirements of LTE.

Technical advance and maturity in LTE has claimed the dawn of a new generation of communication networks. With enlarged access range and broadened bandwidth, users are capable of keeping connected anywhere and anytime, and enjoying high quality service on the move. Such a vision will be confronted by a number of technical challenges before fully fulfilled.

Despite the improvement of LTE in network capacities, it is still difficult to guarantee the qualities of services under the environments with large number of subscribers. Therefore, to efficiently make use of resources defined in LTE remains a good strategy to gain better network performance. In this paper, we propose a novel scheduling algorithm for LTE to provide better network services. The reasons why we aim at LTE to design the corresponding scheduling algorithm include:

- LTE has promised to provide users with considerable data rates. A deficient scheduling algorithm would lead to more serious resource wastage in the network with better capacities.
- As showing in Figure 1, both frequency domain and time domain division can be achieved in LTE, which means that the resource utilization in LTE is more flexible than in other previous protocols.

These two reasons motivate us to design a well scheduling algorithm to make good use of the benefits of LTE.

Throughput and fairness are two indexes usually adopted to evaluate a scheduling algorithm. However, the tradeoff between system throughput and user fairness is a well-known challenge and has been studied for a long time. A scheduling algorithm aiming at maximizing the system throughput may lead to starvation of users with lower received signal strength. A fair scheduler may spend more resources to take care of users at the coverage border of the eNodeB and may diminish the bandwidth efficiency.

In this paper, the scheduling algorithm is proposed to balance the tradeoff between system throughput and fairness and to maintain better qualities for services with time constraints. Taking the characteristics of LTE into account and benefiting from its flexibility, the proposed algorithm can assign a resource block to a suitable user so as to maintain better system throughput. On the other hand,


doi: 10.4304/jnw.8.1.189-196

for the users with packets tend to expire, the proposed algorithm can assign additional resource blocks which lead to less impact on the system throughput to them so that the packets may be transmitted on time. In addition, we design the scheduling algorithm in a cross-layer manner to cope with the rapid variation of channel state. The cross-layer cooperation scheduler takes into account both states of eNodeB and UE, and involves the considerations of CQI [1][2], queue length, bit error rate, priority, and so on, to catch the immediate states of both sides.

TABLE I.
PERFORMANCE REQUIREMENTS OF LTE

| *Index* | *Requirement* |
|---|---|
| Peak Data Rate | Downlink: 100 Mbit/s |
| | Uplink: 50 Mbit/s |
| Spectrum | Downlink: 20 MHz |
| | Downlink: 20 MHz |
| Cell Size | 30 km with reasonable performance |
| Cell Capacity | 200 active users |
| Latency | User plane: < 5 ms |
| | Control plane: < 50 ms |

Therefore, the proposed scheme can well response to the changes of network states and can maintain better performance.

The remainder of this paper is organized as follows. In Section 2, we introduce some representative works related to our study. In Section 3, we detail the methodology of the proposed scheduling algorithm. The performance of the proposed algorithm is evaluated by simulations, and the results are given in Section 4. Finally, we draw some conclusions of our findings in Section 5.

## III. RELATED WORKS

Many works related to scheduling have been proposed. The objectives of these works are usually quite different. A scheduling algorithm may be designed to improve the system throughput, shorten the transmission delay, or maintain the fairness among users. However, these objectives are often in conflict with the others. It is difficult to achieve two or more objectives in a single algorithm.

Some works consider the delay as the major concern of the scheduling algorithm such as the largest weighted delay first algorithm (LWDF) [3], delay threshold priority queuing algorithm (DTPQ) [4], and earliest deadline first (EDF) [5]. Trying to shorten the average delay in the network, these algorithms adopt the delays to determine the priorities of users. The longer delay a user has suffered, the higher priority the user receives. Therefore, real-time services can benefit from this kind of algorithms. There are some works trying to treat each user fairly. The fairness is the major consideration of these works [6]-[14]. Some of these works [11]-[14] are designed for LTE. However, each UE would have different channel quality according to its position and mobility pattern. When fading, interference, and distance are taken into account, a scheduling algorithm is difficult to let the users have equal performance in wireless networks. Therefore, these works usually try to maintain the proportional fairness among users.

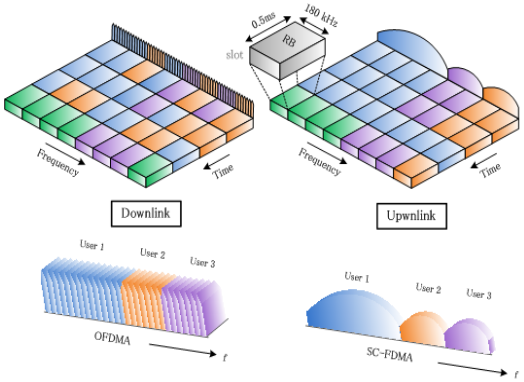



Figure 1. The time division and frequency division in LTE.

Aiming at the overall system throughput, some works [15]-[18] give higher priorities to the users whose channel qualities are better than others. Therefore, the resources would be occupied by the users near to the eNodeB. Users far away from eNodeB may starve due to their poor signal-to-noise ratios. The philosophy of these works is completely different from the works aiming at user fairness.

Some works [19][20] apply the concept of cross-layer cooperation to improve the performance of LTE. Authors of [19] design a cross-layer approach which combines the functionality of MAC layer and the physical layer to reduce the power consumption of LTE. The work in [20] focuses on improving the service quality of video streams in LTE and proposes a cross-layer scheme. In this scheme, according to the channel quality, the application layer is responsible for coding video frames and the physical layer is responsible for determining the modulation and coding scheme.

## IV. CROSS-LAYER DUAL DOMAINS SCHEDULER

Aiming at 3GPP Long Term Evolution, we propose a novel scheduling algorithm, named cross-layer dual domain scheduler (CDDS), to pursue better performance in LTE. From the point of view in application layer, we adopt scalable video coding (SVC) as the example in the following description. SVC divides a video frame into one based layer and multiple enhanced layers. The based layer includes the most important information of the original frame and must be received by a client for decoding the frame. An enhanced layer may contain more detailed information than a based layer but is not indispensable. The SVC can determine how many enhanced layers should be transmitted in light of the network states. We would like to note that the proposed algorithm is not limited to the traffics of SVC. Common traffics also can benefit from CDDS to gain better performance.

In the proposed CDDS, the statuses of the physical layer and the data-link layer are taken into account to improve the system performance. In addition, not only the statuses at the eNodeB but also the statuses at the user equipment are considered in CDDS. As showing in Figure 2, the physical layer in CDDS is responsible to monitor

and information than a based layer but is not essential. The SVC can determine how many enhanced layers should be transmitted in light of the network states. We would like to note that the proposed algorithm is not limited to the traffics of SVC. Common traffics also can benefit from CDDS to gain better performance.

In the proposed CDDS, the statuses of the physical layer and the data-link layer are taken into account to improve the system performance. In addition, not only the statuses at the eNodeB but also the statuses at the user equipment are considered in CDDS. As showing in Figure 2, the physical layer in CDDS is responsible to monitor andreport the channel conditions. A user equipment estimates the received signal strengths and the noise levels by listening the cell-specific downlink reference signals from its attached eNodeB. After estimating the channel statuses, the user equipment presents these information as a Channel Quality Indicator and reports to its attached eNodeB. Besides reporting the channel conditions, the user equipment also reports some information of the data-link layer to the eNodeB such as propagation delays and its queue length. According to the

propagation delays and the queue length, the eNodeB can determine whether a packet will be overdue or not. Note that the queue lengths, for some services such as video streaming, present how many packets can be consumed by the applications. In other words, the lifetime of the packet in the front of the queue at the eNodeB is determined by the queue length at the UE. The packet loss rates, propagation delays and queue lengths can be reported to the eNodeB by way of the uplink L1/L2 control signal specified in LTE standard.

### *A. Considerations of CDDS*

In CDDS, we utilize following information to capture the network states: channel quality between a UE and the attached eNodeB, the loading of the eNodeB, propagation delay and bit error rate, packet loss rate, the queueing delay in eNodeB, traffic priority and packet lifetime, and the queue length in UE.

Now we detail how to use these information to adapt the proposed CDDS.

- Channel quality between a UE and the attached eNodeB: Channel quality is one of major factor directly affecting the network performance and can be estimated in the form of signal-to-noise ratio (SNR). 3GPP also defines a signal, named

Channel Quality Indicator (CQI) to periodically measure the channel quality between a UE and its attached eNodeB. CDDS can benefit from the CQI signaling to let eNodeB know the channel qualities for all UEs and then adapt the scheduler to take care of users and possibly maximize the overall system throughput. For a UE, the eNodeB can estimate the CQI in each resource block to see which resource block is more suitable for the UE. Furthermore, thanks to the adaptive modulation and coding (AMC), a good function in LTE, eNodeB can determine the most suitable modulation and coding schemes for UEs and can evaluate the achievable maximum throughput of each resource block. Therefore, CDDS then can arrange resource blocks to maximum the system throughput as possible.

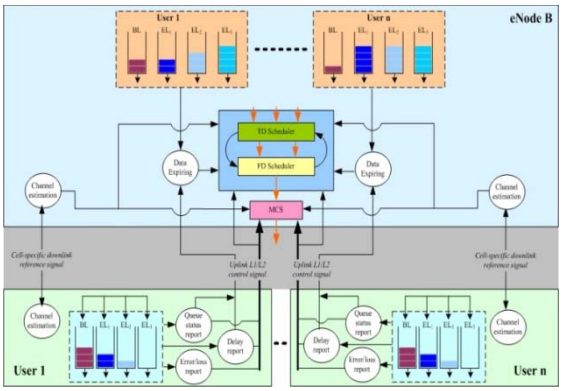



Figure 2. The signaling of the proposed cross-layer downlink scheduler.

- The loading of the eNodeB: The transmission opportunity, named resource block in LTE, is defined by both time and the frequency. A resource block occupies 180kHz bandwidth and lasts 0.5ms. An eNodeB can assign one or more resource blocks to each UE in each round of scheduling. If the total demanded resources of UEs do not exceed the capacity of the eNodeB, the scheduler can assign just enough resource blocks to UEs which report lower CQIs and give more resource blocks to UEs with better channel qualities to enhance the overall system throughput. An eNodeB can know its loading in terms of the number of connections and the requirements of UEs.
- Propagation delay and bit error rate: The bit error rates are measured at UE side and reported to eNodeB. An eNodeB can adjust the modulation and coding scheme for a UE when the estimated channel quality is inconsistent with the reported error rate. Furthermore, to predict whether a packet can be successfully transmitted during its lifetime, eNodeB should gather the propagation delay of each UE. The propagation delay can be measured when the eNodeB receives an acknowledgement from the UE.
- Packet loss rate: A packet loss, different from a bit error, means that a node receives a packet which contains too many error bits to reconstruct by the FEC function or never receive a certain packet. It may be caused by the poor channel qualities or a collision. An eNodeB can calculate the packet loss rate according to the number of retransmission requests from a UE. The packet loss rate presents not only the channel statuses but also the collision problem. The bit error rate only reflects the channel quality when the quality is not so poor that a station cannot receive any bit of a packet. The eNodeB can adjust the modulation and coding scheme or assign different resource blocks to a UE according to its packet loss rate.
- Queueing delay in eNodeB: A packet would suffer a serious delay when the load of the eNodeB is heavy. In the situation, many packets

are buffered in the queue of the eNodeB and wait to be transmitted. The waiting time, referred to as queueing delay, is usually the major latency during transmission when the eNodeB is busy. CDDS estimates the queueing delay for each packet to see whether it may expire or not and then give it an applicable precedence to be transmitted.

- Traffic priority and packet lifetime: Each traffic type has the congenital priority. The higher priority the traffic is, the more transmission opportunities it receives. However, the traffic with higher priority usually means that its packet lifetime is shorter and need to be treated more quickly. For example, some real-time services, such as VoIP and video conferences, have strict delay constraints.
- Queue length in UE: Differing from real-time services, some service, such as video-on-demand (VOD), do not have primitive time constraints. The delay constraints of the packets belonging to such services are determined by the time when the packet will be consumed at the UE. For example, a packet containing the based layer information must be received before the program of the UE tries to decode and play it. If there are still many based layers in the buffer of UE, the eNodeB can transmit the next based layer later. Suppose that $l_a$ stands for the number of based layers in UE's buffer and $\varepsilon$ is the amount of based layers the UE program needs to consume per second. The lifetime of the packet containing the next based layer information, $t_r$, can be as following:

$$t_r = \frac{l_q}{\varepsilon} \quad (1)$$

According (1), we can see that the lifetime of the next based layer information is determined by $l_q$. Therefore, for a streaming service, the queue length in UE is an essential information for determining the packet lifetime. For this reason, a UE is endowed with the duty to report its queue length to the eNodeB in CDDS. The signaling between UEs and the eNodeB is shown as Figure 2.

After obtaining related information mentioned above, the eNodeB can assign resource blocks to each user equipment according to the QoS requirements and current network statuses. As mentioned previously, a resource block in LTE is determined by the time and frequency domains, which is a 0.5ms transmission opportunity over 180-kHz bandwidth. LTE provides flexible arrangement of the resource blocks to meet different QoS requirements and to adapt to rapidly changed network conditions. An eNodeB assigns at least two resource blocks to a user equipment, which is named resource block pair in LTE. In addition, as showing in Figure 3, an eNodeB is allowed to arrange discontinuous resource blocks in both time and frequency domains to a user so that the scheduler of LTE is very elastic. Due to the flexibility of the resource assignment in LTE, a well-designed scheduler can gain better performances in terms of throughputs, delays, and error rates.

In this paper, both the frequency domain and the time domain are considered to design our scheduling algorithm. However, the considerations in the time domain and in the frequency domain are quite different. Therefore, the scheduler of an eNodeB is separated into the time domain (TD) scheduler and the frequency domain (FD) scheduler in CDDS. In the time domain scheduler, the main concern is the packet lifetime. An urgent packet can obtain resource blocks which are advanced in the time domain. A packet with longer lifetime may be assigned later resource blocks in the time domain or even be scheduled in the next round. Differing from the time domain scheduler, the frequency domain scheduler takes into account the channel conditions of the user equipment such as the received signal strength and the noise strength to assign suitable bands to a UE so as to satisfy the QoS requirements of the user equipment and maximize its throughput.

However, the decision order of the time domain scheduler and the frequency domain scheduler may influence the system performance. If the time domain scheduler makes the decision before the frequency domain scheduler, the time constraints of packets can be satisfied.

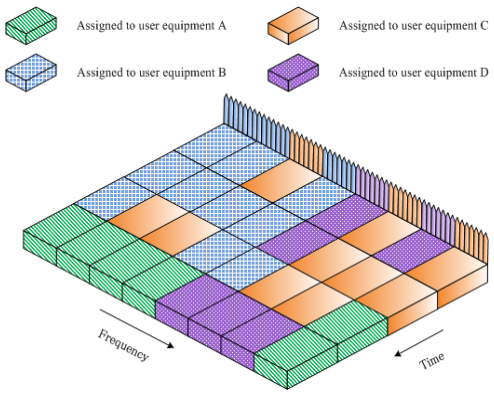



Figure 3. Discontinuous resource assignment in LTE.

In this situation, however, stations with urgent packets would occupy resource blocks which are more suitable to other stations. For example, a station with an urgent packet obtains a resource block, but it is far away from the eNodeB so that the throughput of the resource block is decreased. In the worst case, the transmission may fail due to the weak signal strength. If the frequency domain scheduler makes the decision before the time domain scheduler, each resource block can be assigned to the most suitable stations in terms of channel conditions. The system throughputs can be maximized and the packet loss rate and bit error rate can be reduced. However, the time constraints of real-time services may not be satisfied. Aim at this trade-off, we design a novel scheduling algorithm to satisfy the time constraints of packets and to maximize the system throughput as possible. The details of the proposed algorithm are present in following section.

### B. Scheduling Algorithm Design

Suppose that the signal-to-noise ratio $\gamma$ of the resource block $j$ for the UE $i$ at the observing time $t$ can be expressed as following:

$$\gamma_{i,j}[t] = \frac{P \cdot G_{i,j}[t]}{k \times (I + N_o)} \quad (2)$$

where $P$ is the transmission power of the eNodeB, $G$ stands for the power gain, $k$ is the total number of resource blocks, $I$ and $N_o$ are the interference and noise respectively. At first step, as showing in (3), we seek the UE $x$ which has the highest SNR for a certain resource block and temporarily assign the resource block to the UE to maximize the system throughput. By doing so, as showing in (4), the resource block $j$ is temporarily assigned to UE $x$ when $x$ satisfies (3).

$$\gamma_{x,j}[t] = \arg\max\{\gamma_{x,j}[t] \mid \rho_i = 0\} \quad (3)$$

$$c_j = x \quad (4)$$

In (4), $c_j$ presents the identity of the UE having the highest SNR for resourcr block $j$. So we can obtain the set $C$:

$$C = \{c_1, c_2 \ldots c_k\} \quad (5)$$

Secondly, the eNodeB checks the lifetimes of packets in its buffer. If there is no packet tends to expire, or all UEs with packets tend to expire are listed in set $C$, CDDS will assign resource blocks to UEs according to (5). However, if any UE with packets about to expire doesn't appear in set $C$, CDDS would seek one or more suitable resource blocks for it. Suppose that $L$ is the set of UEs with packets tending to expire, CDDS tries to seek resource blocks for the UEs in the set $Y$:

$$Y = L \cap \bar{C} \quad (6)$$

In this case, as shown in (7), CDDS finds out the resource blocks which are not assigned to the UEs in $L$ and cause the least loss in term of system throughput for the n-th UE in $Y$. Finally, as showing in (8), CDDS assigns this resource block to the n-th UE in $Y$.

$$\delta_j = \arg\min\left\{\delta = \gamma_{c_j,j} - \gamma_{y_n,j}, c_j \in \bar{L}\right\} \quad (7)$$

$$c_j = y_n \quad (8)$$

By doing so, CDDS can take care of the packets with delay constrains and can increase the system throughput as possible.

The methodology of the proposed CDDS can be implemented not only in downlink transmission but also in uplink transmission. The Figure 4 shows the signals when CDDS is adopted in uplink transmission.

As showing in Figure 4, UEs should report their statuses to their attached eNodeB so that the eNodeB can recalculate the scheduling algorithm for them. After listening the cell-specific downlink reference signals from its attached eNodeB, a UE estimates the received signal and noise strengths, and then reports these information to its attached eNodeB by sending a Channel Quality Indicator. Other information, such as propagation delays, the queue length, even the energy level, can be reported to the attached eNodeB by way of the uplink L1/L2 control signal.

## V. Simulation Results

In order to validate the performance of the proposed scheduling algorithm, we introduce some simulation results in this section. Because CDDS is designed to balance the tradeoff between the system throughput and the fairness, we compare the performance of CDDS with the maximum throughput and proportional fair algorithm, the typical algorithms concern for the system throughput and the fairness respectively.

Figure 5 shows the scenario of our simulations. Following 3GPP release 9, the instantaneous downlink peak rate is 100 Mbps in a 20MHz spectrum. The number of UEs attaching to the eNodeB is from 10 to 100. Each UE is served by its corresponding server which locates in a wired domain. A Half of servers transmit video streams to their users. The applications at UEs start to consume the data 10 seconds after the beginning of the simulations. The other servers transmit voice data with 150ms time constraint to their users. Simulation parameters are summarized in Table II.

Figure 6 shows the overall throughput of the eNodeB. We can see that maximum throughput algorithm has the highest throughput. This phenomenon is more obvious when the number of UEs increases.

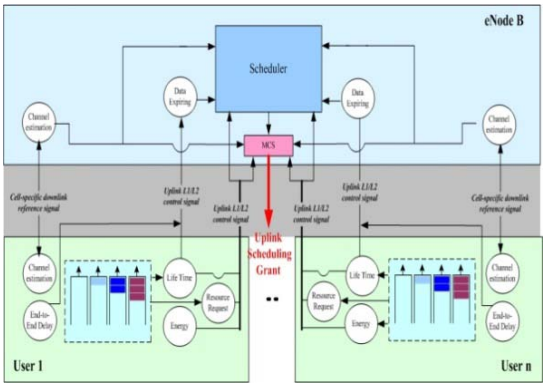



Figure 4. The signaling of the proposed cross-layer uplink scheduler.

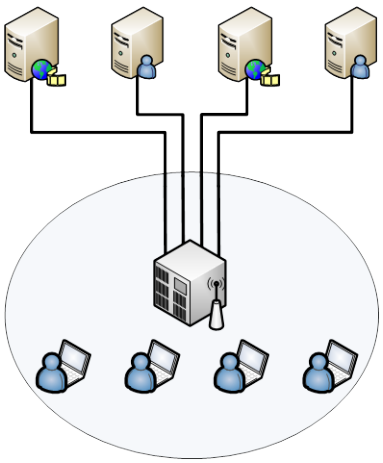


Figure 5. The scenario of simulations

TABLE II.
PARAMETERS IN THE SIMULATIONS

| Parameter | value |
|---|---|
| Number of UEs | 10~100 |
| Peak data rate | 100 Mbps |
| Bandwidth | 20 MHz |
| Coverage of the eNodeB | 2 km |
| Transmission Duration of a RB | 0.5 ms |
| Bandwidth of a RB | 180 KHz |
| Time Constraint of voice data | 150 ms |
| Packet arrival rate of voice stream | 10 packet/s |
| Packet arrival rate of video stream | 5 packet/s |
| Simulation time | 150 s |

When there are less UEs in the network, the total traffics may less than the system capacity and the overall throughput of the eNodeB is dominated by the total traffics. As the number of UEs increases, the system throughput tends to the peak data rate of the eNodeB. The maximum throughput algorithm may provide all resources to the UE with the best channel quality so that it has higher throughput than the others. In the other hand, the proportional fair algorithm may spend more resources to take care of UEs with lower channel qualities so that the overall throughput is quite less than CDDS. We can see that the throughput of CDDS is not much different with the maximum throughput algorithm.

Figure 7 shows the average delays of the three algorithms. Giving higher priorities to UEs with better channel qualities, maximum throughput algorithm may lead to the starvation problem and may result in serious delays at UEs with worse channel qualities. CDDS is designed with the consideration of packet lifetime so that ithas shorter average delays for packets with time constraints. Figure 8 shows the number of expired packets in each algorithm. Due to the starvation problem in maximum throughput algorithm, the number of expired packets is much larger than the other algorithms, especially when the number of UEs increases. Due to the consideration of packet lifetimes, CDDS has less expired packets than the other algorithms.

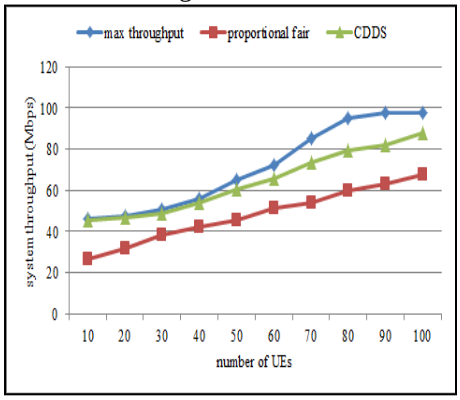



Figure 6. The relationship between the system throughput and the total number of the UEs.

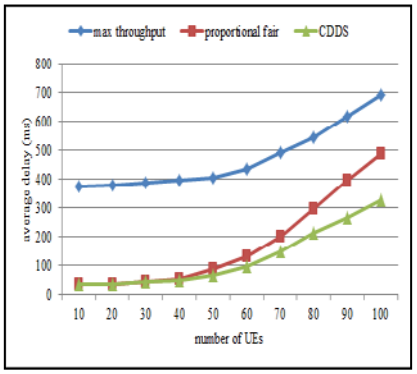



Figure 7. The relationship between the average delay and the total number of the UEs.

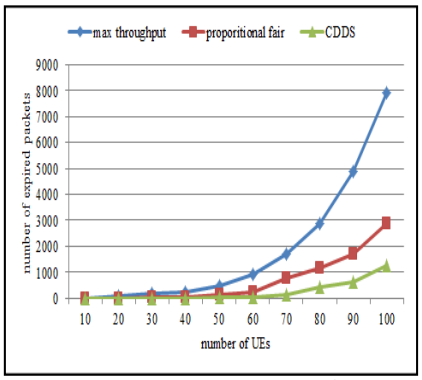



Figure 8. The relationship between the number of the expired packets and the total number of the UEs.

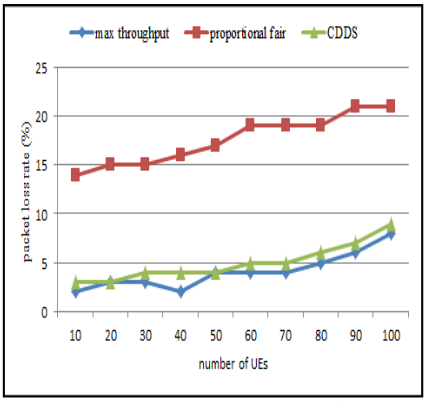



Figure 9. The relationship between the number of the packet loss rate and the total number of the UEs.

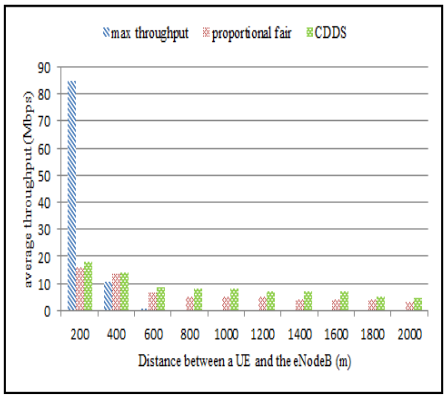



Figure 10. The average throughtput for different distances between a UE and the eNodeB.

Figure 9 shows the packet loss rates of three algorithms. Maximum throughput algorithm assigns resource blocks to the UEs which have higher signal-to-noise ratios so that the packet loss rate of maximum throughput algorithm is lower than other algorithms. Proportional fair algorithm gives more opportunities to the UEs which have lower signal-to-noise ratios to maintain the system fairness and results in higher packet loss rate. The proposed CDDS tries to balance the system throughput and the fairness and has a little higher packet loss rate than maximum throughput algorithm.

We turn our attention to the system fairness. In the simulation of the system fairness, the total number of UEs is 100. We arrange 10 UEs at the locations which are 200 meters away from the eNodeB, another 10 UEs at the locations which are 400 meters away from the eNodeB, and so on. The maximum distance between a user equipment and the eNodeB is 2000 meters. The average throughputs of UEs at each location are summarized in Figure 10. As showing in Figure 10, maximum throughput algorithm arranges almost all resource blocks to the UEs closed to the eNodeB and results in the starvation problem to the UEs far away from the eNodeB. In contrast to maximum throughput algorithm, the proposed CDDS algorithm and proportional fair algorithm are quite fairer. We can see that the UEs can obtain resource blocks even if the distances between UEs and the eNodeB are far. In other words, the proposed CDDS algorithm and proportional fair algorithm can take care the UEs with lower signal-to-noise ratios. Compared with proportional fair algorithm, we can see that the proposed CDDS algorithm has higher throughputs at each location. The reason is that the proposed CDDS algorithm only assigns resource blocks to the UEs with lower signal-to-noise ratios when their packets tend to expire. In addition, CDDS arranges the resource blocks which have less influence in the system throughput to the UEs with urgent packets. Therefore, the proposed CDDS has higher throughput at each location than proportional fair algorithm. As showing in Figure 6 and Figure 10, we can see that the proposed CDDS can maintain well system throughput and well system fairness.

The simulation results show that CDDS can have well performance in terms of overall throughput and transmission delay.

## VI. Conclusion

A novel scheduling algorithm for LTE is proposed in this paper to provide better network services. Taking the characteristics of LTE into account and benefiting from its flexibility, the proposed algorithm assigns each resource block to a suitable user so as to maintain better system throughput. Furthermore, for the users with packets tend to expire, the proposed algorithm would assign additional resource blocks which lead to less impact on the system throughput to them so that the packets may be transmitted on time. The proposed scheduling algorithm is designed in a cross-layer manner to cope with the rapid variation of channel state. It takes into account both states of eNodeB and UE, and involves CQI, queue length, bit error rate, priority, and so on, to catch the immediate states of both sides. Therefore, the proposed scheme can well response to the changes of network states and can maintain better performance.

## Acknowledgment

This work is partially supported by the National Science Council, Taiwan, under the grant No. NSC 100-2219-E-110-001, No. NSC 99-2221-E-110-068, and No. NSC 99-2221-E-110-065. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## References

[1] S. N. Donthi, N. B. Mehta, “Joint Performance Analysis of Channel Quality Indicator Feedback Schemes and Frequency-Domain Scheduling for LTE”, IEEE Transactions on Vehicular Technology, vol. 60, pp. 3096–3109, 2011.

[2] S. N. Donthi, N. B. Mehta, “An Accurate Model for EESM and its Application to Analysis of CQI Feedback Schemes and Scheduling in LTE”, IEEE Transactions on Wireless Communications, vol. 10, pp. 3436–3448, 2011.

[3] A. L. Stolyar and K. Ramanan, “Largest Weighted Delay First Scheduling: Large Deviations and Optimality”, *Annals of Applied Probability*, vol. 11, pp. 1–48, 2001.

[4] D. H. Kim and C. G. Kang, “Delay Threshold-based Priority Queueing Packet Scheduling for Integrated Services in Mobile Broadband Wireless Access System, “*IEEE International Conference of High Performance Computing and Communications*, pp. 305–314, 2005.

[5] H. Sariowan, R. L. Cruz, G. C. Polyzos, “SCED: a generalized scheduling policy for guaranteeing quality-of-service”, *IEEE/ACM Transactions on Networking*, vol. 7, iss. 5, pp. 669–684, Oct. 1999.

[6] J. Wu, J. Mo, and T. Wang, “A Method for Non-Real-Time Polling Service in IEEE 802.16 Wireless Access Networks”, *IEEE Vehicular Technology Conference*, pp. 1518–1522, 2007.

[7] H. Kim and Y. Han, “A proportional fair scheduling for multicarrier transmission systems”, *IEEE Communication Letter*, vol. 9, pp. 210–212, Mar. 2005.

[8] F. Hou, P. Ho, X. Shen, and A. Chen, “A Novel QoS Scheduling Scheme in IEEE 802.16 Networks”, IEEE Wireless Communication and Networking Conference, pp. 2457–2462, 2007.

[9] P. Bender, P. Black, M. Grob, R. Padovani, placeN. Sindhushayana, and A. Viterbi, “CDMA/HDR: A Bandwidth-Efficient High-Speed Wireless Data Service for Nomadic Users”, *IEEE Communication Magazine*, vol. 38, pp. 70–77, Jul. 2000.

[10] N. Ruangchaijatupon and Y. Ji, ’”Simple Proportional Fairness Scheduling for OFDMA Frame-Based Wireless Systems”, *IEEE Wireless Communication and Networking Conference*, pp. 1593–1597, 2008.

[11] M. Proebster, C. M. Mueller, H. Bakker, “Adaptive fairness control for a proportional fair LTE scheduler”, IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications, pp. 1504–1509, 2010,

[12] E. Yaacoub, Z. Dawy, “A Game Theoretical Formulation for Proportional Fairness in LTE Uplink Scheduling”, IEEE Wireless Communications and Networking Conference, pp. 1–5, 2009.

[13] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Young-Doo Kim, “Fairness-aware radio resource

management in downlink OFDMA cellular relay networks", IEEE Transactions on Wireless Communications, vol. 9, pp. 1628–1639, 2010.

[14] S. Ali, M. Zeeshan, "A Delay-Scheduler Coupled Game Theoretic Resource Allocation Scheme for LTE Networks", 2011 Frontiers of Information Technology, pp. 14–19, 2011.

[15] S. Shakkottai, R. Srikant, and A. Stolyar, "Pathwise Optimality and State Space Collapse for the Exponential Rule", 2002 *IEEE Int. Symp. Information Theory*, pp. 379, 2002.

[16] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas", *IEEE Transaction on Information Theory*, vol. 48, pp 1277–1294, Jun 2002.

[17] V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks", *IEEE Wireless Communication and Networking Conference*, vol. 2, pp. 984–990, 2006.

[18] S. Schwarz, C. Mehlfuhrer, M. Rupp, "Low complexity approximate maximum throughput scheduling for LTE", the Forty Fourth Asilomar Conference on Signals, Systems and Computers, pp. 1563–1569, 2010.

[19] R. Torrea-Duran, C. Desset, A. Dejonghe, "A cross-layer approach to save energy in 3GPP-LTE terminals", 2010 Future Network and Mobile Summit, pp 1–8, 2010.

[20] L. Haiyan, C. Song, W. Dalei, W. Jianjun, T. Hui, "Quality-driven cross-layer optimized video delivery over LTE", IEEE Communications Magazine, vol. 48, pp 102–109, 2010.

**Wei Kuang Lai** (S'88–M'96–SM'08) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1984 and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1992. In August 1992, he joined the faculty of the Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan, where he is currently a Professor. His research interests include high-speed and wireless networks.

**Kai-Ting Yang** received the M.S. degree from the National Sun Yat-Sen University, Kaohsiung, Taiwan, in 2006.
He is currently working toward the Ph.D. degree with the Department of Computer Science and engineering, National Sun Yat-Sen University, Kaohsiung, Taiwan His research interests include mobile computing, algorithm design, machine learning and 4G networks.

# Data Aggregation Scheme based on Compressed Sensing in Wireless Sensor Network

Guangsong Yang
School of Information Technology, Jimei University, Xiamen, China
Email: gsyang@jmu.edu.cn

Mingbo Xiao, Shuqin Zhang
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China
Email: mingbo@hdu.edu.cn

***Abstract*—Wireless sensor network (WSN) consisting of a large number of nodes, are usually deployed in a large region for environmental monitoring, security and surveillance. The data collected through high densely distributed WSN is immense. To improve measure accuracy and prolong network lifetime, reducing data traffic is needed. Compressive sensing (CS) is a novel approach to achieve much lower sampling rate for sparse signals. In order to reduce the number of data transmissions and save more energy, we apply CS theory to gather and reconstruct the sparse signals in energy-constrained large-scale WSN. Instead of sending full pair-wise measurement data to a sink, each sensor transmits only a small number of compressive measurements. The processes of CS aggregation in WSN are given, including sparse presentation of signal, observation matrix and reconstruction algorithm design. The relationship between observations and reconstruct MSE are also discussed. Simulation result shows that our scheme can recovery the unknown data with acceptable accuracy as well as reduce global scale cost.**

***Index Terms*—Compressed Sensing; Wireless Sensor Network; Aggregation**

## I. INTRODUCTION

The wireless sensor networks are consisted of hundreds to thousands of inexpensive wireless nodes, each with some computational power and sensing capability, operating in an unattended mode. They are intended for a broad range of environmental sensing applications from vehicle tracking to habitat monitoring [1]. The problem of efficiently transmitting or sharing information from and among a vast number of distributed nodes makes a great challenge to the energy and computation consumption of the sensor nodes.

Compressive sensing [2][3]is a collection of recently proposed sampling methods in information theory which deals with estimating an unknown signal with fewer measurements than the Nyquist sampling theorem dictates. Compressed sensing (CS) theory has recently become widely popular to improve system efficiency in the field of image processing, geophysics, medical imaging, computer science, as well as in Wireless Sensor Network (WSN). The two key ideas of CS are sparsity and incoherence. The former depends on single itself, the latter depends on both single and sensing environment.

Wireless sensor networks (WSN) gather sensing data from spatially deployed sensor nodes through wireless communications. However, the enormous energy consumption and communication costs are challenges inherent in a large-scale WSN because of a number of sensor nodes. It is well known that proper data aggregation techniques [4] may reduce the amount of data transmission load carried by a WSN and may hence improve its performance in every aspect. However, conventional aggregation techniques just extract some statistical characteristics from sensing data and loss some features [5], other technology such as Slepian-Wolf coding [6], need using non-cooperative data compress, without prior knowledge of the data correlation structure could render it impossible to perform the coding operations. Collaborative in-network compression makes it possible to discover the data correlation structure through information exchange [7], the resulting high computation and communication load may potentially offset the benefit of this aggregation technique.

CS has been envisioned as a useful technique to improve the performance of WSN [8][9]. It can use for single processing, signal detection [10], channel estimation [11] [12], etc. Current works for aggregation mainly consider single-hop aggregation [13] and data distribution [14]. In this paper, we consider the application of a new decentralized compression technology known as compressed sensing (CS), to in-network data aggregation.

The remaining of our paper is organized as follows. In Section II, we discuss some related work about data aggregation in WSN. In section III, we review the basic CS concepts that are relevant to our problem. In section IV, we discuss the applications of Compressive Sensing for Wireless sensor network. In section V, we discuss the concept of Orthogonal Matching Pursuit. We then describe the work process of CS data aggregation for WSN in section VI. In Section VII, we show the simulation results. Finally, we conclude the paper in Section VIII.


doi: 10.4304/jnw.8.1.197-204

## II. Data Aggregation and Related Work

### A. Data Aggregation

One of the basic distributed data processing procedures in the wireless sensor networks is data aggregation. Data aggregation has been put forward as an essential paradigm for wireless routing in sensor networks [15]. An important topic addressed by the WSN community over the last several years has been in-network aggregation. In typical sensor network scenarios, data is collected by sensor nodes throughout some area, and needs to be made available at some central sink node (s), where it is processed, analyzed, and used by the application. Given the application area, network resource constraints, and the fact that local computation often consumes significantly less energy than communication, in-network data aggregation and management are at the very heart of sensor network research.

In WSN, a sensing field usually exhibits high correlation between the measured data and can be compressible in some transform domains. In many cases, data generated by different sensors can be jointly processed while being forwarded towards the sink,.

Data aggregation is recognized as one of the basic distributed data processing procedures in wireless sensor net-works for saving energy and reducing medium access layer contention. The aggregation operation is also helpful in reducing the contention for communication resources.

In-network data aggregation can be considered a relatively complex functionality, since the aggregation algorithms should be distributed in the network and therefore require coordination among nodes to achieve better performance. Also, we emphasize that data size reduction through in-network processing shall not hide statistical information about the monitored event.

### B. CS and its Applation in Data Aggregation

In wireless sensor network (WSN) there are two main problems with conventional compression techniques. Firstly, the compression performance relies heavily on how the routes are organized. In order to achieve the highest compression ratio, compression and routing algorithms need to be jointly optimized. Secondly, efficiency of an in-network data compression scheme is not solely determined by the compression ratio, but also depends on the computational and communication overheads.

Compressive data aggregation technique helps to solve these problems. By the CS technique, data are gathered at some intermediate node where the data size is reduced by applying compression technique without losing any information of complete data. Compressive data aggregation technique requires each node in the WSN to send exactly k packets irrespective of what it has received, which means, compared with traditional techniques, more load for the nodes which are far away from the sink and less load for the nodes that are close to the sink. Data compression and aggregation technique have the potential to improve WSN energy efficiency and minimize communication.

The applications of compressive sensing for data gathering have been studied in a few papers. [16] proposed that every sensor in the field computes and stores sparse random projections in a decentralized manner and sends its aggregates randomly within the network. In [17] Lee et al. investigated CS for energy efficient data gathering in a multi-hop wireless sensor network. In [18], Luo et al. applied compressive sensing theory for efficient data gathering in a large scale wireless sensor network. They showed that the proposed scheme can substantially save communication cost and increase network capacity. In [19], the authors propose two different ways (plain-CS and hybrid-CS) of applying CS to WSNs at the networking layer, in the form of a particular data aggregation mechanism.

## III. Compressed Sensing Basis

### A. Compressed Sensing

A simplified explanation of compressed sensing is as follows: A signal projected linearly onto a lower dimensional space can be used to reconstruct the original higher-dimensional signal with high probability if the signal is sparse and the projection matrix satisfies the restricted isometry property (RIP, explained in section III, D).

Traditionally, one is required to acquire the full N-sample of signal to compute the complete group of transform coefficients. The traditional compression techniques suffer from an important inherent inefficiency since it computes all N coefficients and records all the nonzero, although $K << N$ .

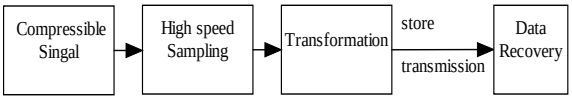



Figure 1. Conventional compress method.

CS can be explain simply as follows: A signal projected linearly onto a lower dimensional space can be used to reconstruct the original higher-dimensional signal with high probability if the signal is sparse and the projection matrix satisfies the restricted isometry property (RIP).The work process is shown in Figure 2. Compressed sensing conducts a signal-independent linear projection of the original signal onto a lower-dimensional space (M-dimensional, $M \le S \le N$ ), which yields poor compression performance. Therefore, it can be stated that the conventional compression method is superior to compressed sensing for the known signals in terms of compression performance.

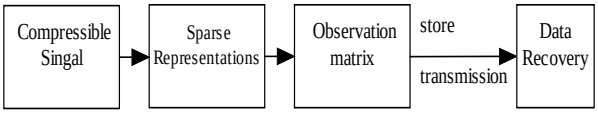



Figure 2. Compressed sensing method

### B. Sparse Representations

The data should be compressible and can be transformed in some sparse representations. A signal is called sparse when it is represented by a small number of nonzero coefficients in any convenient domain [20].If a signal is represented in a non-orthogonal domain, including any user-defined domain, the signal is also called sparse.

The natural signals in time domain are not sparse, but they are sparse when transform to some domain. Take a figure for example, almost pixels are nonzero, but they tend to be zero in wavelet domain, most information of the original figure can be expressed by several big coefficients. A 1-D discrete time domain signal $f$ with length $N$ can be expressed by a set of Linear Combinations of Orthonormal Basis

$$f = \sum_{i=1}^{N} x_i \psi_i \text{ or } f = \Psi x \tag{1}$$

where, $\Psi = [\psi_1, \psi_2, \ldots, \psi_N]$, $\psi_i$ is column vector, $x_{N*1}$ is weighted coefficients of $f$, $x_i = \langle f, \psi_i \rangle = \psi^T{}_i f$. $x$ is the visible equivalent representation of the signal $f$, if $x$ has only a small number of large coefficient, then the signal $f$ is compressible. If $x$ have $k$ nonzero element, we call $x$ is a $k$ sparse representations of the signal $f$.

### C. Obervation Matrix

We first consider First consider the general problem of signal reconstruction, given a observation matrix $\Phi \in R^{M*N} (M << N)$ and the linear observed value of a unknown signal $x$ in this matrix $y \in R^N$

$$y = \Phi x \tag{2}$$

In Eq. (1), can be considered as a linear projection of $x$ from the $\Phi$, so we can reconstruct $x$ from $y$. From (1) we know that the dimension of $y$ is far less than x, so it has infinite many solutions.

It is hard to recovery $x$ from $y$ because (1) is an indeterminate equation. However, if $x$ is k-sparse, and $y$ and $\Phi$ satisfy RIP (Restricted Isometry Property) condition, $x$ can be reconstructed by solve the optimal problem of $l_0$ norm

$$\hat{x} = \arg\min \|x\|_0 \quad s.t. \quad \Phi x = y \tag{3}$$

where, $\| \|_0$ is the $l_0$ norm of vector, which express the numbers of nonzero. The observe numbers of $M$ (dimension of $y$) should meet $M = O(K \log(N))$, $\Phi$ must meet RIP.

### D. Restricted Isometry Property

RIP casts an NP-hard problem onto an l1-norm minimization method. When an M×N measurement matrix $\Phi$ and a N×N transform domain matrix $\Psi$ satisfy the following inequality for all S-sparse vectors, matrix $\Theta = \Phi\Psi$ is said to obey RIP of order S with

$$(1-\delta)\|s\|_2^2 \le \|\Phi\Psi s\| \le (1+\delta)\|s\|_2^2 \tag{4}$$

$\|s\|_2 = \sum_i S_i^2$ and $\delta (0 \le \delta \le 1)$ are the smallest constant that satisfies the above inequality. The RIP show that Euclidian distance between any two N×1 S-sparse vectors is approximately preserved after the measurement as long as $\Theta$ obeys RIP.

### E. Reconstruction

If $X$ is represented by $N \times 1$ coefficient vector S in the $N \times N$ transform domain $Y$, it is expressed as $Y = \Phi\Psi S$. It has been shown that $S \cdot \log(N/S)$ measurements are enough for the reconstruction of an S-sparse signal when measurement matrix $\Phi$ consists of random ensembles. So reconstruction of the compressed sensing refers to finding $N \times 1$ S-sparse vector s from known $M \times 1$ measurement output matrix $Y$, $M \times N$ measurement matrix $\Phi$, and N × N transform matrix $Y$ when $Y = \Phi\Psi S$ (S<M<<N). Reconstruction of the original signal s can be found via the l1 minimization problem, as described below, as long as RIP is satisfied.

$$\min \|\hat{s}\|_{l_1} \text{ subject to } Y = \Phi\Psi\hat{s},\ \hat{s} \in R^N$$

$$\text{if RIP satisfies } \hat{s} = s, \tag{5}$$

where $\|\hat{s}\| = |s_1| + |s_2| + \cdots + |s_n|$, and $R^N$ is the set of $N \times 1$ vector.

An intuitive explanation of RIP is as follows: the Euclidian distance between any two $N \times 1$ S-sparse vectors is approximately preserved after the measurement as long as Q obeys RIP. The l1-norm minimization problem finds $\hat{s}$ that minimizes the l1-norm subject to $Y = \Phi\Psi S$ among all possible S-sparse vectors.

## IV. Appliation of Compressive Sensing for Wireless Sensor Network

### A. WNS without Compressed Sensing

A WSN consists of a large number of nodes with small devices each with sensing, processing, communication and controlling abilities to monitor the real environment. A WSN with $N$ nodes, each having information or data $x_i, i = 1, 2, 3 \cdots N$ .Each $x_i$ has a scalar value and therefore network data is arranged in a vector as [21]

$$X = [x_1, x_2 \cdots x_n]^T \tag{6}$$

As mentioned in Section 2, one is required to acquire the full $N$-samples of signal □ to compute the complete group of transform coefficients although K<<N (K is nonzero coefficients of information). The network data vector is very large and it is a problem of processing in a WSN with thousands or millions of nodes.

*B. WNS with Compressed Sensing*

The phrase compressed sensing refers to the problem of realizing a sparse input $X$ using few linear measurements that possess some incoherence properties. The key objective in compressed sensing (also referred to as sparse signal recovery or compressive sampling) is to reconstruct a signal accurately and efficiently from a set of few non-adaptive linear measurements.

CS theory asserts that one can recover certain signals from far fewer samples than what have been acquired from the sensors, if those signals can be sparsely represented in a proper basis [22]. It also can be used in a WSN scenario with n nodes, each node acquiring a sample (e.g., temperature) $x_i$. Our goal is to collect the vector $\boldsymbol{X} = [x_1, x_2 \cdots x_n]^T$ at the sink. We say x has an m-sparse representation if there exists a proper basis $\Psi = [\psi_1, \psi_2 \cdots \psi_n]^T$, s.t. $X = \sum_{i=1}^{m} S_i \psi_i$ and $m << n$ .Now the CS theory suggests that, under certain conditions, instead of collecting $X$, we may collect $Y = \Phi X$, where $\Phi = \{\phi_{j,i}\}$ is a $k \times n$ "sensing" matrix whose entries are i.i.d. zero-mean random variables with variance $1/k$. Consequently, we can recover $X$ from $Y$ by solving the convex optimization problem

$$\min \|S\|_{l_1} = \min \sum_i |s_i| \text{ subject to } Y = \Phi \Psi S \qquad (7)$$

and letting $X = \Phi \hat{S}$, with $\hat{S}$ being the optimal solution of (2).

The condition that guarantees the correctness of this recovery is given by $k > C \times m \times \log n$, where C is some small constant. The sink needs to collect only m<< n samples to reconstruct the sensory data represented by the n samples.

According to the description above, $\Phi$ and $\Psi$ are two keys to applying CS in WSN. Using pseudo-random number generators to produce the entries of $\Phi$, we can meet the i.i.d. criterion while avoiding actually transmitting $\Phi$ by seeding the generators using publicly known numbers. For example, if we associate a specific generator with a node *i*, the *i*-th column of $\Phi$, $\phi_i$ can be generated anywhere with consistent output. Although the $\Psi$ that yields the sparsest representation of $X$ may not be known, wavelets are in general considered as a good candidate for $\Psi$.

## V. Orthogonal Matching Pursuit algor ithm

There are lots of common CS reconstruction algorithms including both the LASSO (Least Absolute Shrinkage and Selection Operator) and LARS (Least Angle Regression) algorithms and the Orthogonal Matching Pursuit (OMP) algorithm. There have been two distinct major approaches to sparse recovery that each present different benefits and shortcomings. The first, ℓ1-minimization methods (see appendix B) such as Basis Pursuit (BP), use a linear optimization problem to recover the signal. This method provides strong guarantees and stability, but relies on Linear Programming, whose methods do not yet have strong polynomially bounded runtimes. The second approach uses greedy methods that compute the support of the signal iteratively.

The sparse vector $\Theta$ can be accurately recovered from $Y$ using the reconstruction techniques, one of those is Orthogonal Matching Pursuit (OMP) [23], which solves the reconstruction problem by identifying the component of the sparse signal in each iteration which is most coincident with the sampling value. The most potent is fast implementation and easy to realize, especially apply to energy limited WSN. When SINK received the M values, it can use OMP algorithm to rebuild $\hat{y}$, and then compute the Mean Squared Error (MSE). The work process is below

**Input:** $M * N$ **observation matrix** $\Phi_{M*N}$, $N$ Dimensional sparse vector $Y_{N*1}$, sparsity level $k$

**Output:** $\hat{\Theta}$, which is the estimation of $\Theta$.

Index matrix $\Lambda_1 \cdots \Lambda_N$;

$y_k$, which is the estimation of $y$;

residual $r_k = y - y_k$.

The work process is below

1) Initialize the residual $r_0 = y$, index set $\Lambda_0 = \varnothing$ and counter $t = 1$;

2) Find the column vector $\lambda_t$, that is mostly correlated with the residual: $\lambda_t = \arg\max_{j=1,\ldots,N} \left| \left\langle r_{t-1}, \varphi_j \right\rangle \right|$

3) Reconstruct the matrix $\Lambda_t = \Lambda_{t-1} \cup \{\lambda_t\}$ and $\Phi_t = [\Phi_{t-1}\ \varphi_{\lambda_t}]$, $\Phi_0$ is empty matrix;

4) Solve the least-square problem

$$e_t = \arg\min_e \|\Phi_t e - y\|_2 \qquad (8)$$

5) Compute the estimations and residual in this round $y_t = \Phi_t e_t$

$$r_t = y - y_t \qquad (9)$$

6) $t = t + 1$, if $t < k$ return to step 2)

7) Nonzero index of $\hat{\Theta}$ store in $\Lambda_k$.

## VI. Aggregation in WSN based on CS

In WSN, data generated by different sensors can be jointly processed while being forwarded toward the sink. Data aggregation is the simplest type of in-network processing which combines data from different sources or nodes into a single entity. In sensor networks, the data gathered by spatially close sensors are usually correlated, so the node information is compressible. We can use CS technology to eliminate data redundancy.

We model the WSN as a set of nodes, which consisted of $N$ nodes and a sink. Each node is associated with a geographical location. We assume that all nodes send sensory data to the sink with the same rate, time is slotted and all the nodes are synchronized, and the network is operated in a conflict-free and scheduled manner. We

assume that all the nodes have the same transmit power and the same data-rate. We use clustering scheme, the cluster head collect data from cluster member and sent it to sink. The algorithm process is as follows.

*A. Sparse Representations and Transformation Matrix*

From the analysis in section III we know, $\Psi$ is used for making compressible single sparse. We use wavelet basis compressed algorithm or 5/3 wavelet lifting algorithm in index A. Assume there is a perfect optimal basis $\Psi$, which can make original data $X_{N*1}$ sparse, the compressed data is $\Theta_{N*1}$ with $k$ nonzero values from beginning.

$$\Theta_{N*1} = [\theta_1 \quad \theta_2 \cdots \theta_k \quad 0 \cdots 0]^T_{N*1} \tag{10}$$

*B. Generation of Observation Matrix*

The $N(0,1)$ independent identically distributed observation matrix meet the requirement of CS theory [24], the element in matrix almost not correlated with sparse element, the observation matrix $\Phi$ is bellow:

$$\Phi = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,N} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,N} \\ \vdots & \vdots & & \vdots \\ \phi_{M,1} & \phi_{M,2} & & \phi_{M,N} \end{bmatrix}_{M*N} \tag{11}$$

The sink needs to store all the n seeds, such that it can generate $\Phi$ in order to process the compressed data. In N nodes, the nodes which store nonzero value (the former k nodes), generates M (where, $M = k \log N$ ) elements which are $N(0,1)$ distribution and mutually independent $\{\phi_{i,j}\}_{i=1}^{M}$,

$$\Phi = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,k} & 0 & \cdots & 0 \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,k} & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & & & \vdots \\ \phi_{M,1} & \phi_{M,2} & \cdots & \phi_{M,k} & 0 & \cdots & 0 \end{bmatrix}_{M*N} \tag{12}$$

*C. Clustering and Data Transmission*

The clustering is an efficient mechanism in large scale Wireless Sensor Network. In our model, cluster work in slotted way. Cluster heads will aggregate the data of slave nodes, and then sent it to sink.

According to the principle of observation, the cluster head just need the data from its slave nodes. In a cluster with N nodes, $\{Node \quad i\}_{i=1}^{k}$ generate $M$ values, multiply it by its self data $\theta_i$ as below

$$[\theta_1 * \phi_{1,1} \quad \theta_2 * \phi_{2,1} \quad \cdots \quad \theta_k * \phi_{M,1}]^T$$

$$[\theta_1 * \phi_{1,2} \quad \theta_2 * \phi_{2,2} \quad \cdots \quad \theta_k * \phi_{M,2}]^T \tag{13}$$

$$[\theta_1 * \phi_{1,k} \quad \theta_2 * \phi_{2,k} \quad \cdots \quad \theta_k * \phi_{M,k}]^T$$

In slot 1, $\{Node \quad i\}_{i=1}^{k}$ sent $\theta_i * \phi_{1,i}$, cluster head receive the information from different slave nodes,

$$\sum_{i=1}^{k} \theta_i * \phi_{1,i} = \theta_1 * \phi_{1,1} + \theta_2 * \phi_{1,2} + \cdots + \theta_k * \phi_{1,k} \tag{14}$$

So the data received cluster is

$$y = [\sum_{i=1}^{k} \theta_i * \phi_{1,i} \quad \cdots \quad \sum_{i=1}^{k} \theta_i * \phi_{M,i}]^T \tag{15}$$

The clusters sent data to sink by the same method.

*D. Algorithm of Reconstruction*

After the Step 1 and Step 2, the selected node generates locally values, The $k$ column includes $M$ elements $\{Nodei\}_{i=1}^{k}$ which generate by node $k$, and multiply $\{\phi_{i,j}\}_{i=1}^{M}$ by itself data $\theta_i$, i.e. $\Phi * \Theta$, and send it to sink.

So the data received by SINK is

$$y = [\sum_{i=1}^{k} \theta_i * \phi_{1,i} \quad \cdots \quad \sum_{i=1}^{k} \theta_i * \phi_{M,i}]^T \tag{16}$$

The sparse vector θ can be accurately recovered from y using the reconstruction techniques, one of those is Orthogonal Matching Pursuit (OMP), which most potent is fast implementation and easy to realize, especially apply to energy limited WSN. When SINK received the M values, it can use OMP algorithm to rebuild $\hat{y}$, and then compute the Mean Squared Error (MSE).

## VII. Simulation Results

Assume there are $N$ nodes in a 2 dimensional area show in Figure 3, each node store a sensing value in Figure 4, so the original data are $N$ dimensional vector $X_{N*1}$ .We set $N = 256$ , $k = 60$ ( $M = k \log N$ , i.e. $M = 144$ ).

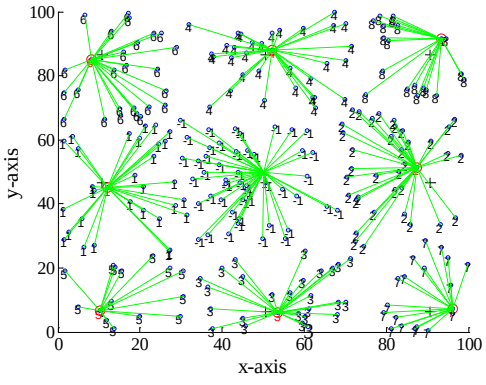



Figure 3. Simulation scenario.

Firstly, we transform the original data in figure 4 to k-sparse matrix by transform matrix $\Psi$ , so the $N$ dimensions $X$ will be compressed to k-sparse single $\Theta$ in Figure 5.

The data received by SINK is show in Figure 6.

Finally, when SINK received y, recovery the data by OMP algorithm, we can estimate the $\hat{\Theta}$ from $\hat{\Theta}$, Figure 7

is the comparison between original data and reconstructed data. By comparison between original data and reconstructed data in Figure 7, we can see that MSE of reconstructed accuracy less than 0.005.

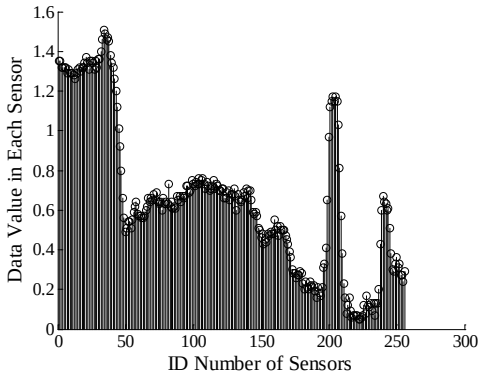



Figure 4. Original data $X$

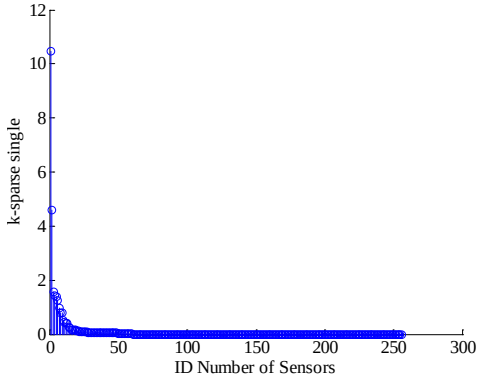



Figure 5. Sparse coefficient.

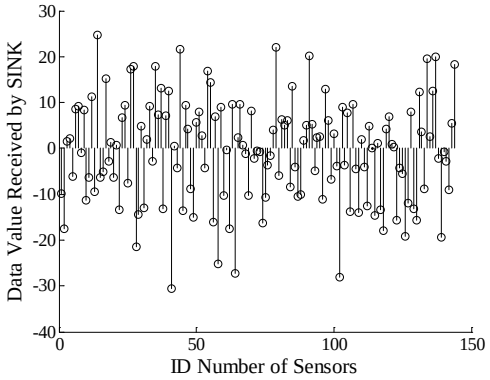



Figure 6. Received data by SINK.

We select different observations $M$ , and get different reconstructed precision. The more the observations, the more accurate of data be reconstructed. Figure 8 is the relationship between observations and reconstruction MSE. From Figure 8 we can see, the MSE is below 0.02 when observations larger than 80, we can recovery the data efficiently.

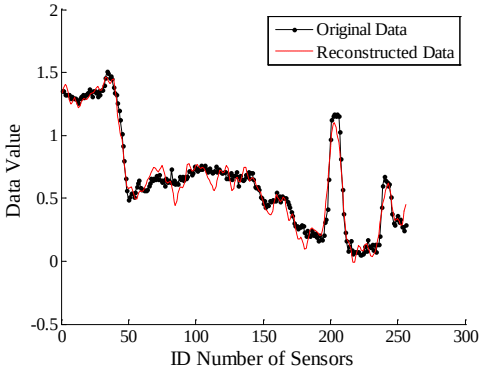



Figure 7. Comparison between original data and reconstructed data

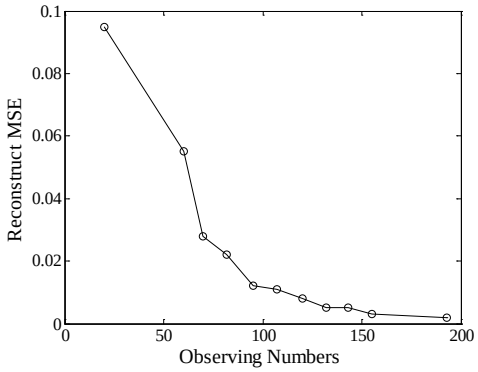



Figure 8. Observations and reconstruct MSE

## VIII. Conclusion

In the paper, we give the processes of CS aggregation in WSN, including sparse presentation of signal, observation matrix and reconstruction algorithm design. We also discussed the relationship between observations and reconstruct MSE are also discussed. The transmitted data are reduced due to the sparsity of sensing signal, the communication overload of cluster head and slaves also can be reduced which can reduce the energy consumption and prolong the lifetime of the whole WSN.

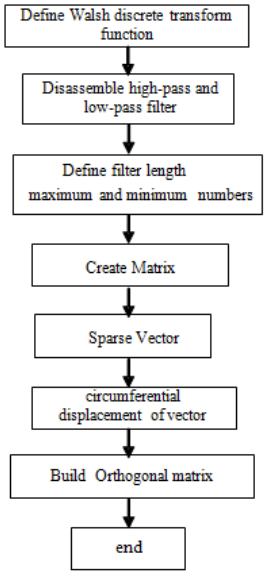
As we know, how to select an optimal transformation basis is directly related to the selection of the follow-up observation matrix, and it will also affect the reconstruction quality. In this paper, we did not do much introduction to the design of the transformation matrix. This should be our research work in future.

## Appendix A Wavelet Transform Frame

Wavelet transformation is a fast-developing and popular signal analysis method. Wavelet analysis allows the use of long time intervals for more precise low frequency information, and short regions for high frequency information.

Lifting scheme is a new method for constructing wavelets and performing wavelet transform based on wavelet theory in recent years. It has many advantages for example simple structure, fast calculation, in-place operation. At the same time, it can realize integer to integer wavelet transform and can reconstruct original signal accurately. So it has been extensively used in the

field of signal process. Lifting scheme based on wavelet transform is also used in our paper. The main work process is shown below.




## APPENDIX B MINIMUM $l_1$-MINIMIZATION ALGORITHM

Suppose there exists an unknown signal $x_0 \in \square^n$, a measurement vector $b \in \square^d (d < n)$, and a measurement matrix $A \in \square^{d \times n}$, such that $A$ is full rank and $b = Ax_0$. Recovering $x_0$ given $A$ and $b$ constitutes a non-trivial linear inversion problem, since the number of measurements in $b$ is smaller than the number of unknowns in $x_0$. A conventional solution to this problem is the linear least squares, which finds the minimum $l_2$-norm solution (or the solution of least energy) to this system. However, if $x_0$ is sufficiently sparse and the sensing matrix $A$ is incoherent with the basis under which $x_0$ is sparse (i.e., the identity matrix in the standard form), then $x_0$ can be exactly recovered by computing the minimum $l_1$-norm solution. The $l_1$-minimization refers to finding the minimum -norm solution to an underdetermined linear system $b = Ax_0$. The $l_1$-minimization algorithm can be described below.

If we replace $l_0$ -norm by $l_1$ -norm, we obtain,

$$\hat{x} = \arg\min \|x\|_1 \quad s.t.\ \Phi x = y$$

This is a convex optimal problem, and can be solved by transform to a Linear Programming problem. Its Computational-complexity is $O(N^3)$, it also called BP (Basis Pursuit) algorithm. If we consider the reconstruction error, the problem become

$$\min \|x\|_1 \quad s.t.\ \|\Phi x - y\|_2 \le \varepsilon$$

It can be solved by second-order circular con programming.

At present, the basic CS theory includes recoverability and stability: the former quantifies the central fact that a sparse signal of length p can be exactly recovered from far fewer than p measurements via l1-minimization or other recovery techniques, while the latter specifies the stability of a recovery technique in the presence of measurement errors and inexact sparsity. So far, most analyses in CS rely heavily on the Restricted Isometry Property (RIP) for matrices.

## ACKNOWLEDGMENT

The work of this paper is supported by Fujian Science Foundation (No. 2011J01356), and Foundation for Young Professors of Jimei University, China (No. 2008B002).

## REFERENCES

[1] Ian F. Akyildiz, Weilian Su, Yogesh Sankara, and E. Cayrci, "Wireless sensor network: a survey", *Computer Networks*, vol. 38, no. 4, pp. 393-422, March 2002,

[2] D. L. Donoho, "Compressed sensing", *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp.1289–1306, 2006.

[3] Y. C. Eldar and G. Kutyniok, "Compressed Sensing: Theory and Applications, "*Cambridge University Press*, 2011.

[4] Liu Xiang, Jun Luo, Vasilakos, "A Compressed data aggregation for energy efficient wireless sensor networks Sensor", Mesh and Ad Hoc Communications and Networks (SECON), 2011 8th Annual IEEE Communications Society Conference on Digital Object Identifier, pp.46 - 54, 2011.

[5] J. Gao, L. Guibas, N. Milosavljevic, and J. Hershberger, "Sparse Data Aggregation in Sensor Networks", *in Proc of the 4th ACM IPSN*, 2007.

[6] Z. Tu, J. Li, and R. Blum, "An efficient turbo-binning approach for the Slepian-Wolf source coding problem, "*Eurasip Jour. on Applied Signal Processing - Special Issue on Turbo Processing*, 2003.

[7] P. Zhang, Peng, C., Liu, Minrun, "The application of compressed sensing in wireless sensor network", *2009 International Conference on Wireless Communications and Signal Processing*, no. 5 2009.

[8] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed Sensing for Networked Data", *IEEE Signal Processing Mag*, vol. 25, no. 3, 2008.

[9] W.Bajwa, J.Haupt, A.Sayeed, Nowak R, "Compressive wireless sensing", *Proceedings of the International conference on Information Processing in Sensor Networks*, Washington D. C, USA, pp.134–142, 2006.

[10] C. Li, T.Sun, K.F. Kelly, Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing", *IEEE Trans Image Process*, 21 (3), pp.1200–1210, 2012.

[11] Taubock G, Hlawatsch F., "A compressed sensing technique for OFDM channel estimation in mobile environments: Exploiting channel sparsity for reducing pilots", *Proceedings of the IEEE International conference on Acoustics, Speech, and Signal Processing*. Washington D. C., USA, pp.2885–2888, 2008.

[12] BajwaWU, Jarvis Haupt G R, Nowak R., "Compressed channel sensing", *Proceedings of Conference on Information Sciences and Systems*, Washington D. C., USA, pp.5–10, 2008.

[13] M.Rabbat, J.Haupt, A.Singh, and R.Nowak, "Decentralized Compression and Predistribution via Randomized Gossiping", in Proc. of the *3th ACM IPSN*, 2006.

[14] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive Wireless Sensing", *in Proc. of the 3th ACM IPSN*, 2006.
[15] Kalpakis, K. "Everywhere sparse approximately optimal minimum energy data gathering and aggregation in sensor networks", *ACM Transactions on Sensor Networks (TOSN), vol.* 7, no. 1, pp.9-14, 2010
[16] W. Wang, M. Garofalakis, and K. Ramchandran, Distributed sparse random projections for refinable approximation, *In Proceedings of the 6th international conference on Information processing in sensor networks*, pp.331–339. 2007.
[17] S. Lee, S. Pattem, M. Sathiamoorthy, B. Krishnamachari, and A. Ortega, "Spatially-localized compressed sensing and routing in multi-hop sensor networks", in *Proc.* vol. 5669, *Third International Conference on Geosensor Networks*, pp. 11-20, 2009.
[18] C. Luo, F. Wu, J. Sun, and C. W. Chen, Compressive data gathering for large-scale wireless sensor networks, in Proc.ACM Mobicom, pp. 145-156, 2009.
[19] Luo, Jun.; Liu, Xiang.; Rosenberg, Catherine. Luo, Jun.; Liu, Xiang.; Rosenberg, Catherine. "Does compressed sensing improve the throughput of wireless sensor networks?" *Proceedings of the IEEE International Conference on Communications*, 2010
[20] A. Cohen, W. Dahmen, and R. Devore, "Compressed Sensing and Best k-term Approximation", *Journal of the American Mathematical Society*, Vol. 22, No. 1, pp. 211–231, Jan. 2009.
[21] Chun Tung Chou, R. Rana and Wen Hu, "Energy efficient information collection in wireless sensor networks using adaptive compressive sensing", *in IEEE 34th Conference on Local Computer Networks (LCN 2009)*, pp. 443-450, 2009.
[22] E. Cand`es and M. Wakin, "An Introduction To Compressive Sampling", *IEEE Signal Processing Mag*, vol. 25, no. 3, 2008.
[23] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit", *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.
[24] Chen S B, Donoho D L, Saunders M A, "Atomic decomposition by basis pursuit", *SIAM Journal on Scientic Computing*, pp. 33–61, 1998.

**Guangsong Yang** He graduated and received BS degree from Guilin University of electronic science and technology in 1990, and received MS degrees form Institute of Computer Science, Guizhou University in 2002, He then pursued the Ph.D. degree in Xiamen University and received the degree in Dec.2005. He is currently a vice professor in school of information engineering, and is the vice director of the communication and information technology institute, Jimei University. His research interests are in wireless communications, with a focus on wireless ad hoc network, energy efficient sensor networks, and distributed signal processing.

**Mingbo Xiao** He received his Ph.D. degree from Purdue University in 2002. He was a Post doc Fellow in Rice University from May 2002 to July 2003. He is currently a Professor in the Department of Communication Engineering at Xiamen University. His research interests fall in the general area of wireless networking, spanning from the network layer to the physical layer. His research draws on a combination of tools in networking theory (e.g., traffic engineering and queuing theory), wireless communications, information theory, and probability theory. A principal goal of his current research is to establish a framework for cross-layer optimization and design in wireless networks, particularly in mobile ad hoc networks and cellular networks.

# Detection of Underwater Carrier-Free Pulse based on Time-Frequency Analysis

Yunlu Ni and Hang Chen
School of Marine, Northwestern Polytechnical University, Xi'an, China
Email: niyunlu@mail.nwpu.edu.cn

***Abstract*—Carrier-free short pulse widely employed in UWB radar is brought into high-resolution sonar system, which has unique advantages: attaining more target information, restraining fluctuation of reverberation envelop efficiently in short-range detection and achieving accurate estimation. In essence such pulse is transiently short in time domain and wide in frequency domain, and as such it is difficult to separate signal to noise based on Fourier Transform spectrum. So as to seek for detection methods of short pulse, minor differences of energy distribution of time-frequency characteristics are presented on three time-frequency methods such as Short Time Fourier Transform, Wavelet Transform and Hilbert-Huang Transform. With these results, a tri-channel detector is established for such underwater short pulse in noise environment, which is generally suitable not only for detection module of underwater sonar system but also that of radar system.**

***Index Terms*—carrier-free pulse, time-frequency analysis, characteristics, detection**

## I. INTRODUCTION

Carrier-free short pulse as Ultra-Wideband Signal [1] is kind of transient signal and widely applied in detection, imaging, accurate orientation and target identification, as has relatively large instantaneous bandwidth with considerable information, anti-jamming capability, good stealth and anti-stealth effect, etc. Taken into Sonar system, its has unique advantages [2-3]: 1) detecting invisible or quiet target because of broadband characteristics of both low-frequency and high-frequency; while stealthy design of the target is always effective on certain frequency band, as long as Sonar system has a sufficiently wide signal bandwidth invisible target can be detected; 2) high range resolution, distinguishing between the target scattering points; separation of multipath channel signals; 3) a good ability of identification, separation of the response of different target in different areas, highlighting the target characteristics for identification; 4) ultra-short-range detection capability; 5) to reduce the reverberation fluctuation; 6) anti-jamming performance.

Carrier-free pulse is not only kind of Ultra-Wideband (UWB) signal as definition, but also transient signal. Detection method of UWB signal is gradually mature on derived Raker receivers [4] and has applications in communications at low SNR [5]; some are correlators using the optimal deterministic template to optimize a deterministic template to achieve the maximum energy capture capability in the sense of ensemble average [6]. Novel detection methods not based on correlation are also investigated. A subspace-based detection method [7] is proposed for analog space-time codes wedded with UWB transmissions. Without estimating the channels at the receiver, the proposed algorithm yields the estimation of transmitted symbols by minimizing some quadratic form built on the orthogonality between signal and noise subspaces. Dai [8] researches Cramer-Rao lower bound on the basis of the Maximum Likelihood rule and gives precision of vital signal detection. A compressed sensing theory is developed to correctly detect original transmitted signal in Ref [9], while the estimation performance is improved by 4.5dB. There detection methods are not completely suitable for underwater detection because of different pulse width and transmission channel. Except for traditional detection method for underwater transient signal, such as correlation detection [10] and peak energy detection [11], novel methods fit for underwater conditions are also investigated. Hartley Phase Cepstrum offers improved detection and localizations of a sequence of transient events [12]. For Sonar detection and target identification, Wu [13] provided methods of detecting weak transient signal according to the sudden change of prediction error in chaotic background.

Except for detection based on channel or pulse wave, single or combined time frequency analysis is also widely employed. Because of the shortcomings of current feature extraction and fault diagnosis technologies, a new approach based on wavelet packet decomposition (WPD) and empirical mode decomposition (EMD) are combined to extract fault feature frequency and neural network for rotating machinery early fault diagnosis is proposed [14]. A new automatic transient detection and localization for analysis of power quality disturbances is analyzed. The proposed method is based on an over-complete dictionary (OCD) matrix and an ℓ1-norm minimization algorithm. Experiment results show that the proposed method provides accurate time-information of impulsive and oscillatory transients under high levels of noise, and also preserves signature of transient events [15]. Masoum [16] uses PQ (Power Quality), which has become a major concern owing to increased use of sensitive electronic equipment, to detect these disturbance waveforms automatically in an efficient manner. It is the fact that PQ


doi: 10.4304/jnw.8.1.205-212

disturbances vary in a wide range of time and frequency, which make automatic detection of PQ problems often difficult and elusive to diagnose, as is same to underwater short pulse. PQ disturbances have been defined into several categories and software based novel approach techniques for detection of PQ disturbances by time and frequency analysis where wavelet transform are proposed. These techniques detect PQ problems of waveform distortion and provide a promising tool in the field of electrical power quality problems.

As electronic technique develops, complex system is realizable and real-time. The problem faced is high-resolution detection, while it is economic to add new algorithm to original system. Short pulse with short time duration has the characteristics of transient signals. Its time-frequency characteristics are fundamental for target detection and identification which has important applications in the detection of quiet target. As so, three time-domain frequency methods are employed for obtaining spectrum characteristics to compose a tri-channel detector.

The paper is organized as follows. The time-frequency principles and formula are briefly presented in Section II. Underwater target echo is shown in Section III. Spectrum and time-frequency analysis result is described in Section IV. A tri-channel detector is developed and detector performance is shown in Section V. In Section V, the conclusion is collected, while lack of such detector and the future research is presented.

## II. Time Frequency Analysis Principles

Time-frequency analysis plays a significant role in signal detection. Instantaneous wide bandwidth in frequency domain is an important feature of carrier-free short pulse. Traditional Fourier Transform (FT) is a kind of global transformation without detailed frequency information. Three time-frequency methods are employed to analyze short pulse and obtain its time-frequency characteristics, which provides theoretical basis for detection. The principles of Short Time Fourier Transform, Wavelet Transform and Hilbert-Huang Transform are expressed as follows.

### A. Short Time Fourier Transform

With the development of electronics technique, as well as signal detection systems based on short-time Fourier transform become practical and widely applied.

Short Time Fourier Transform (STFT) can be regarded as local spectrum near analysis time, STFT of discrete time signal *x (n)* may then be defined as [17]

$$STFT = \sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{-j\omega_k m} . \qquad (1)$$

Equation (1) shows that $w(n)$, the window function, selectively determines the portion of $x(n)$ which is being analyzed, i. e., $w(n)$ can be considered as a single low-pass filters uniformed spaced in frequency with the determination of all of the properties of the filter bank.

In STFT, discrete signal $x(n)$ first multiplies by analysis window central at time $t$ before Fourier transform. As window function's time shift and frequency shift modulation, it's equivalent to obtain a slice of signal in the vicinity of analysis time point, that is, the local spectrum. Slice at any time $t$ is the local spectrum of the signal at the moment. More important is that once a window has been chosen for the STFT, the time-frequency resolution is fixed over the entire time-frequency plane.

The spectrum of STFT is defined as square of modulus value

$$SPEC(t, f) = \left|STFT(t, f)\right|^2 . \qquad (2)$$

There is few single use of STFT for application, because of its uniformed property as filter, which cannot discover the nonlinear frequency part of wideband signal. However, for energy detection especially for generalized Gaussian pulse, more frequency band is helpful.

### B. Wavelet Transform

To analyze signal structures of very different sizes, it is necessary to use time-frequency analysis with different time supports.

In a Wavelet Transform (WT), the notion of scale is introduced as an alternative to frequency, leading to a so-called time-scale representation and having multi-resolution characteristics, which features are suitable for the detection of transient signals. WT uses short windows at high frequencies and long windows at low frequencies, as is in the spirit of so-called "constant-Q" or constant relative bandwidth frequency analysis. It could effectively focus on the instantaneous part of the signal and overcome the resolution limitation of the STFT. WT is defined as follows [18]

$$WT_s(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} s(t)\varphi^*(\frac{t-b}{a})dt . \qquad (3)$$

Where $a$ and $b$ are the scale parameter and translation parameter, respectively; $\varphi(t)$ is the wavelet function.

The scalogram is defined as the square of WT modulus values

$$S\ CAL(a,b) = \left|WT(a,b)\right|^2 . \qquad (4)$$

Wavelet Transform has found comprehensive application in narrowband and wideband signal estimation. Here simple form of WT is chosen, because of its calculation time. Moreover, it is to verify various methods with different principles can avoid fakes and get better performance in detection although with non-modified method.

### C. Hilbert-Huang Transform

Researches have shown that pre-mentioned STFT and WT are time-frequency analysis based on FT, and inevitably influenced by defects of FT, such as fake frequency and redundant signal component. Empirical

Mode Decomposition (EMD), recently been pioneered by N.E. Huang et al [19], breaks the limitations of FT. The method enables a decomposition of a nonlinear and nonstationary signal into its characteristics scales, so called Intrinsic Mode Function (IMF). The basis of empirical mode decomposition is the assumption that any signal consists of multiple intrinsic modes of oscillation. Especially for transiently and locally signal, Rilling [20] modified EMD to local EMD.

Given a signal $x(t)$, the effective algorithm of local EMD can be summarized as follows:

1) identify all extrema of $x(t)$;

2) interpolate between minima (resp. maxima), ending up with some envelope $e_{\min}(t)$ (resp. $e_{\max}(t)$);

3) compute the mean $m(t) = (e_{\min}(t) + e_{\max}(t))/2$;

4) extract the detail $d(t) = x(t) - w(t)m(t)$;

5) iterate on the residual $m(t)$.

The modified part is the fourth step and $w(t)$ is employed to give extra-iteration for local points with large error, avoiding contaminating other stop–iteration points. After performing EMD operation for all fluctuations composing the entire signal, the same procedure can be performed for the residual, composed of all local trends. Finally the signal $x(t)$ may thus defined as

$$x(t) = \sum_{n=1}^{N} d_n(t) + r_N(t) . \tag{5}$$

Where $d_n(t)$ means the nth IMF, $r_N(t)$ represents the trend, which is a monotonic function with no more than two extrema.

EMD is on a basis of the data itself, making the decomposition flexible and adaptive.

Through sifting process, Intrinsic Mode Function (IMF) are obtained for spectrum as

$$X(t) = \sum_{j=1}^{n} a_j(t) \exp\left(i \int \omega_j(t) dt\right) . \tag{6}$$

According to (7) to calculate the IMFs energy moment and according to (8) if the signal is discrete [21]

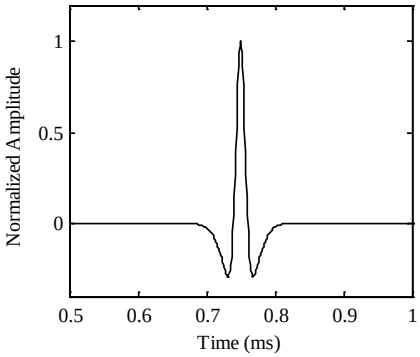



Figure 1. Carrier-free short pulse waveform

$$E_i = \int t \, |c_i(t)|^2 \, dt \tag{7}$$

$$E_i = \sum_{k=1}^{n} (k \cdot \Delta t) \, |c_i(k \cdot \Delta t)|^2 . \tag{8}$$

Where $\Delta t$ is the period of data samples, $n$ is the total number of data samples and $k$ represents the number of data samples. Finally, built the eigenvector and make it as $T$ by normalization

$$T = [E_1, E_2, E_3, \ldots, E_n] / \sum_i E_i \tag{9}$$

$T$ is the percent of the energy of $c_i(t)$ in the whole signal energy. From (7) and (8), we can see that the IMF's energy moment not only contains the size of IMFs energy, but also considers the distribution of IMF's energy with the time parameter $t$. Thus, the IMFs energy moment in the aspect of revealing energy distribution for a signal is better than simple IMF energy.

Without FT and its drawbacks, HHT gives intrinsic mode without losing any information. The low frequency energy residual is higher than FT method.

## III. TARGET ECHO

In this paper, the carrier-free narrow pulses is generalized Gaussian signal, widely used in ultra-wideband signals in the radar and now studied as underwater acoustic emission signal. The waveform is [22]

$$s(t) = \frac{E}{1-a}\left(\exp(-4\pi(t-t_m)^2/\Delta t^2) - a\exp(-4\pi\alpha^2(t-t_m)^2/\Delta t^2)\right) . \tag{10}$$

Where $E$ is peak value of energy, $a$ is Gaussian spread parameters ($s^{-1}$), $\Delta t$ is nominal duration, $t_m$ is the time of peak energy.

Fractional bandwidth is defined as

$$B_f = \frac{2(f_h - f_l)}{f_h + f_l} . \tag{11}$$

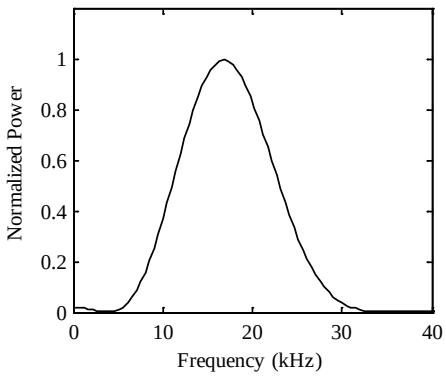



Figure 2. Normalized Power of Target Echo

With $E = 1$, $\alpha = 1$, $\Delta t = 0.1ms$, $\Delta t = 0.1ms$, the waveform of such short pulse is shown in Figure 1. The signal has instantaneous 3dB bandwidth of 18.8kHz and central frequency is 17.6kHz, in accordance with the definition of ultra-wideband signal in (7), whose

fractional bandwidth $B_f$ is much larger than 0.2, as Figure 2.

Such short pulse signal is broadband and cannot neglect the transmission loss in the ocean. The absorption coefficient is half-experienced summed up by

$$a_c = A\frac{Sf_r f^2}{f_r + f} + B\frac{f^2}{f_r}. \quad (12)$$

where $A = 1.89\times10^{-5}$; $B = 2.72\times10^{-5}$; $S$ $(‰)$ means salinity of ocean; $f$ is sound frequency with its unit *kHz*, $f_T(kHz)$ is relaxation frequency, which is multiplicative inverse of relaxation time $T$, related to temperature as

$$f_r = 21.9\times10^{6-\frac{1520}{T}}. \quad (13)$$

Here $T$ means thermodynamic temperature. The unit of $a_c$ is *km/dB*.

Precise target echo cannot be obtained only by analyzed or experiments. The analyzed method is so ideal that complex environment is simplified, while method based on experiments contains different noise with different environment. The target echo is not precisely to obtained yet in anechoic pool. Generalized Gaussian pulse is emitted by programmable signal source without amplifier at the depth of 3m in anechoic pool, direct wave and target echo wave are received by hydrophone with 58dB gain as in Figure 3. Small noise still can be observed.

Combing analyzed pulse and experiment pulse, target echo pulse is fitted at short range. Through it, simulated target echo pulse is obtained. Considering transmission loss in underwater channel and target effect, pulse width becomes wider provided the target has only one highlight point. The negative peak increases relatively to positive peak. Target echo simulation depends on [23], while target echo of such pulse can be seen Figure 4. Analysis results based on time-frequency is discussed as follows.

## IV. TIME-FREQUENCY ANALYSIS RESULT

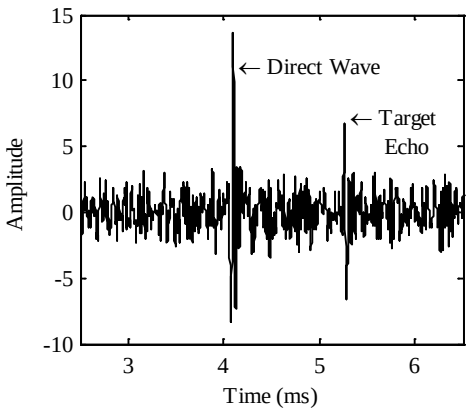



Figure 3. Direct wave and target echo in experiment

Short Gaussian pulse has transient time span and band width. The spectrum power distributes widely at low frequencies with Gaussian envelop, which is shown in Figure 2. Just from FT method, it is insufficient to detect short pulse. To know how the time and frequency varies at the same time, time-frequency method is performed.

### A. STFT Result

After STFT for short pulse, the spectrum is presented in Figure 5. The time-frequency distribution focuses on pulse time, while the energy distribution at different frequency is observable. From the contour line as Figure 6, detailed information and peak energy time are obtained. Energy is higher as the contour line is whiter. Because STFT has the same resolution at different frequency blocks, small energy of short pulse can be seen, such as the part of 30kHz-35kHz and 0kHz-5kHz. Energy summation in pulse width in contour can be employed for detection.

STFT spectrum has full frequency domain information, while the energy distributes uniformly. At a time point, more than one frequency bin appear. For transient short pulse, it is more complex and harder for detection than harmonic wave.

### B. WT Result

Morlet Wavelet Transform is performed on underwater carrier-free echo pulse, and then the scalospectrum is demonstrated in Figure 7.

Because of the imperfect matching of Morlet wavelet, pulse width extends wider. However, Morlet WT gives prominence to low frequency. For contour and scalogram, time-frequency distribution is waterdrop-shaped, energy focus in low frequency in more time duration. Small energy of some frequency is lost but not important for time-frequency distribution. Different energy distribution in pulse width can be used for such short pulse detection.

In Figure 8, detailed energy distribution is presented. The peak energy can be seen not as a point but a circle in the contour. The central part has more energy when the contour lines get wider.

One time point corresponds with different frequencies and one frequency bin corresponds with different time points. Low frequency and high frequency with low energy is drop-out.

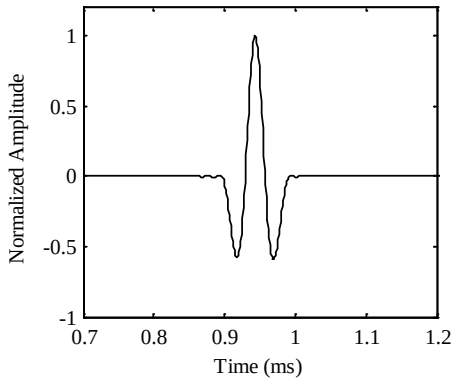



Figure 4. Simulated target echo without noise

### C. HHT Result

While local EMD is acted on underwater carrier-free echo pulse, HHT spectrum is shown in Figure 9. In this

part there are three main IMFs presented. Because EMD acts approximately as dyadic filter, peak frequency of third IMF is almost half of that of second IMF. IMFs with correlations are chosen to be viewed and prepared for detection.

HHT has the advantages of fine resolution, while HHT spectrum is approximately equivalent to contour information, checking fig9 and Figure 10. Different IMFs have same low frequency part and extraordinary high frequency part. As this matter, special fine resolution can be used for target recognition. In three consecutive IMF line, when short pulse appears, the frequency relative to time varies consecutively and differently. The frequency different between consecutive frequency point is large.

For HHT spectrum, the frequency bins separated in different IMFs. In each IMF, one time corresponds with one frequency. IMF is an intrinsic method to separate complex time and frequency correspondence appeared in other time-frequency methods. However, seriously the analysis loses high-frequency part.

Combining three time-frequency characteristics, a tri-channel detector can be employed for short pulse detection and recognition, with pulse width energy based on STFT, energy distribution in pulse width based on WT and precise recognition based on HHT.

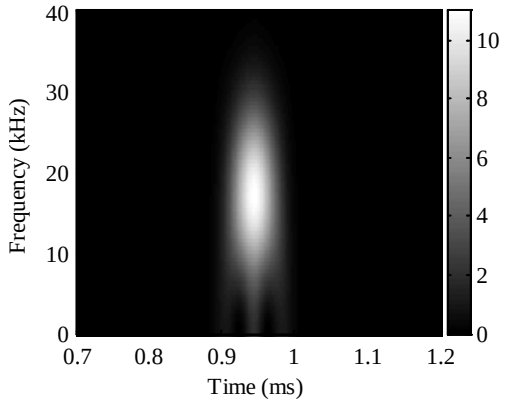



Figure 5. STFT spectrum

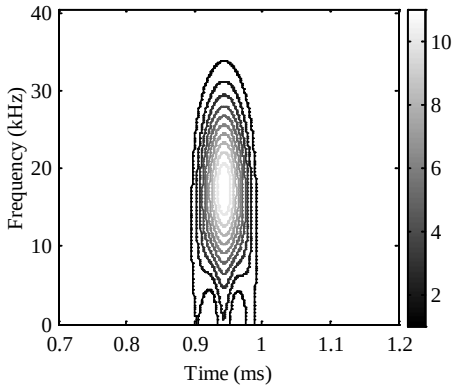



Figure 6. Contour of STFT spectrum

## V. TRI-CHANNEL DETECTOR

Time-frequency characteristic is presented in Section IV. With these characteristics, a tri-channel detector is developed to detect short pulse. The flow of the detector is demonstrated as Figure 11.

The criterion of decision device is as follows. Decision device 1 has the characteristics: the highest energy with consecutive frequency distribution and lower energy in the vicinity; in the signal frequency domain, very high and very low frequency still have part energy. Decision device 2 has the characteristics: the same highest energy as STFT, but low frequency energy is higher than low energy. Decision device 3 has the characteristics: like-pulse IMF and at pulse time the frequency changes not so as consecutive frequency change as STFT and WT. The decision devices are summation or product of transforms.

Single detection with STFT, WT and HHT with separate decision devices, the probability of detection (PD) comparing summation energy of different frequency points to the product ones is in Figure 12, Figure 13 and Figure 14 at different SNR. Line with ‘ * ’ dot presents the product detection, while line with ‘·’ presents the summation detection.

For single detection, STFT spectrum product detection has better performance than the other two methods at both summation and product detection. As is that frequency spectrum distributes widely in frequency domain of signal spectrum domain. Except for maximum energy and its vicinity, the spectrum of echo signal exists in lots of frequency band. More correlation frequency band product enhances the occurrence of echo.

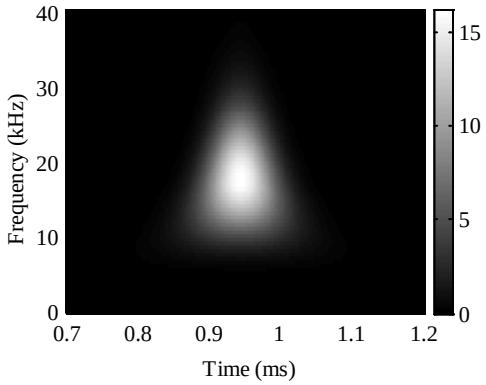



Figure 7. Scalogram of Morlet WT

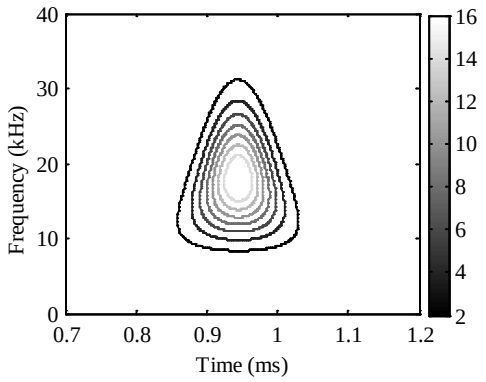



Figure 8. Contour of Scalogram of Morlet WT

WT spectrum product detection enhances not the same as STFT spectrum product detection, because the low frequency points extend energy domain, which brings the wider of pulse at low frequency point and the energy does not focus at maximum energy point with fake ones. WT method without modification become well when SNR>0, but WT method cost lots of calculation time.

HHT spectrum product method is not as good as that of STFT or WT even at low-SNR, because the processing is only based on transient signal. When underwater sound pulse is buried in noise, the short pulse cannot be seen as transient signal.

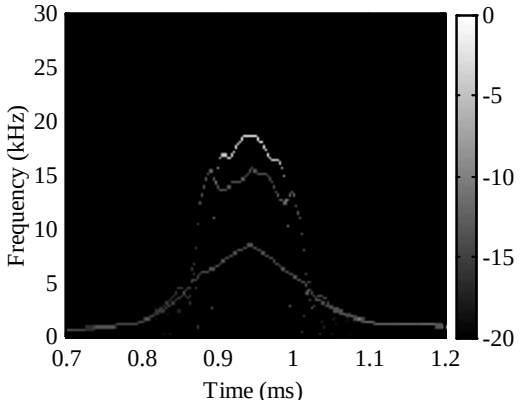



Figure 9. Contour of HHT spectrum

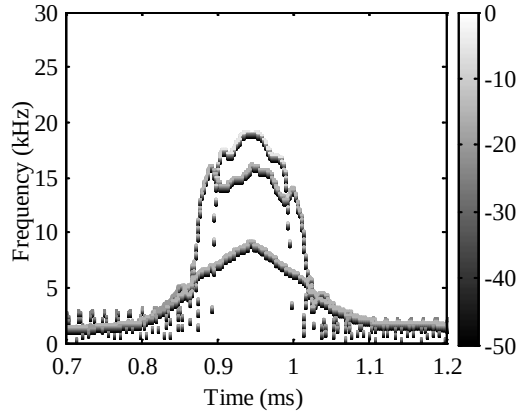



Figure 10. Flow of tri-channel detector

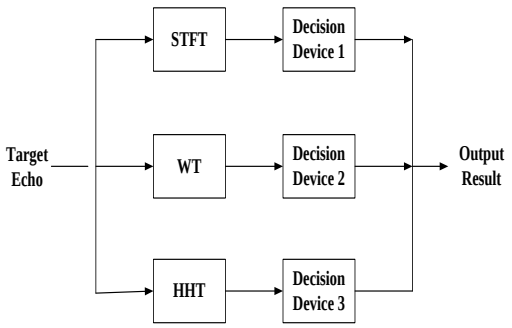



Figure 11. Flow of tri-channel detector

Combing the three time-frequency analysis methods with decision devices, the probability of detection is in Figure 15 at different SNR.

Totally the summation methods are no better than product methods, although when the SNR increases its PD increases rapidly. Because product of signal component by transform reduce the correlation of noise and keep the components' correlation.

It is not traditionally discovered that detection based on STFT is better than WT transform. We can find the reason from the figures given. Figure 16 is demonstrated that much frequency bins give contribution to energy detection.

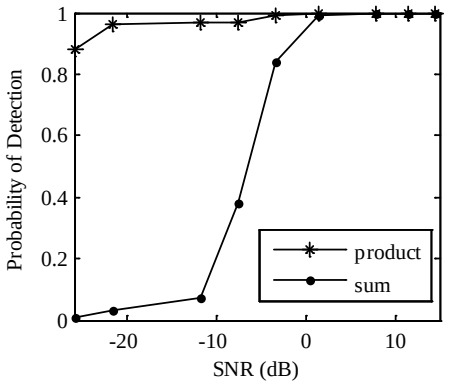



Figure 12. PD of sum and product STFT spectrum at different SNR

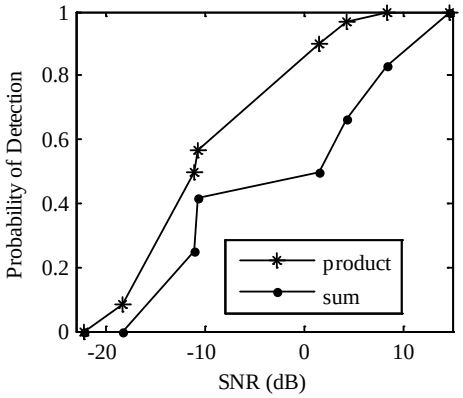



Figure 13. PD of sum and product WT spectrum at different SNR

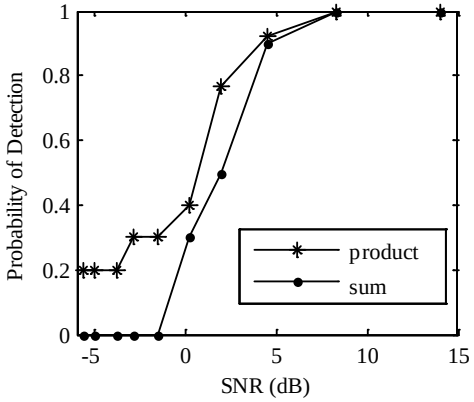



Figure 14. PD of sum and product HHT spectrum at different SNR

WT obtain higher energy at the maximum frequencies but perform worse than STFT. From Figure 17 and Figure 18, correlation coefficient of frequency bins at 0.1 kHz, 2 kHz, 4.3 kHz, 9 kHz with other frequency bins are presented. Correlation coefficient of STFT is larger than that of WT. Well number of IMFs is smallest comparing with the other method. So the STFT gets best

performance as a single detection method and raise detection probability at a whole, especially with low SNR.

From the correlation of different components by transform, it is known that more correlation components with correlation make detection method well.

## VI. CONCLUSION

Carrier-free short pulse generalized used in UWB radar is taken into underwater detection in high-resolution Sonar. Considering its transient relative wideband and short time duration, echo pulse of target at near field is simulated considering transmission loss of underwater channel.

It is still confirmed that short pulse is extended by transmission distance. Negative peak amplitude relative to positive peak amplitude is increasing visible. Simulated target echo is more practical for ocean environment. To get more information for detection, time-frequency analysis of such short pulse is presented in this paper to get precise result.

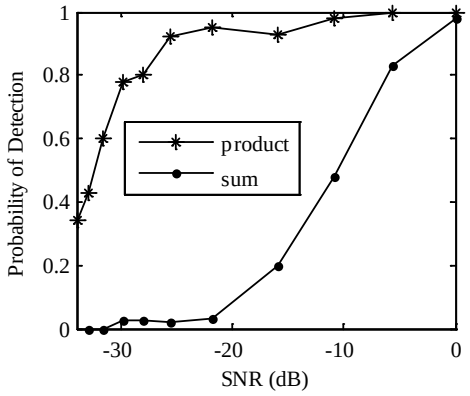



Figure 15. The tri-channel detector PD at different SNR

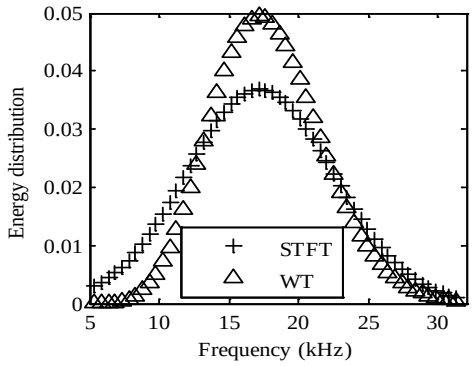



Figure 16. Energy distribution at different frequency bins

After time-frequency analysis of short pulse, it is demonstrated the characteristics of three time-frequency analysis method, such as Short Time Fourier Transform, Wavelet Transform and Hilbert Huang Transform. With these characteristics, the decision devices are obtained with different criterion. Figure s are viewed to explain the different performance of single detectors because of correlation between frequency bins. STFT detector gets best performance although it is kind of uniformed filter.

A tri-channel detector is established to detect such short pulse. Comparing with single detector and tri-channel detection, probability of detection gives the performance of such detector.

It is demonstrated that more channels with different time-frequency methods merge more information and overcome the limitation of single detector. Although it brings complexity of systems, recent electronics development makes it practical and real-time.

With basic transform without modification, probability of detection works well. Furthermore, with modification, it is probable to obtain better performance of detector, which can be investigated in the future work.

## ACKNOWLEGEMENT

Thanks for assistant from the teacher Hu Yang in theory and colleague Daojiang Li in discussion. The work is supported by the Office of College of Marine Code 624.

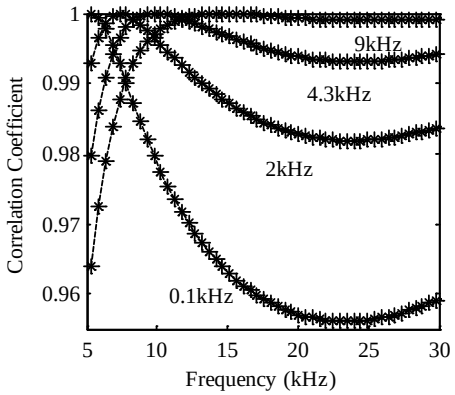



Figure 17. Correlation coefficient separately between 1kHz, 5kHz, 10kHz, 20kHz and other frequency bins by STFT

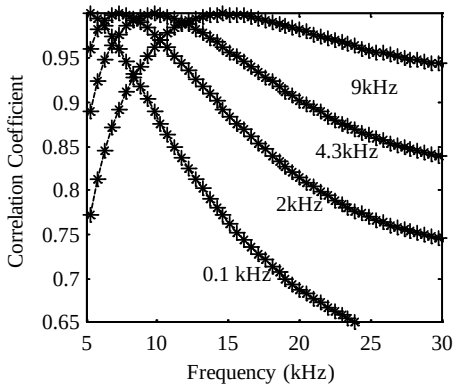



Figure 18. Correlation coefficient separately between 1kHz, 5kHz, 10kHz, 20kHz and other frequency bins by WT

## REFERENCES

[1] M. Z. Win, R. A. Scholtz, “Impulse radios: how it works”, IEEE Communication letters, vol. 2, pp. 36–38, 1998.

[2] H. Chen, Y. S. Chen, H. Yang and Y. L. Ni, “A method using stochastic derivative for analyzing fluctuation of reverberation envelop”, Sci. China Ser. G., vol. 39, pp. 1584–1588, 2009.

[3] Y. S. Chen, “Research on torpedo homing with Ulra Wideband pulse”, unpublished.

[4] F. E. Aranda, N. Brown, H. Arslan, “Rake receiver finger assignment for Ultra - wideband radio”, Signal Processing

Advances in Wireless Communications. SPAWC 2003, 4th IEEE Workshop on, pp. 239–243, 2003.
[5] Z. H. Jing, J. Hua, X. N. Yang, "A new method for detecting negative SNR UWB Signal", Communication Countermeasures. vol. 4, pp. 17–21, 2007.
[6] Q. T. Zhang and S. H. Song, "Eigen-Based Receivers for the Detection of Random UWB Signals", IEEE Transactions on Communications, vol. 54, no. 7, pp. 1184–1189, July 2006.
[7] X. T. Cheng, W. L. Zhu, "A subspace detection method of space-time block codes for Ultra-wideband communications", Journal of Electronics & Information Technology, vol. 28, no. 7, pp. 1295–1297, July 2006.
[8] X. Dai, G.Y. Fang, "Cramer-Rao lower bound of vital signal frequency detection for Ultra Wideband Radar," Journal of Electronics and Information Technology, vol. 33, pp. 701–705, 2011.
[9] X. Y. Zhang, Y. L. Liu, K. Wang, "Ultra Wide-Band channel estimation and signal detection through compressed sensing", Journal of Xi'an Jiao Tong University, vol. 44, pp. 88–91, Feb. 2010.
[10] M. Li, L. Yao, "Detection of transient signal based on short-time correlation", Proceedings of National Acoustics Conference, China Institute of Acoustics Press, pp. 101–102, 2002.
[11] J. Wang, Z. Z. Huang, W.D. Wang, D. J. Wang, "A new peak detection UWB impulse radio receiver", Journal of University of Science and Technology of China, vol. 38, pp. 1168–1172, 2008.
[12] I. Paraskevas, A. Kokkosis, I. Liaperdos and M. Rangoussi, "Detection of transient events using the Hartley Phase Cepstrum for improved power quality", 7th Mediterranean Conference and Exhibition on Power Generation, Transmission, Distribution and Energy Conversion, Agia Napa, Cyprus, pp. 1–5, November 2010.
[13] H. P. Wu, W. W. Liu, J. J. Lou, X. Q. Wang, "Application of chaos in Sonar detection", Mechanics Automation and Control Engineering (MACE), 2011 Second International Conference, pp. 4774–4777, July 2011.
[14] X. M. Zhang, C. Q. Cai, J. H. Zhang, "A transient signal detection technique based on flatness measure", the 6th International Conference on Computer Science & Education (ICCSE 2011), SuperStar Virgo, Singapore, pp. 310–312, August 2011.
[15] P. Kathirvel, M. S. Manikandan, P. Maya and K. P. Soman, "Detection of power quality disturbances with overcomplete dictionary matrix and ℓ1-norm minimization", Power and Energy Systems (ICPS), 2011 International Conference on, pp. 1–6, Dec. 2011.
[16] M.A.S. Masoum, S. Jamali, N. Ghaffarzadeh, "Detection and classification of power quality disturbances using discrete wavelet transform and wavelet network", IET Sci. Meas. Technol., vol. 04, pp. 193-205, 2010.
[17] B.J. Alllen, R.L. Rabiner, "A unified approach to Short-Time Fourier analysis and synthesis", Proceeding of the IEEE, vol. 65, pp. 1558--1564, 1977.
[18] O. Rioul, M. Vetterli, "Wavelets and signal processing", IEEE Signal Processing Magazine, vol. 8, pp. 14–38, 1911.
[19] N.E. Huang, Z. Shen, S.R. Long,, M.C. Wu, H.H. Shih, Q. Zheng, et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", Proc. R. Soc. Lond. A., vol. 454, pp. 903–995, 1998.
[20] G. Rilling, P. Flandrin, P. Gonçalv`es, "On Empirical mode decomposition and its algorithm", Proceedings of the 6th IEEE/EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP \'03), Grado, Italy, 2003.
[21] G. F. Bin, J. J.Gao, X. J. Li, B. S. Dhillon, "Early fault diagnosis of rotating machinery based on wavelet packets—Empirical mode decomposition feature extraction and neural network", Mechanical Systems and Signal Processing, vol. 27, pp. 696–711, 2012.
[22] M.-G.M. Hussain, "Principles of high-resolution Radar based of nonsinusoidal wave--part I: signal representation and pulse compression", IEEE Trans. Electromagn. Compat., vol. 31, pp. 359–367, 1989.
[23] Y. L. Ni, H. Chen, "High-resolution Sonar based on carrier-free narrow pulse: I. transmission characteristic", Proceedings of Acoustics 2012 Hong Kong, in press.

**Yunlu Ni** was born in China in 1983. She received a B.Sc. degree in applied physics and M.Sc. degree in radio physics from China University of Petroleum in 2005 and 2008 respectively. Now she is applying for a doctor degree in Northwestern Polytechnical University. Her major is sound information and signal processing.

**Hang Chen** was born in China in 1956. He received the M.Sc. degrees and Ph.D. degrees in sound engineering from Northwestern Polytechnical University, Xi'an, China, in 1989 and 2004 respectively.

Since 1984, he has been active in sound engineering for over twenty years, with the past research in sound theory and current interests in sound information and signal processing.

Since 2000, Dr Chen has been member of Shaanxi Acoustics Institute.

# Speed Sensorless Control of PMSM using Model Reference Adaptive System and RBFN

Wei Gao
Naval Bengbu Petty Officer Academy/Department of Weaponry Engineering, Bengbu, China
Email: hjgaowei.@126.com

Zhirong Guo
Naval Bengbu Petty Officer Academy/Department of Weaponry Engineering, Bengbu, China
Email: guo1976.love@163.com

***Abstract*—In the speed sensorless vector control system, the amended method of estimating the rotor speed about model reference adaptive system (MRAS) based on radial basis function neural network (RBFN) for PMSM sensorless vector control system was presented. Based on the PI regulator, the radial basis function neural network which is more prominent learning efficiency and performance is combined with MRAS. The reference model and the adjust model are the PMSM itself and the PMSM current, respectively. The proposed scheme only needs the error signal between *q* axis estimated current and *q* axis actual current. Then estimated speed is gained by using RBFN regulator which adjusted error signal. Comparing study of simulation and experimental results between this novel sensorless scheme and the scheme in reference literature, the results show that this novel method is capable of precise estimating the rotor position and speed under the condition of high or low speed. It also possesses good performance of static and dynamic.**

***Index Terms*—permanent magnet synchronous motor, vector space control, sensorless, model reference adaptive system, radial basis function neural network**

## I. INTRODUCTION

Closed-loop speed control in high performance permanent magnet synchronous motor vector control system is essential. There are two ways to achieve speed detection, one is to mount the sensor directly to speed detection and the other is not installed sensors, but indirectly through the motor parameters to estimate and detect rotor speed. Objective point of view, speed sensor makes the design simple, accurate signal, it should be preferred PMSM system design. But in some special occasions and field of application, whether to install the sensor has become a prominent issue. How to adapt to the working environment, increasing system reliability has become a system research and design problems. In order to overcome the mechanical sensor the flaw, to research and develop the control method with high reliable non-mechanical sensor, becomes one of electrical machinery control area of technology research hot spots.

At present, the sensor PMSM vector control has been proposed in a number of ways to estimate the motor rotor position and speed. In [1], direct calculation formula of calculation method of rotor position angle was given. Rotor position and speed are obtained by actual measurement. Calculation process is simple and direct, without the use of complex convergence algorithms, faster dynamic response. Due to the methods used in the calculation of the differential current, measured error significant impacts on the accurate observation of the rotor position. Moreover, it is an open-loop algorithm, which is not possible to get the correct estimated results by the noise or parameter changes. By the phase voltage and phase current of motor, the phase induction of the stator is calculated in real time [2-4]. By comparing the calculated inductance value and the actual measured inductance value, the estimated rotor position is calculated. In the transient and at low speed, it is difficult to measure the back EMF accurately, so there is the error in the inductance value, resulting in the error of estimated rotor position. In addition, estimated speed is affected by the parameter change, especially when the inductance is saturated, the value of the angle will be a great error by table look-up. When the motor steady-state is running, the stator and rotor flux synchronous rotate. By calculating the phase angle of stator flux, the rotor position is estimated. Using voltage equation, stator magnetic flux can be evaluated by the back-EMF [5]. But this method is dependent on the motor parameters, when the motor parameters changing caused by the temperature change, the magnetic saturation effects, the accuracy of the speed observer is dropping. In the stator two-phase coordinate system, the voltage equation of the salient PMSM is more complex than the one of the non-salient PMSM. The salient pole PMSM methods of observation based on the speed of the back-EMF cannot be directly applied to the salient pole PMSM. In [6-8], the adaptive method based on the extended back-EMF to observe the motor speed was proposed; the simulation and the experimental results have proven the effectiveness of the method. By constructing the back-EMF, the motor model cans be seen as the unified mathematical model of both the salient pole PMSM and the non-salient pole PMSM, which apply the unified method to estimate the speed when the motor is running with high speed. But the


doi: 10.4304/jnw.8.1.213-220

extended EMF contains the $i_d$ and $i_q$ under the rotor synchronous coordinate, which are changing in the motor dynamic process. So the extended back EMF is not constant value, the estimated rotor speed and position is not accurate. In [9-11], the method using saliency-tracking method can be used for a wide range of speeds and even better results can be achieved at low speeds, but the speed is estimated dependent on the motor saliency effects. The Kalman filter law estimates the rotational speed, which is complex, and causes the structure the adjustment and the parameter design quite is all difficult [12-15]. Moreover, the Extended Kalman Filter is on the application of PMSM sensorless control system. The simulation model of position sensorless speed control system is established. Tracking characteristics of the motor speed and rotation were simulated, adopting the attenuation factor to improve the speed of fast-track performance. However, the method in not applicable to the closed-loop speed control system with the larger viscosity coefficient. Model reference adaptive system (MRAS) is a relatively common method to estimate the rotor position and speed, which is simple, easy to achieve in the digital control system. The MRAS method of reference model and adjustment model based on the stator flux vector is adopted [16-18], which calculating the stator flux linkage separately with the voltage model and the current model, using the adaptive algorithm to adjust above two kind of model computation the stator flux linkage, and observes the electrical machinery rotational speed. Disadvantage of the method is the accuracy of the observations dependent on the accuracy of the motor parameters, particularly on voltage model, motor temperature rise of stator resistance significant impact on calculation of stator flux of change.

The novel method based on the documents [19-20] proposed the estimated speed based on MRAS and RBFN, which estimates the position and speed for PMSM sensorless vector control system. The Proposed method adopts PMSM itself as a reference model and the PMSM current model as the adjustable model. The error signal which is difference between *q*-axis estimated current and *q*-axis actual current fed into the RBFN regulator to obtain estimated speed. Simulation and experimental results verify that the method presented in this paper the effectiveness and viability.

## II. Speed Sensorless Control using MRAS and RBFN

### A. Model Reference Adaptive System

Model reference adaptive control system scheme was first proposed by Professor Witark of the United States in 1958. The main characteristic of such system is the use of the reference model, which provides for the performance required by the system. Model reference adaptive control and self-adaptive control are very different in principle and structure. This kind of system performance requirement is not expressed with a target function, but is expressed with the output of reference model or the condition response. Reference model or the status of the output is equivalent to a given dynamic performance, by comparing the output of a controlled object and reference model or the state respond to obtain error information.

According to certain law (adaptive law) to amend the actual system parameters (parameter adaptive), so that actual output or system state as far as possible follow the output or state of the reference model. Parameters of the amended law or auxiliary input signal are generated by the application of agency. In 1989, the scholars used the model self-adaptive system to estimate the speed for the first time, its basic structure as shown in Fig. 1.

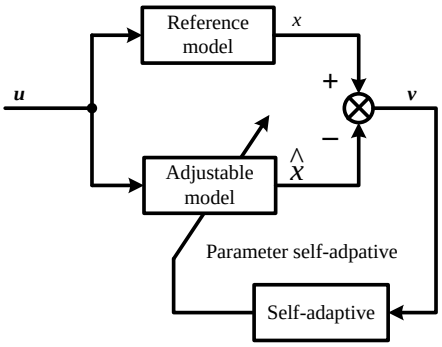



Figure 1. The basic construction of MRAS.

In Fig. 1, reference model and adjustable model (adaptive model) are incited by external input at the same time, $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ are the state vector of the reference model and adjustable model, respectively. The reference model with its state (or output) provides for a given performance, which is comparing with the measured performance $\hat{\boldsymbol{x}}$ of adjustable system. The difference vector $\boldsymbol{v}$ input adaptive mechanism to modify the parameters of adjustable model. Its state $\hat{\boldsymbol{x}}$ is fast and stable to approach the state $\boldsymbol{x}$, which make the difference $\boldsymbol{v}$ approaches zero. One of the key issues to constitute a high quality adaptive control system is the implementation of adaptive law in the adaptive mechanism in Fig. 1.

There are optimal design methods of local parameters, the Lyapunov Stability Theory and the Popov Super stability design method about the design of the adaptive law, which are complex, therefore very difficult to obtain the widespread promotion in reality. The neural network has the characteristics of self-learning, adaptive capacity and the ability to fully approaching the arbitrarily complex non-linear relationship, so neural network combines with the model reference adaptive system to form the neural network reference adaptive control.

### B. Model Reference Adaptive System Based on RBFN

In the rotor rotating reference frame, the PMSM stator current model is described as follows:
where

$$\frac{di_d}{dt} = -\frac{R_s}{L} i_d + \omega i_q + \frac{u_d}{L} \tag{1}$$

$$\frac{di_q}{dt} = -\frac{R_s}{L} i_q - \omega i_d - \frac{\psi_f}{L}\omega + \frac{u_q}{L} \tag{2}$$

where

$i_d$, $i_q$ *d*-, *q*-axis stator current;

$u_d$, $u_q$ *d*-, *q*- axis stator voltage;

$R_s$ stator resistance;

$L$ stator inductance;

$\psi_f$ rotor permanent magnet flux.

In [8], by estimating *d*-, *q*- axis currents, measured the estimated speed using MARS is:

$$\hat{\omega} = \int_0^t k_1 \left[ (i_d \hat{i}_q - i_q \hat{i}_d - \frac{\psi_f}{L}(i_q - \hat{i}_q) \right] d\tau + k_2 \left[ (i_d \hat{i}_q - i_q \hat{i}_d - \frac{\psi_f}{L}(i_q - \hat{i}_q) \right] + \hat{\omega}(0) \tag{3}$$

where

$\hat{\omega}$ estimated rotor angular speed;

$\hat{i}_d$, $\hat{i}_q$ *d*-, *q*- stator axis estimated current;

$k_1$, $k_2$ PI regulator coefficients;

$i_d$, $\hat{i}_d$ are *d*- axis feedback current and estimated current respectively, which are the number of non-zero.

But for the $i_d = 0$ vector control strategy of PMSM system, $i_d$, $\hat{i}_d$ can be approximated to zero.

From the mathematical equations of the motor, motor speed is only related to the equation (2). Therefore, according to the equation (2), adjustable model can be gotten and the actual motor body is as the reference model. With parallel structure, estimated speed can be formulated as

$$\frac{di_q}{dt} = -\frac{R_s}{L} i_q - \frac{\psi_f}{L}\omega + \frac{u_q}{L} \tag{4}$$

$$\frac{d\hat{i}_q}{dt} = -\frac{R_s}{L} \hat{i}_q - \frac{\psi_f}{L}\hat{\omega} + \frac{u_q}{L} \tag{5}$$

Equation (5) is discretized and $T_s$ is sample time.

$$\frac{d\hat{i}_q}{dt} = \frac{\hat{i}_q(k) - \hat{i}_q(k-1)}{T_s} = -\frac{R}{L}\hat{i}_q(k-1) - \frac{\psi_f}{L}\hat{\omega}(k) + \frac{u_q(k-1)}{L} \tag{6}$$

Thus, from (***k-1***)th the sampled data, estimated speed can be formulated as

$$\hat{i}_q(k) = \alpha \hat{i}_q(k-1) + \beta u_q(k-1) - \gamma \hat{\omega}(k) \tag{7}$$

where $\alpha = 1 - \frac{T_s R}{L}$, $\beta = \frac{T_s}{L}$, $\gamma = \frac{T_s \psi_f}{L}$.

In the (***k-1***)th sample time of actual control, the stator axis voltage $u_q$ can be calculated. The estimated current $\hat{i}_q(k)$ can be measured by equation (7), and the actual current $i_q(k)$ can be measured in the ***k***th sample time.

If the deviation of the rotor estimated speed $\hat{\omega}$ and the actual speed $\omega$ is large, the deviation between the estimated current $\hat{i}_q(k)$ and the actual current $i_q(k)$ will be obvious when the motor is running.

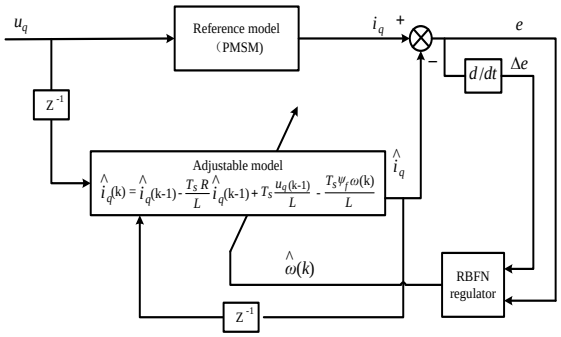



Figure 2. Model reference adaptive system based on RBFN.

Therefore takes the RBFN regulator using the deviation and the deviation rate of change the input, namely

$$e = i_q(k) - \hat{i}_q(k) \tag{8}$$

$$\Delta e = e(k) - e(k-1) \tag{9}$$

The regulator is constituted to get the MRAS based on equation (8) and (9). PMSM itself is as the reference model, and equation (7) is as an adjustable model. Speed is the adaptive variable $\hat{\omega}$, and *q*-axis voltage is the combination of inputs.

The stator *q*-axis measured current is as the output of the reference model, and the calculated value is as the output of the adjustable model. After comparing, the result is as the input of RBFN regulator to adjust the adaptive parameters $\hat{\omega}$.

The system block diagram is shown in Fig. 2.

### *C. Structure and Algorithms of Radial Basis Function*

Due to the wide range of adaptability and leaning ability, artificial neural networks have been widely used in the prediction of the nonlinear system.

The feed forward neural network is the most widely used artificial neural network. At present, the BP neural network to get more applications in the earthquake prediction, but the convergence of BP neural network learning process is closely related with the initial value. The radial basis function neural network is a good performance of feed forward artificial neural network.

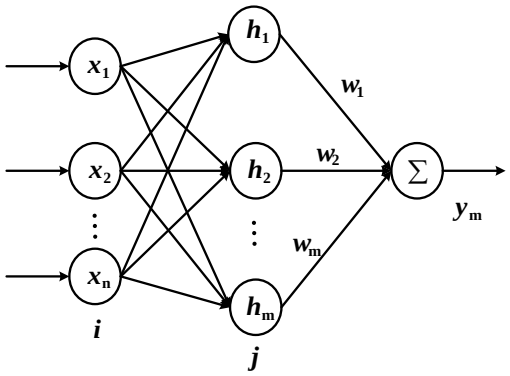



Figure 3. RBF neural network structure.

Radial basis function neural network is three-layer feed forward network, which is a single hidden layer. From the input to the output mapping is nonlinear whereas from the

hidden layer space to output space mapping is linear, so it can greatly accelerate the learning speed and avoid local minima.

RBF network structure is shown in Fig. 3.

In the RBF network structure, $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$ is the input vector. In this paper, there are two inputs, which are $q$-axis current deviation $e$ and deviation rate of change $\Delta e$.

The Gaussian function is adopted as the membership function as follows:

$$h_j=\exp\left(-\frac{\left\|\boldsymbol{X}-\boldsymbol{C}_j\right\|^2}{2b_j^2}\right),\ j=1,2,\cdots,m \tag{10}$$

where the $\boldsymbol{B}=[b_1,b_2,\cdots,b_m]^{\mathrm{T}}$ and $\boldsymbol{C}_j=[c_{j1},c_{j2},\cdots,c_{jn}]^{\mathrm{T}}$ are the standard deviation and mean, respectively.

The Input layer to hidden layer weight value is 1.0, and the weight vector from hidden layer to output layer is

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_m]^{\mathrm{T}} \tag{11}$$

The RBF network output is

$$y_m(k)=w_1h_1+w_2h_2+\cdots+w_mh_m \tag{12}$$

The energy function $E$ is defined as

$$E(k)=\frac{1}{2}\left[i_q(k)-\hat{i}_q(k)\right]^2=\frac{1}{2}e(k)^2 \tag{13}$$

And the learning algorithms about the weight value, standard deviation and mean are as follows:

$$w_j(k)=w_j(k-1)+\eta\left[i_q(k)-\hat{i}_q(k)\right]h_j+\alpha\left[w_j(k-1)-w_j(k-2)\right] \tag{14}$$

$$\Delta b_j=\left[i_q(k)-\hat{i}_q(k)\right]w_jh_j\frac{\left\|\boldsymbol{X}-\boldsymbol{C}_j\right\|^2}{b_j^3} \tag{15}$$

$$b_j(k)=b_j(k-1)+\eta\Delta b_j+\alpha\left[b_j(k-1)-b_j(k-2)\right] \tag{16}$$

$$\Delta c_{ji}=\left[i_q(k)-\hat{i}_q(k)\right]w_jh_j\frac{x_i-c_{ji}}{b_j^2} \tag{17}$$

$$c_{ji}(k)=c_{ji}(k-1)+\eta\Delta c_{ji}+\alpha\left[c_{ji}(k-1)-c_{ji}(k-2)\right] \tag{18}$$

where $\eta$ is the learning rate and $\alpha$ momentum factor.

## III. Simulation Results

Matlab is both a powerful computational environment and a programming language that easily handles matrix and complex arithmetic. It is a large software package that has many advanced features built-in, and it has become a standard tool for many working in science or engineering disciplines. Matlab simulation is to verify the proposed method, which is comparison with the method in the literature [19]. In this paper, the method proposed this paper is method 1 and the method proposed the literature [19] is method 2.

The simulations were made in Matlab-Simulink environment, for a PMSM, with the following parameters:

Rated torque: $T_n$=3$Nm$;

Number of pole pairs: $P$=2;

Stator resistance: $R_s$=2.875 Ω;

$d$ axis inductances: $L_d$= 8.5$mH$;

$q$ axis inductances: $L_q$=12.5 $mH$;

Flux induced by magnets: $\Psi_m$=0.175$Wb$.

In the simulation, three cases were simulated.

Case 1: start form the static to 1000r/min.

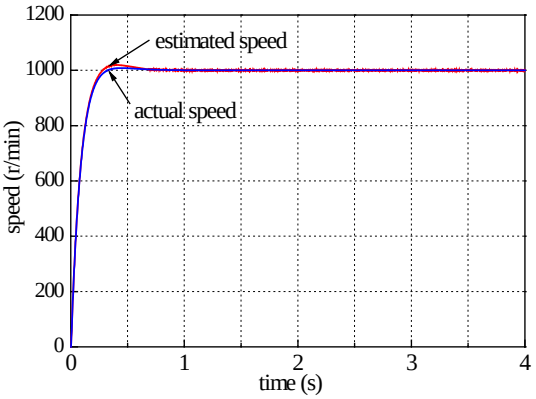
Case 2: start form the static to 50r/min.

Case 3: form the 500r/min to 1000r/min.

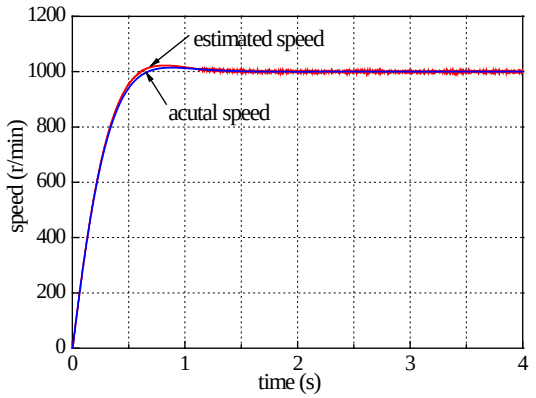
There are high speed and low speed situations in the two cases. In each case, the estimated speed and actual speed curves of the two methods were given. And the error curves of the two methods were given.

### A. Staring Process from the Static to 1000r/min

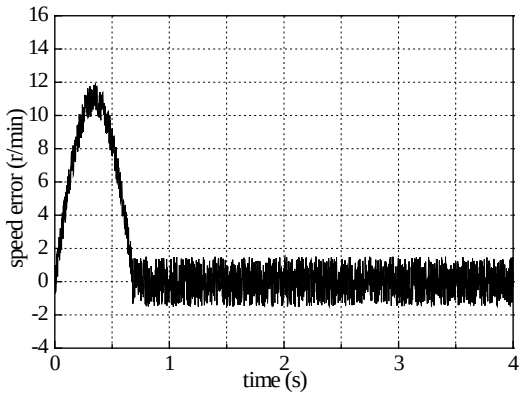
Fig. 4 shows the simulation wave of form static to 1000r/min in sensorless control.



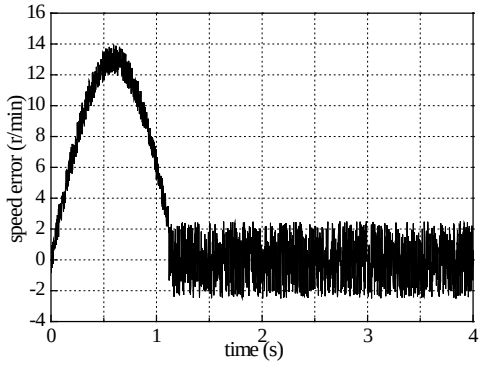

(a) Actual and estimated speed with method 1.




(b) Actual and estimated speed with method 2.




(c) Speed error curve with method 1.




(d) Speed error curve with method 2.

Figure 4. Simulation results from static to 1000r/min

Fig. 4(a) is the comparison of the actual speed and estimated speed with method 1.

Fig. 4(b) is the comparison of the actual speed and estimated speed with method 2.

Fig. 4(c) and Fig. 4(d) are the error curve in method 1 and method 2, respectively.

In method 1, it is about 0.9s from start to 1000r/min, while the estimated speed tracking time is 0.68s. The estimated error does not exceed 2r/min.

But in method 2, it is about 1.5s from start to 1000r/min, while the estimated speed tracking time is 1.12s. The estimated error does not exceed 3r/min.

From the graph and data analysis, faster tracking speed and higher estimation accuracy with method 1 are than with method 2.

### *B. Staring Process from the Static to 50r/min*

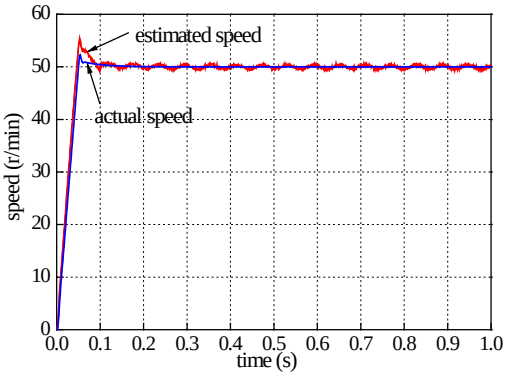
Fig. 5 shows the simulation wave of from static to 50r/min in sensorless control.

Fig. 5(a) is the comparison of the actual speed and estimated speed with method 1.

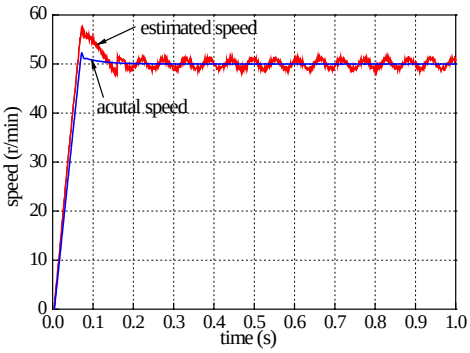
Fig. 5(b) is the comparison of the actual speed and estimated speed with method 2.

Fig. 5(c) and Fig. 5(d) are the error curve in method 1 and method 2.

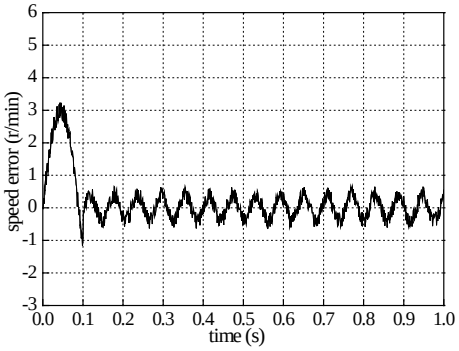
In method 1, it is about 0.15s from start to 50r/min , while the estimated speed tracking time is 0.1s. The estimated error does not exceed 1r/min.



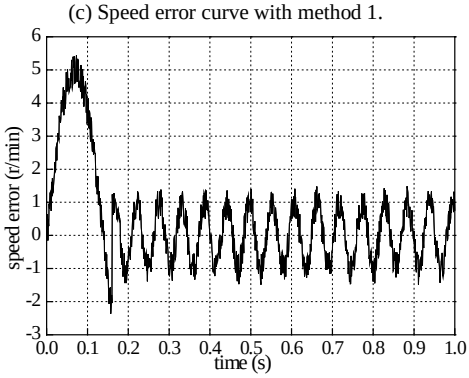

(a) Actual and estimated speed with method 1.




(b) Actual and estimated speed with method 2.




(c) Speed error curve with method 1.




(d) Speed error curve with method 2.

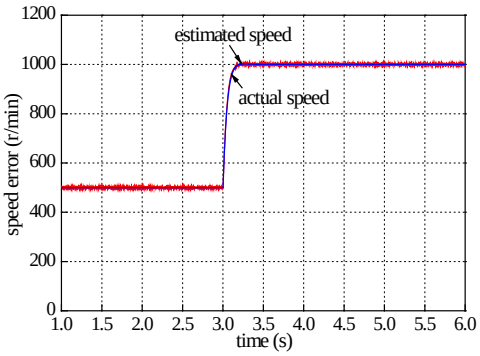
Figure 5. Simulation results from static to 50r/min.

But in method 2, it is about 0.22s from start to 50r/min, while the estimated speed tracking time is 0.16s. The estimated error does not exceed 2r/min.

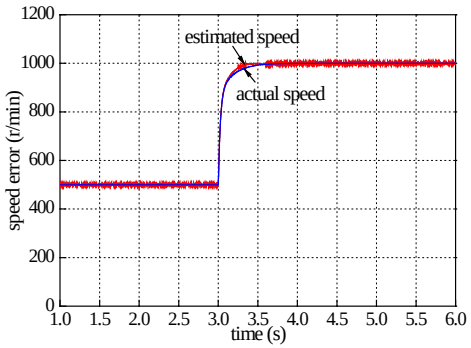
From the graph and data analysis, faster tracking speed and higher estimation accuracy with method 1 are than with method 2.

### *C. Dyanmic Process form 500r/min to 1000r/min*

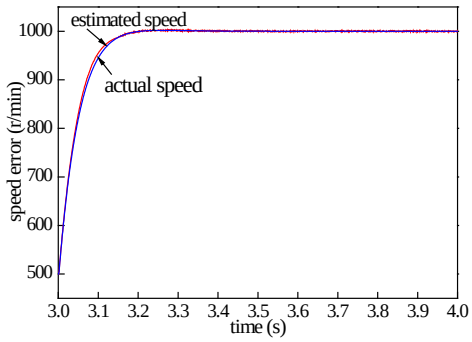
In order to compare the dynamic performance of the two methods, the simulation of speed change form 500r/min to 1000r/min was carried out. Fig. 6 shows the simulation wave of form 500r/min to 1000r/min in sensorless control. Fig. 6(a) is the comparison of the actual speed and estimated speed with method 1.



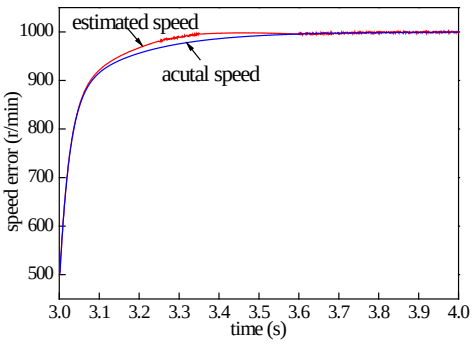

(a)Actual and estimated speed with method 1.




(b) Actual and estimated speed with method 2.




(c) Enlarged dynamic waveform with method 1.




(d) Enlarged dynamic waveform with method 2.

Figure 6. Simulation results form 500r/min to 1000r/min.

Fig. 6(b) is the comparison of the actual speed and estimated speed with method 2.

Fig. 6(c) and Fig. 6(d) are the enlarged dynamic waveforms in method 1 and method 2. The dynamic response time of motor is about 0.3s in method 1, while the time is 0.6s in method 2. The dynamic response in method 1 is quicker, moreover the accuracy of estimated speed in method 1is higher in the dynamic process.

## IV. Experimental Results

In order to examine basic performance of the proposed sensorless vector control method using the model reference adaptive system and RBFN, extensive experiments were carried out using the 600W salient pole PMSM. The control algorithm of the motor drive is implemented in a digital signal processor (DSP), which can be programmed using C language. The sampling frequency and the pulse width modulation (PWM) switching frequency of the system is 5 kHz.

Experimental results are to verify the proposed method, and the results with method in the literature [19] are compared. In this paper, the method proposed this paper calls method 1 and the method proposed the literature [19] calls method 2.

In the experiment, three cases are tested.

Case 1: start from the static to 1000r/min.

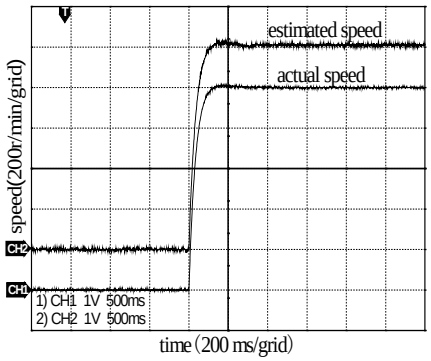
Case 2: start from the static to 100r/min.

Case 3: form 500r/min to 1000r/min.

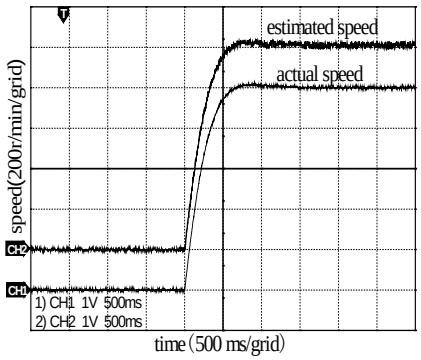
There are high speed, low speed situations and speed change in the three cases. In each case, the estimated speed and actual speed curves of the two methods were given.

### *A. Starting Process from the Static to 1000r/min*

Fig. 7 shows the experimental waveform of from static to 1000r/min with sensorless control.




(a) Actual and estimated speed with method 1.




(b) Actual and estimated speed with method 2.

Figure 7. Experimental results of sensorless at 1000r/min.

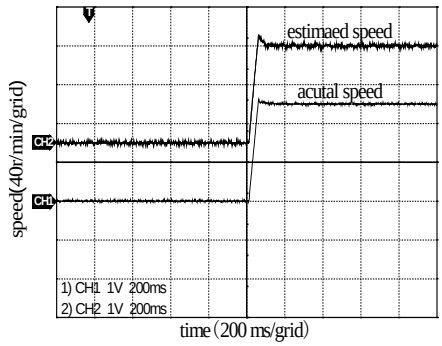
Fig. 7(a) is the comparison of the waveforms of the actual measured speed and estimated speed with method 1. Fig. 7(b) is the comparison of the waveforms of the actual measured speed and estimated speed with method 2.

By analyzing the waveforms, the error between the estimated speed and the actual speed with method 1 is smaller than with method 2 in the process of dynamic. The experimental results are the same as the simulation results.

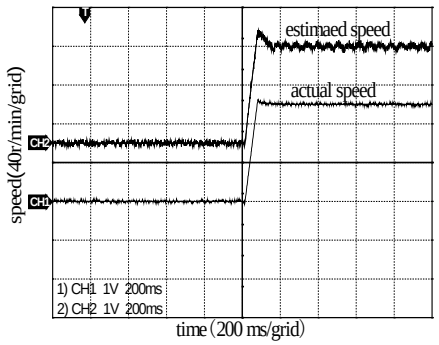
### *B. Starting Process from the Static to 100r/min*

Fig. 8 shows the experimental waveform of from static to 100r/min with sensorless control.

Fig. 8(a) is the comparison of the waveforms of the actual measured speed and estimated speed with method 1. Fig. 8(b) is the comparison of the waveforms of the actual measured speed and estimated speed with method 2. By analyzing the waveforms , Two kinds of methods in 100r/min can still identify the motor speed . The error between the estimated speed and the actual speed with method 1 is smaller than with method 2 in the process of dynamic. The experimental results are the same as the simulation results.




(a) Actual and estimated speed with method 1.



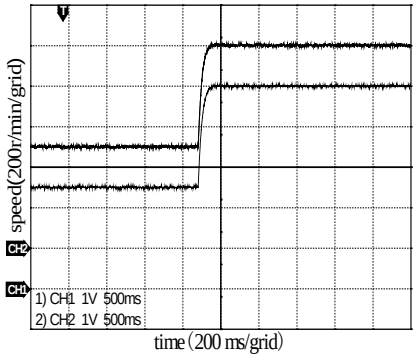

(b) Actual and estimated speed with method 2.

Figure 8. Experimental results of sensorless at 100r/min.

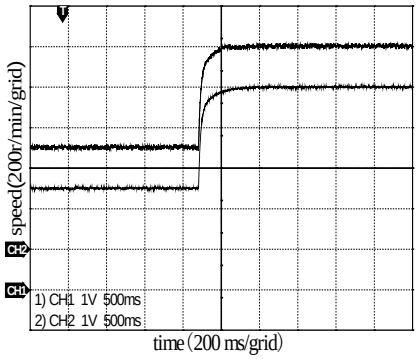
### *C. Dynamic Process from 500r/min to 1000r/min*

Fig. 9 shows the experimental waveform of from 500r/min to 1000r/min with sensorless control.

Fig. 9(a) is the comparison of the waveforms of the actual measured speed and estimated speed with method 1. Fig. 9(b) is the comparison of the waveforms of the actual measured speed and estimated speed with method 2. The dynamic response with method 1 is faster 40ms than with method 2, and the error with method 1 is smaller than with method 2. Because the base of these two kinds of identification methods is the rotor speed is constant, in the dynamic process two methods cannot guarantee the estimated speed is consistent with the actual speed and the process is gradual convergence.




(a) Actual and estimated speed with method 1.




(b) Actual and estimated speed with method 2.

Figure 9. Experimental results of sensorless from 500r/min to 1000r/min.

## V. CONCLUSION

This paper presents a model reference adaptive speed identification scheme, which is based on the existing literature [19]. It only needs the error signal between *q* axis estimated current and *q* axis actual current. Then estimated speed is gained by using RBFN regulator which adjusted error signal. Comparative study of simulation and experimental results between this novel sensorless scheme and the scheme in reference literature [19] are presented in this paper. The results show that this novel method is capable of precise estimating the rotor position and speed under the condition of high or low speed. It also possesses good performance of static and dynamic.

## ACKNOWLEDGMENT

The authors would like to acknowledge the Qun Pang. This work was supported in part by a grant from Zhenyun Pang.

## REFERENCES

[1] O. Wallmark, S. Lundberq, and M. Bonqrorno, “Input admittance expressions for field-oriented controlled salient

PMSM Drives", *IEEE. Trans. Power. Electron.*, vol. 27, pp. 1514–1520, March 2012.
[2] H. Kim, J. Son, and J. Lee, "A high speed sliding mode observer for the sensorless speed control of PMSM" , *IEEE. Trans. Ind. Electron.*, vol. 58, pp. 4069–4077, September 2011.
[3] L. Changsheng and M. Elbuluk, "A sliding mode observer for sensorless control of permanent magnet synchronous motors", *IEEE IAS Annual Meeting*, vol. 2, pp. 1273-1278, September 2001.
[4] M. J. Corley and R. D. Lorenz, "Rotor position and velocity estimation for a salient-pole permanent magnet synchronous machine at standstill and high speeds", *IEEE Trans. on Industry Applications*, vol. 34, no. 4, pp. 784-789, August 1998.
[5] Y. S. Han, J. S. Choi, and Y. S. Kim, "Sensorless PMSM drive with a sliding mode control based adaptive speed and stator resistance estimator", *IEEE Trans. on Magnetics*, vol. 36, no. 5, pp. 3588-3591, September 2000.
[6] K. Paponpen and M. Konghirun, "An Improved sliding mode observer for speed sensorless vector control drive of PMSM", *Proc. of CES/IEEE 5th International Power Electronics and Motion Control Conf.*, vol. 2, pp. 1-5, August 2006.
[7] Y. S. Kim, S. L. Ryu, and Y. A. Kwon, "An improved sliding mode observer of sensorless permanent magnet synchronous motor", *Proc. of SICE 2004 Annual Conference in Sapporo*, August 2004.
[8] A. Accetta, M. Cirrincione, M. Pucci, and G. Vitale, "Sensorless control of PMSM fractional horsepower drives by signal injection and neural adaptive-band filtering", *IEEE Transaction on Industrial Electronics*, vol. 59, pp. 1355-1366, March 2012.
[9] C. C.Ku, and K. Y. Lee, "Diagonal recurrent networks for dynamic systems control", *IEEE Trans. Neural Networks*, vol. 6, pp. 144–156, March. 1995.
[10] F. J. Lin, R. F. Fung, and R. J. Wai, "Comparison of sliding mode and fuzzy neural network control for motor-toggle servomechanism", *IEEE/ASME Trans. Mechatron.*, vol. 3, pp. 302-318, June 1998.
[11] T. S. Radwan and M. M. Gouda, "Intelligent speed control of permanent magnet synchronous motor drive based-on neuro-fuzzy approach", *Proc. IEEE PEDS*, vol. 1, pp. 602-606, April 2005.
[12] M. Comanescu, "Rotor position estimation of PMSM by sliding mode EMF observer under improper speed", *2010 IEEE International Symposium on Industrial Electronics*, pp. 1474-1478, June 2010.
[13] G.B.Lee and Y.A Kwon, "High-performance sensorless-control of PMSM using back-EMF and reactive power", *Trans. Korean Inst. Electr. Eng.*, vol. 59, pp. 740–742, April 2010.
[14] I. Boldea, G. D. Andreescu and F. Blaabjerg, "Very low speed performance of active flux based sensorless control: interior permanent magnet synchronous motor vector control versus direct torque and flux control", *Electric Power Application*, vol. 3, pp. 551-561, May 009.
[15] F.Genduso, R. Miceli, C. Rando and G. Galluzzo, "Back EMF sensorless control algorithm for high dynamic performance PMSM", *IEEE Transactions on Industrial Electronics*, vol. 57, pp. 2092-2100, June 2010.
[16] G. Foo, "Sensorless direct torque and flux-controlled IPM Synchronous motor drive at very low speed without signal injection", *Industrial Electronics*, vol. 57, pp. 395-403. October 2010.
[17] Y. Hua, "Improved sensorless control of a permanent magnet machine using fundamental pulse width modulation excitation", *Electric Power Applications*, vol.5, pp. 359-370. June 2011.
[18] D. Raca, P. Garcia, D. Reiqosa, F. Briz and R.D. Lorenz, "Carrier signal selection for sensorless control of PM synchronous machines at zero and very low speeds", *IEEE Transactions on Industry Applications*, vol. 46, pp. 167-178. January-February 2010.
[19] F. Poltschak and W. Amrhein, "Influence of winding layout of PMSM on inductance and its impact on suitability for sensorless vector control", *2010 International Symposium on Power Electronics, Electrical Drives, Automation and Motion*, pp. 1002-1007. June 2010.
[20] S.Shinnaka, "A new sensorless vector control method using high-order filters with speed-varying bandwidth for permanent magnet synchronous motors", *Electronics and Communications in Japan*, vol. 93, pp. 1-14. August 2010.

**Wei Gao** was born in Jiang su, China, in 1973. He received BS and MS from Naval University of Engineering, China, in 1994 and 2002, respectively. Now he is a professor in Naval Bengbu petty officer Academy, China. His research interests include computer application technology and six-dimensional force/torque sensor technology.

**Zhirong Guo** was born in An hui, China, in 1976. He received BS, MS and Ph.D degree from Naval University of Engineering, China, in 1999, 2006 and 2010, respectively. Now he is a vice professor in Naval Bengbu petty officer Academy, China. His main research interest is automatic testing and measurement.

# Detection System of Clone Attacks based on RSSI in Wireless Sensor Networks

Xiancun Zhou
School of Information Engineering, West Anhui University, Lu'an, China
Email: zhouxcun@mail.ustc.edu.cn

Yan Xiong and Mingxi Li
School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
Email: {yxiong, lim}@ mail.ustc.edu.cn

***Abstract*—Most of the existing node clone attack detection methods in wireless sensor networks depend on the accurate position information and the synchronization clock information of the network node. However, it is often very difficult to guarantee real-time node location information and synchronization clock information in actual network. Therefore, those detection schemes are either difficult to achieve or is with high cost and low availability. A new detection method based on distance measurement is proposed for detecting node clone attacks in wireless sensor networks, in which three detection rules are defined. It has provided that the design and implementation of the detection system. The system has low requirements for the hardware configuration of the node and is easy to implement; it can make real-time detection of clone node for network security. The analysis and the experiment show the system is secure and reliable.**

***Index Terms*—wireless sensor network; node clone attacks; distance measurement based on RSSI**

## I. INTRODUCTION

Wireless Sensor Networks (WSNs) is an integrated intelligent information system with information acquisition, information transmission and information processing. It is deployed with large amount of small sensor nodes and is placed in the monitoring area to be in perception, acquisition and processing of the information of the object in geographic region of network coverage. And it releases the information to the monitor that can be used in military and civil fields widely [1][2]. As the work environment is unattended, adversaries may easily capture and compromise sensors and deploy unlimited number of clones of the compromised nodes. Since these clones have legitimate access to the network (legitimate IDs, keys, other security credentials, etc.), they can participate in the network operations in the same way as a legitimate node, and thus launch a large variety of insider attacks [3]-[5], or even take over the network. If these clones are left undetected, the network is unshielded to attackers and thus extremely vulnerable. Therefore, clone attackers are severely destructive and effective and efficient solutions for clone attack detection are needed to limit their damage.

Nevertheless, detecting clone attacks is not trivial at all. The fundamental challenge comes from the fact that the clone nodes own all the security information (ID, keys, codes, etc.) of the original compromised sensor. Thus, they can pass all the identity/security check and escape from being distinguished from a legitimate sensor. In addition, a "smart" clone may try to hide from being detected by all means. Furthermore, clones may collude to cheat the network administrator into believing that they are legitimate. When WSN is deployed in enemy region for military purposes, an adversary may distribute clone nodes anywhere in the network. A number of protocols have been proposed so far to tackle node clone attacks. However, to the best of our knowledge, the existing node clone attack detection methods rely heavily on precise positioning and synchronization of the network clock. However, the existing hardware and technical conditions often requires investment of very high cost, it can't basically meet the requirements of precise positioning and clock synchronization until coupled with extremely demanding technical support. Therefore, the measurement accuracy of existing node clone attack detection method may be difficult to meet the requirements, or its principle is complex with high cost, and it has high requirements for relevant techniques. If a technology is to be used in large scale of the actual production, it is bound to require low cost, easy to achieve with high reliability and security. The expensive cost and high technology bottleneck has become the reason that the existing detection method cannot be used in real production life widely.

According to the fact the existing clone detection method cannot be used in large-scale in real production life, we has creatively put forward nodes clone attacks detection algorithms based on the distance measurement. We have designed and achieved of WSN clone attack automatic detection system based on distance measurement. The core of the system is the detection algorithm, and the algorithm can detect and identify clone node in the network through inter-node distance. The system has achieved the identification of the clone node at the MAC layer, and it has offered the underlying security detection for network security. The nodes in the


doi: 10.4304/jnw.8.1.221-228

network transfer the neighboring node information to the data terminals, and the data terminals will make control check to the neighboring node address table based on the three rules designed by us in turn. In this way to determine whether there is a clone node in the network. If there's clone node in the network, then the system is also able to obtain the ID number of the node, and it will inform whole network to control the normal node in the network node for the filtering of clone node data. Namely it is to achieve shield to the clone node to ensure information security of the network.

## II. Related Work

The node clone attack is one of the most vexing problems in WSNs. There are some schemes proposed for preventing and detecting node clone attacks. To our best knowledge, the first non-naive detection scheme against node clone attacks is to employ witness nodes, which are randomly selected among the nodes in the network, to undertake the task of node clone attack detection [6]. The scheme is based on location information and has some defects. For instance, if the location claim of a non-compromised node is tampered by the adversary or a malicious node fabricated the location claim of a non-compromised node, the non-compromised node would be detected as a clone node. A few distributed solutions for detection of node clone attacks in wireless sensor networks has been designed, and the solutions need system information claims, which include system synchronization time, precise node location information and so on. In literature [6]-[10], detection schemes are based on time-location claims. These schemes require system synchronization time and precise node location information. The cost is so significant that many wireless sensor networks cannot afford, in view of the fact that it considerably reduces their lifetime; Some detection schemes are based on data encryption to protect WSN from clone attacks [11]-[13]. Encryption protocols have been proposed so far to tackle node clone attacks. However, these schemes require more computing power of the nodes. Some schemes are based on other characteristic data of networks [14]-[15]. Due to the restricted resources of nodes, these schemes are not suitable for detection of node clone attacks.

The detection method proposed in this paper is based on ranging between nodes. Taking into account the factor that the node is limited on power and size, many application systems of WSN are often use low-precision ranging module. In addition, the higher precision ranging the more cost should be paid for a system. There are a number of schemes can be used to estimate the distance between nodes. Most localization methods depend on three types of physical variables measurement for localization: Time of Arrival, Time Difference of Arrival, Angle of Arrival and Received Signal Strength Indicator (RSSI) [16]. We tackle the problem of using RSSI as a distance estimator for detection method based on distance measurement, as it is available on most commercial platforms, such as those which implement ZigBee; consumes little energy and is highly scalable.

The remainders of the paper are organized as follows: In Section III, we provide three detection rules to detect clone node; We described detail issues in implementation of the detection system in Section IV; Section V is security and reliability analysis of the system; In Section VI, we provide the experiment result; Finally, we make conclusions in Section VII.

## III. Proposed Scheme

### A. Description of the Problem

Assume that the area of the application zone for a random deployment of wireless sensor network is S, and the size of the network nodes is Q (with the nodes number of n), and the stable communication distance of the node is L with the node number as x, x1, x2,....., xn-1, assume the system exists the clone nodes x' and it has the same ID number xi with node x, xi ∈ {x, x1, x2,....., xn-1}. Through the use of the multi-node distance measurement, the node identification xi of the attacked node will be detected without the need of the support of precise position information and system synchronization clock of the nodes.

### B. Principle of Detection

When the system is randomly deployed in some place, it usually requires regular topology maintenance of the whole network to ensure the normal work of the network. The ranging between nodes can be finished during the inter-communication of the nodes, records their own distance as the node ID number, and sends it to the data terminal. The data terminal will make the analysis and processing of the data to determine whether there's clone node in the network and records the ID number of the clone node. The adjacent node information table is as shown in the following:

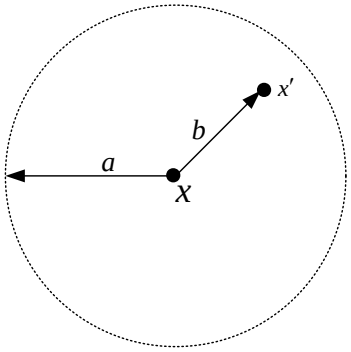
TABLE I.
Neighbor Information Table

| Flag | 01 | 00 | 00 | … | 00 |
|---|---|---|---|---|---|
| ID | 6 | 9 | 8 | … | 7 |

We use R denotes the detection range of each node. In the table, the figure "00" indicates the node distance is in the range of 0~ R/2; the figure "01" indicates the node distance is within the range of R/2~R. Nodes in the network periodically broadcast own neighbor information table. By comparing nodes' neighbor information table, we can detect clone attacks. In the following, we present three detection rules. When any of rules is in force, it means that clone attacks have been detected.

(1) Local unique ID rule

As shown in figure 1, when the clone node $x'$ is within the detection range of the node x, namely the distance between $x'$ and x is less than the detection radius R, the two nodes can probe each other. Therefore, the neighbor node information table of the two adjacent nodes can produce the contradiction equals with its own ID number. It means clone attacks have been detected.




a=R, b≤R,

Figure 1. Local unique ID rule

(2) Neighbor unique ID rule

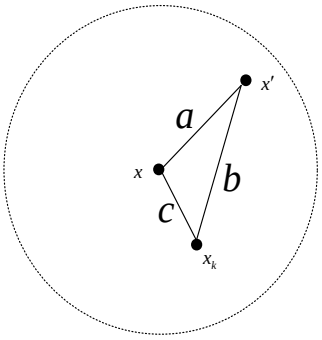
In general, clone node and compromised node cannot detect each other. Then the neighbor unique ID rule might be valid if we can find a node xk in network that can detect both replica $x'$ and compromised node x, meanwhile $x'$ and x are different neighbor of xk as shown in Figure 2. In the neighbor information table of xk, the same ID number of the nodes $x'$ and x can be detected. Then the rule would divide into two symmetrical situations described as the following:

Situation 1:

$$R<\left|x-x'\right|<\frac{3R}{2},\ \frac{R}{2}\le\left|x-x_k\right|\le R \text{ and } \left|x'-x_k\right|\le\frac{R}{2};$$

Situation 2:

$$R<\left|x-x'\right|<\frac{3R}{2},\ \frac{R}{2}\le\left|x'-x_k\right|\le R \text{ and } \left|x-x_k\right|\le\frac{R}{2};$$

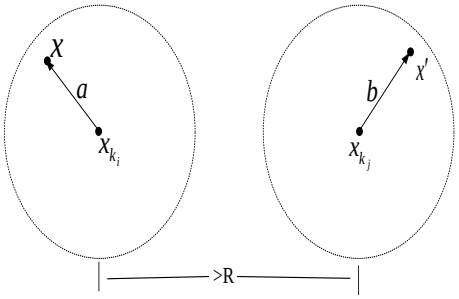



$R<a<\frac{3R}{2},\ \frac{R}{2}\le b\le R,\ c\le\frac{R}{2}$

Figure 2. Neighbor unique ID rule

(3) Global unique ID rule

Considering that the clone node $x'$ is in the monitoring range of a node $x_{k_i}$, and the node x is in the detection range 0~ R/2 of another node $x_{k_j}$ In order to detect the clone nodes, we should use neighbor information tables of different nodes. As shown in Figure 3, $x_{k_i}$ and $x_{k_j}$ find a same identity ( $x'$ or x) as a close neighbor, however, they cannot detect each other.




$a\le\frac{R}{2},\ b\le\frac{R}{2}$

Figure 3. Global unique ID rule

The nodes $x_{k_i}$ and $x_{k_j}$ can find $x'$ (or x) is recorded in its neighbor information table with flag "00" in such case. According to triangle theory, $x_{k_i}$ and $x_{k_j}$ might be neighbors, otherwise, the global unique ID rule should be in force.

The three rules mentioned above are independent of each other, which can be tested in turn in detection of clone attacks. So we can compute the total probability of detection. There are many routing protocols can be used to transmit neighbor information tables of sensor nodes. In order to achieve comparisons of neighbor information tables without storing any received information, other complex rules by analyzing more than two nodes' neighbor information tables will not be mentioned in this paper.

## IV. The Implementing of the Detection System

### A. Distance Measurement based on RSSI

Range-based localization must reassure distance between neighboring nodes. Received Signal Strength Indication (RSSI), featuring low communication overhead and low complexity, is our basis of localization on the energy constrained sensor nodes. In this method, it has known that launch node and launch signal strength, according to receive signal strength, receiving node can calculate the signal transmission loss and transform loss into distance using the theory and experience model.

In fact, there is no explicit analytical expression about the value of distance and signal strength loss. The universal theory model in wireless signal transmission is Shadowing model; the expression is given as following:

$$PL(d)=PL(d_0)-10n\lg(d/d_0)-X_\sigma \qquad (1)$$

where, $d$ denotes distance between launcher and receiver, $d_0$ denotes the reference distance, n denotes path loss index, which relies on the surrounding environment and building types, $X_\sigma$ denotes normal random variables with the standard deviation $\sigma$, which mainly influenced by the sensors' perception accuracy, $PL(d_0)$ denotes signal strength value received in the position of $d_0$ meters. Based on a lot of range-based localization

obtained, distance measurement based on RSSI is in line with the curve shown in Figure 4.

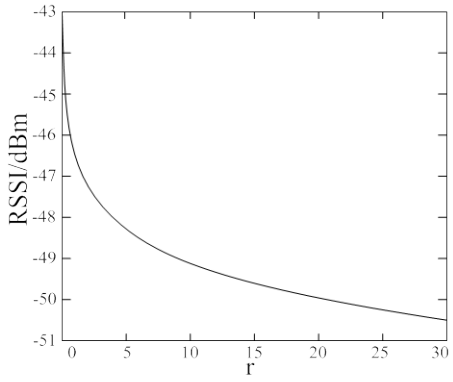



Figure 4. Curve of distance measurement based on RSSI

### *B. System Scheme*

The detection system consists of sensor nodes, data collection terminals, embedded systems and data analysis software. The system components and architecture is shown in Figure 5.

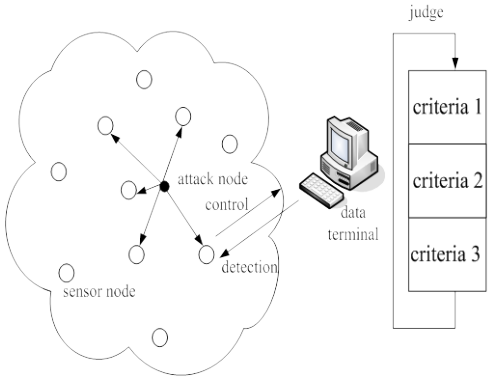



Figure 5. Architecture diagram of the system

The solid circle in the figure indicates the clone attack nodes, and the hollow circle indicates the normal nodes in the network. According to the position relationship of the clone nodes and normal nodes in the network, we have worked out three detection rules. The data terminal will make the detection to the address table of the neighboring node. In this way to determine whether there's the contradictions between the ID numbers of nodes of the neighboring address nodes to conclude a conclusion whether there's a clone node.

The data collection terminals and sensor nodes containing multiple functional modules, and they are used to achieve the wireless communications, data processing, data acquisition, and other functions, just as is shown in Figure 6. Of which, the sensing module is responsible for the indication value of the signal strength when communicating; and the calculation module is responsible for processing the collected data and packs it in frame for transmission and processes the tasks of the other nodes and control commands; the communication module is responsible for the transceiver of the wireless data and real-time channel listen; and the data receiver module is used to receive and process sensor node of the returned data frame by extracting the information of neighboring nodes; the data processing module will make the analysis and process of the extracted neighboring node information, and read the ID number of the clone node, then it will transfer the data to the data display module; and the data display module will make real-time display of the verification results of the clone nodes in multimedia ways.

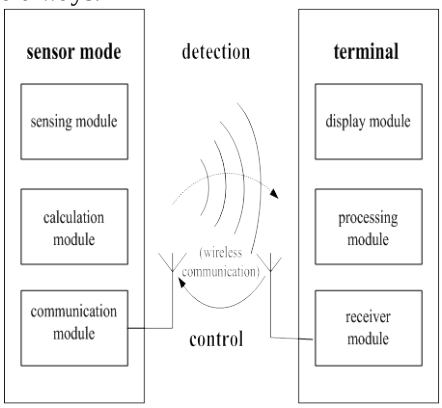



Figure 6. Structure diagram of the system module

### *C. Implementation of the System*

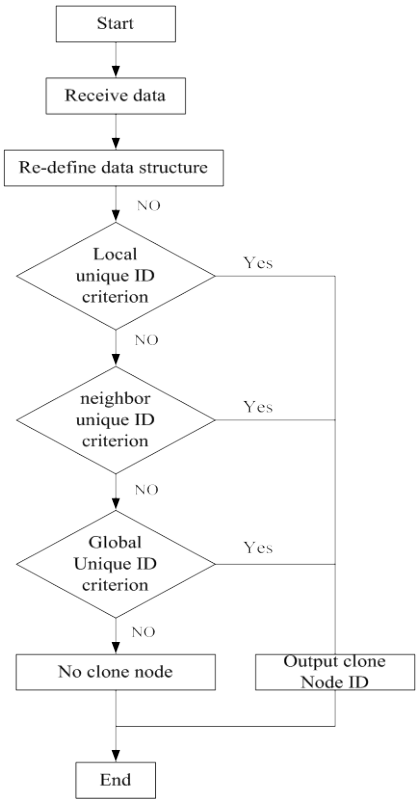



Figure 7. Data comparison detection algorithm flow diagram

The implementation of the clone node detection system is divided into two parts: embedded software and data analysis software. Of which, the embedded software runs on the sensor nodes; and data analysis software runs on data terminal. The two parts run independently and achieve the data interaction through the inter-network communications. System uses data comparison detection

algorithm to analyze the neighbor node information table of each node, and then identify the clone node in the network. Detection algorithm flow chart is shown in figure 7.

The nodes collects the ranging data of signal strength compared with the set threshold, and sends the data to the base station after processing. The computer terminal receives the data via the base station, after the re-definition of the data structures to determine whether there is a clone node in the system according to the three rules. If there's the existence of clone node, recording the ID number of the clone node, and the multi-media terminal will display dynamically and report the clone node information. The terminal will automatically send the instruction to the base station. The base station will broadcast ID number of the clone node to all nodes in the wireless sensor network, and each node will automatically shield the packet of the clone node. Therefore, the clone attack will fail.

## V. The Analysis of System Security and Reliability

To improve the detection performance, the probabilities of three rules are analysis in this section. Assume $P_1$ is the detection probability of local unique ID rule, $P_2$ is the detection probability of neighbour unique ID rule, $P_3$ is the detection probability of global unique ID rule, and we can define the total detection probability as $P = P_1 + P_2 + P_3$.

(1) $P_1$: the detection probability of local unique ID rule

Because nodes ( $x, x_1, x_2 \cdots x_{n-1}$) obeys the uniformly independent distribution in the region, the probability of distributed in a position can be describe as $f(x_i)$.

$$f(x_i) = \begin{cases} \frac{1}{s}, & x_i \in S_{re} \\ 0, & x_i \notin S_{re} \end{cases} \quad (2)$$

According to the definition of Local unique ID rule (that is $|x - x'| \le R$), the detection probability of local unique ID rule is as following:

$$P_1 = \iint_{|x-x'| \le R} f(x) \cdot f(x') dS_{re} \quad (3)$$

Substituted equation (2) into equation (3), we can get equation (4).

$$P_1 = \iint_{|x-x'| \le R} \frac{1}{s} dS_{re} = \frac{\pi R^2}{s} \quad (4)$$

(2) $P_2$: the detection probability of neighbor unique ID rule

When the clone node can't be detected by neighbor unique ID rule, it might effect in the situation mention in Section IV. We define situation 1 and situation 2 as following:

Situation 1:

$$\begin{cases} |x_k - x| \le \frac{R}{2}, & (Condition \quad I) \\ \frac{R}{2} < |x_k - x'| \le R, & (Condition \quad II) \end{cases}$$

Situation 2:

$$\begin{cases} |x_k - x'| \le \frac{R}{2}, & (Condition \quad I') \\ \frac{R}{2} < |x_k - x| \le R, & (Condition \quad II') \end{cases}$$

As given in equation (3), the probability of condition $I$ (or condition $I'$) can be computed as following:

$$P(I) = \iint_{|x_k - x| \le R/2} f(x_k) \cdot f(x) d_{re}S = \frac{\pi \left(R/2\right)^2}{s} = \frac{\pi R^2}{4s} \quad (5)$$

The probability of condition $II$ (or condition $II'$) can be computed as following:

$$P(II) = \iint_{R/2 < |x_k - x'| \le R} f(x_k) \cdot f(x') dS_{S_{re}} = \frac{\pi R^2 - \pi \left(R/2\right)^2}{s} = \frac{3\pi R^2}{4s} \quad (6)$$

In view of that condition $I$ and condition $II$ are independent of each other, so the probability of situation 1 (or situation 2) is as following:

$$P(I) \cdot P(II) = \frac{\pi R^2}{4s} \cdot \frac{3\pi R^2}{4s} = \frac{3\pi^2 R^4}{16 s^2} \quad (7)$$

In the same way, we can obtain $P(I') \cdot P(II')$.

Assume that $P_2'$ denotes the detection probability of neighbor unique ID rule, it can be computed as following:

$$P_2' = \left[ P(AB) + P(A'B') \right] \cdot (1 - P_1) = \frac{3\pi^2 R^4}{8 s^3} \cdot \left( s - \pi R^2 \right) \quad (8)$$

There are n sensor nodes have been deployed. In addition to the node x, another nodes are similarly to detect the clone node, thus, the detection probability of neighbor unique ID rule is the probability that there is at least one node can satisfy both condition $I$ (or condition $I'$) and condition $II$ (or condition $II'$).

The detection probability of neighbor unique ID rule can be obtained by the Bernoulli equation:

$$P_2 = 1 - \left(1 - P_2'\right)^{n-1} \quad (9)$$

(3) $P_3$: the detection probability of global unique ID rule

Nodes in the network periodically broadcast own neighbor information table and compare the received table to its own table for detection of clone attacks. But local unique ID rule and neighbour unique ID rule aren't in force. We should find a pair of nodes can satisfy the conditions describe as follows.

Situation 1:

$$\begin{cases} \left|x_{k_1}-x\right| \le \frac{R}{2}, & (Condition \quad I) \\ \left|x_{k_2}-x'\right| \le \frac{R}{2}, & (Condition \quad II) \\ \left|x_{k_1}-x_{k_2}\right| > R, & (Condition \quad III) \end{cases}$$

Situation 2:

$$\begin{cases} \left|x_{k_1}-x'\right| \le \frac{R}{2}, & (Condition \quad I') \\ \left|x_{k_2}-x\right| \le \frac{R}{2}, & (Condition \quad II') \\ \left|x_{k_1}-x_{k_2}\right| > R, & (Condition \quad III') \end{cases}$$

Similar to the equation (5), the probabilities of condition $I$ , condition $II$ and condition $III$ can be computed as follows:

$$P(I) = \iint\limits_{\left|x_{k_1}-x\right| \le R/2} f\left(x_{k_1}\right) \cdot f(x) dS_{re} = \frac{\pi\left(R/2\right)^2}{s} = \frac{\pi R^2}{4s} \quad (10)$$

$$P(II) = \iint\limits_{\left|x_{k_2}-x'\right| \le R/2} f\left(x_{k_2}\right) \cdot f(x') dS_{re} = \frac{\pi\left(R/2\right)^2}{s} = \frac{\pi R^2}{4s} \quad (11)$$

$$P(III) = \iint\limits_{\left|x_{k_1}-x_{k_2}\right| > R} f\left(x_{k_1}\right) \cdot f\left(x_{k_2}\right) dS_{re} = 1 - \frac{\pi R^2}{s} \quad (12)$$

Considering that condition $I$ , $II$ and $III$ are independent of each other, the probability of situation 1 (or situation 2) is as the following:

$$P(I) \cdot P(II) \cdot P(III) = \frac{\pi R^2}{4s} \cdot \frac{\pi R^2}{4s} \cdot \left(1 - \frac{\pi R^2}{s}\right) = \frac{\pi^2 R^4}{16 s^3}\left(s - \pi R^2\right) \quad (13)$$

Similarly, we can get $P(I') \cdot P(II') \cdot P(III')$ . $P_3'$ can be computed as equation (14).

$$\begin{aligned} P_3' &= \left[P(ABC) + P(A'B'C')\right] \cdot \left(1 - P_1\right) \cdot \left(1 - P_2\right) \\ &= \frac{\pi^2 R^4}{8 s^4}\left(s - \pi R^2\right)^2 \cdot \left(1 - P_2'\right)^{n-1} \end{aligned} \quad (14)$$

According to Bernoulli equation, $P_3$ can be obtained as follows.

$$P_3 = 1 - \left(1 - P_3'\right)^{C_{n-1}^2} \quad (15)$$

In equation (15), $P_3$ is the detection probability that there are at least one pair of nodes can meet condition $I$ (or condition $I'$ ), condition $II$ (or condition $II'$ ) and condition $III$ (or condition $III'$ ) in the same time.

In conclusion, we can get the total detection probability $P$ .

$$\begin{aligned} P &= P_1 + P_2 + P_3 \\ &= \frac{\pi R^2}{s} + 1 - \left(1 - P_2'\right)^{n-1} + 1 - \left(1 - P_3'\right)^{C_{n-1}^2} \\ &= \frac{\pi R^2}{s} - \left(1 - P_2'\right)^{n-1} - \left(1 - P_3'\right)^{C_{n-1}^2} + 2 \end{aligned} \quad (16)$$

Where

$$P_2' = \frac{3\pi^2 R^4}{8 s^3} \cdot \left(s - \pi R^2\right),$$

$$P_3' = \frac{\pi^2 R^4}{8 s^4}\left(s - \pi R^2\right)^2 \cdot \left(1 - P_2'\right)^{n-1}.$$

As given in equation (16), the total detection probability $P$ can be computed by the parameters of $n$ , $s$ and $R$ . Thus, we can use the function to determine suitable parameters for different applications, and then the range-based method can get a high detection probability. Therefore, we can select the appropriate parameters $n$ , $s$ and $R$ to ensure that the system is with a high detection probability, in this way to ensure system reliability and security.

## VI. EXPERIMENTAL VERIFICATION

### A. Experiment Steps

32 nodes (30 normal nodes; a base station node; a clone node) and 1 PC.

First of all, on a rectangular area of size 10m x 8m, to build the experimental environment, we place 30 sensor nodes evenly for composition of WSN. The system test environment is indoor stadium. Then, open all the nodes, which begin to work normally. Each node records the nodes flag and ID in its communication distance in the R value. Next, we place clone node in different positions of the network to verify whether the network can accurately detect clone node and shield it immediately.

### B. Experiment Results

The experiment is to test the clone attack against the node 13. There are three situations.

(1) Rule one

TABLE II.
NEIGHBOR NODE INFORMATION TABLE OF NODE 13

| Node ID | Flag | ID | Flag | ID | Flag | ID | Flag | ID |
|---|---|---|---|---|---|---|---|---|
| 3 | 00 | 02 | 00 | 04 | … | … | 01 | 0E |
| 4 | 01 | 02 | 00 | 03 | … | … | 01 | 0F |
| 5 | 01 | 03 | 00 | 04 | … | … | 01 | 10 |
| 6 | 01 | 04 | 00 | 05 | … | … | 01 | 10 |
| 7 | 00 | 02 | 01 | 03 | … | … | 01 | 12 |
| 8 | 01 | 02 | 00 | 03 | … | … | 01 | 13 |
| 9 | 01 | 02 | 00 | 00 | … | … | 01 | 14 |
| 0A | 01 | 03 | 01 | 00 | … | … | 01 | 15 |
| 0B | 01 | 04 | 01 | 00 | … | … | 00 | 0A |
| … | … | … | … | … | … | … | … | … |
| 13 | 01 | 08 | 01 | 09 | … | … | 00 | 13 |
| … | … | … | … | … | … | … | … | … |
| 2 | 00 | 03 | 01 | 04 | … | … | 01 | 0D |

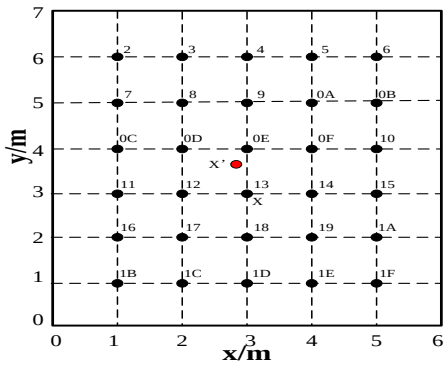



Figure 8. Rule one

As shown in Figure 8, clone node is in detection radius of the node 13, then, there is the same ID with the node 13 in its neighbor node information table. According to local unique ID rule, the clone attack can be found.

(2) Rule two

As shown in figure 9, clone node is in detection radius of node 9.There is two records about node 13 with the conflicting flag. So, the clone attack can be found.

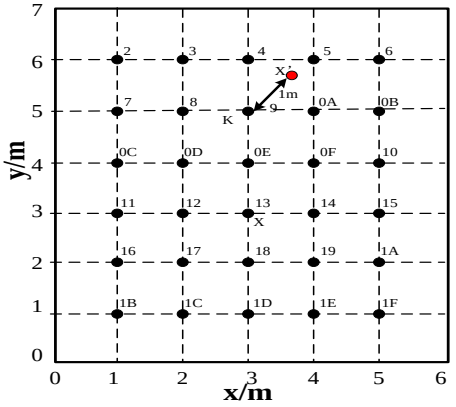



Figure 9. Rule two

TABLE III.
NEIGHBOR NODE INFORMATION TABLE OF NODE 9

| Node ID | Flag | ID | … | Flag | ID | Flag | ID | Flag | ID |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 00 | 02 | … | … | … | 01 | 0E | … | … |
| 4 | 01 | 02 | … | … | … | 01 | 0F | … | … |
| 5 | 01 | 03 | … | … | … | 01 | 10 | … | … |
| 6 | 01 | 04 | … | … | … | 01 | 10 | … | … |
| 7 | 00 | 02 | … | … | … | 01 | 12 | … | … |
| 8 | 01 | 02 | … | … | … | 01 | 13 | … | … |
| 9 | 01 | 02 | … | 01 | 13 | 01 | 14 | 00 | 13 |
| 0A | 01 | 03 | … | … | … | 01 | 15 | … | … |
| … | … | … | … | … | … | … | … | … | … |
| 2 | 00 | 03 | … | … | … | 01 | 0D | … | … |

(3) Rule 3

As shown in Figure 10, clone node is in the detection radius of node 3. There are records about node 13 in the neighbor node information table of the node 3 and node 12. However, node 3 and node 12 can't detect each other. The same network has two ID number of the same node, so, the clone attack can be found.

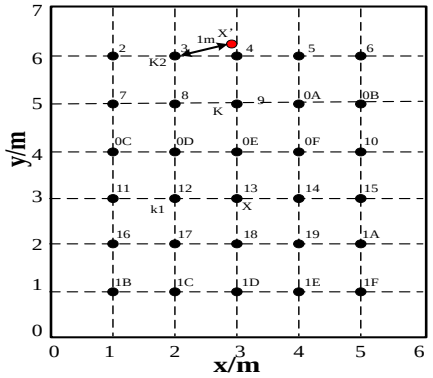



Figure 10. Rule three

TABLE IV.
NEIGHBOR NODE INFORMATION TABLE OF NODE 9 AND NODE 12

| Node ID | Flag | ID | … | Flag | ID | Flag | ID | Flag | ID |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 00 | 02 | … | … | … | 01 | 0E | 00 | 13 |
| 4 | 01 | 02 | … | … | … | 01 | 0F | … | … |
| 5 | 01 | 03 | … | … | … | 01 | 10 | … | … |
| 6 | 01 | 04 | … | … | … | 01 | 10 | … | … |
| 7 | 00 | 02 | … | … | … | 01 | 12 | … | … |
| 8 | 01 | 02 | … | … | … | 01 | 13 | … | … |
| … | … | … | … | … | … | … | … | … | … |
| 12 | 01 | 07 | … | 00 | 13 | 01 | 14 | … | … |
| … | … | … | … | … | … | … | … | … | … |
| 2 | 00 | 03 | … | … | … | 01 | 0D | … | … |

### *D. The Analysis of Experimental Results*

From the experimental data, we can see that the detection algorithm can effectively detect clone node, and the actual experiment verified it.

## VII. CONCLUSION

The node clone attack detection with the use of node features should rely on the node positioning and network synchronization clock information. It leads to great enhancement of the cost of the detection system and it is also very rigor in technology requirements, so it is with low practicality and reliability. We have proposed to use the distance measurement based on RSSI to detect the clone attack. The method has the following advantages: (1)It has low requirements for the hardware configuration of the node, and is easy to implement and transplant to other systems; (2)It can make real-time detection of clone node with short testing cycle; (3) It has no requirements towards the clock synchronization and it has no need for node location information, which has greatly reduced the cost; (4) It can be equipped with various sensors and with good scalability; (5) The terminal program can run on computers, handheld, mobile phones as well as various platforms to facilitate real-time control of the networks.

## ACKNOWLEDGMENT

The work presented in this paper is supported by the Nature Science Foundation of Anhui Education Department under grant no. KJ2011B204 and the National Science Foundation of China under grant No. 61170233.

## REFERENCES

[1] X. Q. Chen, K. Makki, K.Yen, and N. Pissinou, "Sensor network security: a survey", IEEE Communications Surveys &Tutorials, vol. 11, pp. 52-73, April 2009.
[2] F.Yang, X.H.Zhou, and Q.Y.Zhang, J. Xie and S. G. Zhang, "A practical traceback mechanism in wireless sensor networks", Acta Electronica Sinica, vol. 37, pp. 202- 206, January 2009 (in Chinese).
[3] B. Parno, A. Perrig, V. Gligor, "Distributed detection node replication attacks in sensor networks", Proceedings of IEEE Symposium on Security and Privacy, pp. 49-63, 2005.
[4] M. Conti, R. D. Pietro, L. V. Mancini, "A randomize, efficient, and distributed protocol for the detection of node replication attacks in wireless sensor networks", Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 80-89, 2007.
[5] K. Chris, W. David, "Secure routing in wireless sensor networks: attacks and counter measures", Proceedings of First IEEE International Workshop on Sensor Network Protocols and Applications, pp. 113-127, 2003.
[6] H. Chan, A. Perrig, "Security and privacy in sensor networks", Computer, vol. 36, no. 10, pp. 103- 105, 2003.
[7] M. Conti, R. D. Pietro, L. V. Mancini, and A. Mei, "Distributed detection of clone attacks in wireless sensor networks", IEEE Transactions on Dependable and secure computing, vol. 8, pp. 685-698, September/October 2011.
[8] X. M. Deng, Y. Xiong, and D. P. Chen, "Mobility-assisted detection of the replication attacks in mobile wireless sensor networks", In Proceedings of the 6th International Conference on Wireless and Mobile Computing, pp. 225-232, 2010.
[9] M. Jadliwala, S. Zhong, S. Upadhyaya, C. M. Qiao, J. P. Hubaux, "Secure distance-based localization in the presence of cheating beacon nodes", IEEE Transactions on Mobile Computing, vol. 9, no. 6, pp. 810-823, June 2010.
[10] Y. P. Zeng, J. N. Cao, J. Hong, and et al, "A secure Monte Carlo Localization algorithm for mobile sensor networks", In Proceedings of the 6th International Conference on Mobile Ad-hoc and Sensor Systems (MASS '09), pp. 1054-1059, 12-15, October 2009.
[11] A. Perrig, R. Szewczyk, V. Wen et al, "SPINS: Security protocol for sensor networks", Wireless Networks, vol. 8, pp. 521-534, 2002.
[12] S. U. Khan, L. Lavagno, and C. Pastrone, "A key management scheme supporting node mobility in heterogeneous sensor networks", In Proceedings of the 6th International Conference on Emerging Technologies (ICET), pp. 364-369, 2010.
[13] C. Bekara, M. Laurent-Maknavicius, "A new protocol for securing wireless sensor networks against nodes replication attacks", In Proceedings of the 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2007), pp. 59-59, 2007.
[14] H. Choi, S. Zhu, T. F. La Porta, "SET: Detecting node clones in sensor networks", In Proceedings of the 3rd International Conference on Security and Privacy in Communications Networks and the Workshops (SecureComm 2007), pp. 341-350, 2007.
[15] W. T. Zhu, "Node replication attacks in wireless sensor networks: Bypassing the neighbor-based detection scheme", In Proceedings of the International Conference on Network Computing and Information Security, pp. 156-160, 2011.
[16] E. Coca, V. Popa, G. Buta", Wireless sensor network nodes performance measurements and RSSI evaluation", In Proceedings of the 15th International Symposium for Design and Technology of Electronics Packages, pp. 105-112, September 2009.

**Xiancun Zhou** received the B.S. degree in electrical engineering from Anhui University, Hefei, China, in 1997, and received the M. S. degree in computer science from Heifei University of Technology, Heifei, China, in 2004. She is now a associate professor at the Department of Information Engineering in West Anhui University, Lu'an, China. Her current research interests include information security, wireless sensor network and identity authentication.

**Yan Xiong** received the M.S. and Ph.D. degrees in computer science and technology, university of science and technology of China, Hefei, China, in 1986 and 1990 respectively. He is currently a professor in School of Computer Science and Technology, University of Science and Technology of China. His main research interests include distributed processing, mobile computation, and computer network and information security.

**Mingxi Li** received the M.S. degrees in China's people's liberation army artillery institute. He is currently pursuing his Ph.D. degree in computer science at University of Science and Technology of China. His research focuses on wireless sensor network.

# Analysis and Improvement for SPINS

Yuan Wang
Department of Computer Science and technology, Jilin University, Changchun 130012, China
Email: wangyuan2109@mails.jlu.edu.cn

Liang Hu, Jian Feng Chu, *Corresponding author Xiao Bo Xu
Department of Computer Science and technology, Jilin University, Changchun 130012, China
Email: hul@jlu.edu.cn, chujf@jlu.edu.cn, 783392216@qq.com

***Abstract*—Wireless sensor network is a new application network and with broad application prospects, which is considered as the leader of the top-ten technologies in changing world in future. Recent years, the wireless sensor network receives much attention and application, its security becomes even more prominent. SPINS only considers the simple main key sharing way in the safe guidance aspect. It always encountered Dos attack. SPINS lacks key update mechanism, does not support the network expansion.We propose a design scheme of secure communication protocol for wireless sensor network based on the hierarchical topology referring to SPINS, which is widely accepted at present. Our scheme has utilized the LEACH algorithm and has made up it in the security insufficiency. During the topology establishment, we add the authentications of the cluster-heads through base station to ensure the validity of the cluster-heads; in addition, we add a one-way hash function and a shared key in the key management, making the encryption key and authentication key change dynamically to enhance the security of network communications. The improved scheme takes into account its own characteristics and limitations of wireless sensor networks, and seeks to meet the security needs of network communications, that is confidentiality, integrity, data freshness, key update, and authentication. Although the improved protocol consumed slightly more energy compared to SPINS, safety performance has improved a lot.**

***Index Terms*—Wireless sensor networks, hierarchy topology, SPINS LEACH, Key timing update mechanism, one-way key hash chain**

## I. INTRODUCTION

A wireless sensor network is a network of distributed autonomous devices that can sense or monitor physical or environmental conditions cooperatively [1].We envision a future where thousands to millions of small sensors form self-organizing wireless networks. How can we provide security for these sensor networks? Security is not easy; compared with conventional desktop computers, severe challenges exist —these sensors will have limited processing power, storage, bandwidth, energy and so on. The detection approaches are generally realized by the Intrusion Detection System (IDS), such as DTRAB in [2]. A review of current literature on WSNs reveals that only a few user authentication schemes have been adequately addressed [3-9] at the application layer Based on the credible base station, Perrig and others [10] proposed SPINS; ZigBee [11] is a popular set of protocols in the industry currently, which provides a high security for WSNs, but with high energy consumption; The SPINS is suitable for all kinds of wireless sensor networks. So a number of related researches are based on the SPINS.

SPINS (Security Protocols for Sensor Network), is popular and practical at present. It takes into account the context of data in confidentiality, integrity, fresh, certification and other aspects.However, this security system only considers the simple main key sharing way in the safe guidance aspect, so that its security is fully dependent on the base station. In response to this issue, many researchers have proposed key pre-distribution methods, of which the random key pre-distribution scheme is one of the research results. This method weakens the security dependence on the base station to a certain extent, thus the base station's position in the security system reduces to the same position with the general node. Even if the site has been captured, you can also use base station network to re-organize. However a number of issues have not been sufficiently taken into consideration in the SPINS protocol: Firstly, it does not consider the problem of information leakage (If the communication process design works badly, the attacker can eavesdrop on all of the information); Secondly, if a node is captured, the process method is not comprehensive; thirdly, it does not consider the Dos attack. What is more, it does not take into account key update.

We need to surmount these challenges, because security is so important. Sensor networks will expand to fill all aspects of our lives. Here are some typical applications:

*(1).Energy Management.*

Energy distribution will be better managed when we begin to use remote sensors. For example, the power load that can be carried on an electrical line depends on ambient temperature and the immediate temperature on the wire. If these were monitored by remote sensors and the remote sensors received information about desired load and current load, it would be possible to distribute load better.


doi: 10.4304/jnw.8.1.229-236

*(2).Battlefield Management.*

Remote sensors can help eliminate some of the confusion associated with combat. They can allow accurate collection of information about current battlefield conditions as well as giving appropriate information to soldiers, weapons, and vehicles in the battlefield. Thus, we can make full use of the information to gain more chances to win.

*(3).Emergency Response Information.*

Sensor networks will collect useful information about the status of buildings, people, and transportation pathways. Thus, sensor information must be collected and passed on in meaningful, secure ways to the emergency response personnel. Typically, sensor nodes are grouped in clusters, and each cluster has a node that acts as the cluster head. All nodes forward their sensor data to the cluster head, which in turn routes it to a specialized node called sink node (or base station) through a multi-hop wireless communication as shown in Figure 1. [12]

This article presents a set of secure Protocols for Sensor Networks, SPINS. The main contributions of this paper are as follows:

1 We explore the challenges for security in sensor networks.

2 We propose a design scheme of secure communication protocol for wireless sensor network based on the hierarchical topology referring to SPINS.

3 We use key update mechanism to update key.

## II. The Analysis of the Classic Spins Protocol

*(1).SNEP: Data Confidentiality, Authentication, Integrity, and Freshness*

Data confidentiality is one of the most secure primitives in almost every secure protocol. A simple form of confidentiality can be achieved through encryption, but pure encryption is not sufficient. Another important secure property is semantic security, which ensures that an eavesdropper has no information about the plaintext, even if it gets multiple encryptions of the same plaintext. For example, if an attacker has an encryption of a 0 bit and an encryption of a 1 bit, it will not help distinguish whether a new encryption is an encryption of 0 or 1.

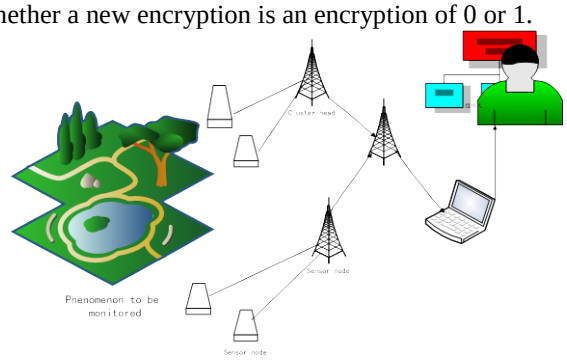



Figure 1. Architecture of a typical wireless sensor network

A basic technique to achieve this is randomization: Before encrypting the message with a chaining encryption function, the sender precedes the message with a random bit string. This prevents the attacker from inferring the plaintext of encrypted messages if it knows cipher.

A good secure design practice is not to reuse the same cryptographic key for different cryptographic primitives; this prevents any potential interaction between the primitives that might introduce a weakness. Therefore, SPINS derives independent keys for encryption and MAC operations.

The combination of these mechanisms forms Sensor Network Encryption Protocol SNEP. The encrypted data has the following format as in (1):

$$E= \{D\} (Kenc, C), \quad (1)$$

where D is the data before encryption, the encryption key is $K_{enc}$, and the counter is C. The MAC as in (2):

$$M= MAC (Kmac, C \mid E) \quad (2)$$

Among them, $K_{mac}$ expresses the news authentication algorithm key, C | E is counter value c and the scrambled text E cementation, indicates the news authentication code is carries on the operation on the counter and the scrambled text. The complete message that A sends to B as in (3):

$$A->B: \{D\} (Kenc, C), MAC (kmac, C|\{D\} (Kenc, C)) \quad (3)$$

SNEP offers the following nice properties:

*1 Data authentication:* If the MAC verifies correctly, a receiver knows that the message originated from the claimed sender.

*2 Weak freshness:* If the message verifies correctly, a receiver knows that the message must have been sent after the previous message it received correctly (that had a lower counter value). This enforces a message ordering and yields weak freshness.

*3 Semantic security:* Since the counter value is incremented after each message, the same message is encrypted differently each time. The counter value is sufficiently long enough to never repeat within the lifetime of the node.

SNEP only provides weak data freshness, because it just enforces a sending order on the messages within node B, but no absolute assurance to node A that a message was created by B in response to an event in node A.

Node A achieves strong data freshness for a response from node B through a nonce $N_A$ (which is a random number so long that exhaustive search of all possible nonce is not feasible). Node A generates $N_A$ randomly and sends it along with a request message $R_A$ to node B. The simplest way to achieve strong freshness is for B to return the nonce with the response message $R_B$ in an authenticated protocol. However, instead of returning the nonce to the sender, we can optimize the process by using the nonce implicitly in the MAC computation. The entire SNEP protocol providing strong freshness for B's response is as in (4) and (5):

$$A ->B: NA, RA \quad (4)$$

B->A: {RB} (Kenc, C), MAC (Kmac, NA | C | {RB} (Kenc, C)) (5)

If the MAC verifies correctly, node A knows that node B generated the response after it sent the request. The first message can also use plain SNEP if confidentiality and data authentication are needed.

*(2) μTESLA: Stream Authentication Protocol*

TESLA [12] is not designed for the limited computing environments we encounter in sensor networks for the following three reasons:

The recently proposed TESLA protocol provides efficient authenticated broadcast. TESLA authenticates the initial packet with a digital signature. Obviously, digital signatures are too expensive to compute on our sensor nodes, since even fitting the code into the memory is a major challenge. For the same reason as we mention above, one-time signatures are a challenge to use on our nodes.

*1 μTESLA overview:*

We give a brief overview of μTESLA, followed by a detailed description.

Authenticated broadcast requires an asymmetric mechanism otherwise any compromised receiver could forge messages from the sender. Unfortunately, asymmetric cryptographic mechanisms have high computation, communication, and storage overhead, making their usage on resource-constrained devices impractical. μTESLA overcomes this problem by introducing asymmetry through a delayed disclosure of symmetric keys, which results in an efficient broadcast authentication scheme.

μTESLA requires that the base station and nodes be loosely time synchronized, and each node knows an upper bound on the maximum synchronization error. To send an authenticated packet, the base station computes a MAC on the packet with a key that is secret at that point in time. When a node gets a packet, it can verify that the corresponding MAC key was not yet disclosed by the base station (based on its loosely synchronized clock, its maximum synchronization error, and the time schedule at which keys are disclosed). Since a receiving node is assured that the MAC key is known only by the base station, the receiving node is assured that no adversary could have altered the packet in transit. The node stores the packet in a buffer. At the time of key disclosure, the base station broadcasts the verification key to all receivers. When a node receives the disclosed key, it can verify the correctness of the key. If the key is correct, the node can now use it to authenticate the packet stored in its buffer.

Each MAC key is a key of a key chain, generated by a public one-way function F. To generate the one-way key chain, the sender chooses the last key $K_N$ of the chain randomly, and repeatedly applies F to compute all other keys as in (6):

$$Ki=F(Ki+1) \quad (6)$$

Each node can easily perform time synchronization and retrieve an authenticated key of the key chain for the commitment in a secure and authenticated manner, using the SNEP building block.

Example Figure 2 shows the μTESLA one-way key chain derivation, the time intervals, and some sample packets that the sender broadcasts. Each key of the key chain corresponds to a time interval and all packets sent within one time interval are authenticated with the same key. In this example, the sender discloses keys two time intervals after it uses them to compute MACs. We assume that the receiver node is loosely time synchronized and knows $K_0$ (a commitment to the key chain). Packets P1 and P2 sent in interval 1 contain a MAC with key K1. Packet P3 has a MAC using key P2. So far, the receiver cannot authenticate any packets yet. Assume that packets P4, P5, and P6 are all lost, as well as the packet that discloses key K1, so the receiver can still not authenticate P1, P2, or P3. In interval 4 the base station broadcasts key K2, which the node authenticates by verifying as in (7):

$$K0=F(F(K2)) \quad (7)$$

The node derives as in (8):

$$K1=F(K2) \quad (8)$$

So it can authenticate packets P1, P2 with P1, and P3 with P2.

*2 μTESLA detailed description:*

μTESLA has multiple phases: security initialization of base station, sending authenticated packets, bootstrapping new receivers, and authenticating packets. We first explain how μTESLA allows the base station to broadcast authenticated information to the nodes, and we then explain how TESLA allows nodes to broadcast authenticated messages.

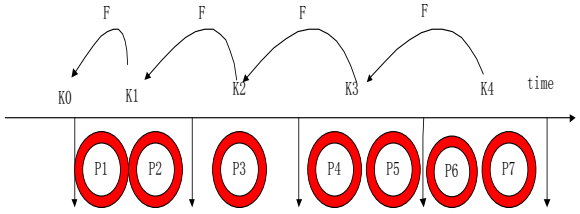



Figure 2. The μTESLA one-way key chain

Security initialization of base station.The sender first generates a sequence of secret keys (a one-way key chain). To generate a one-way key chain of length n, the sender chooses the last key $K_n$ randomly, and generates the remaining values by successively applying a one-way function F as in (9):

$$ki+1=F(ki) \quad (9)$$

Because F is a one-way function, anybody can compute forward, e.g., compute $K_0 \ldots K_i$ given $K_{i+1}$.On the other hand, nobody can compute backward, because the generator function is one-way.

Broadcasting authenticated packets. Time is divided into uniform time intervals and the sender associates each key of the one-way key chain with one time interval. In time interval i, the sender uses the key of the current interval, $K_i$, to compute the message authentication code (MAC) of packets in that interval. In time interval (i + a), the sender

reveals $key_i$. The key disclosure time delay is on the order of a few time intervals, as long as it is greater than any reasonable round trip time between the sender and the receivers.

In a one-way key chain, keys are self-authenticating. The receiver can easily and efficiently authenticate subsequent keys of the one-way key chain using one authenticated key. To bootstrap μTESLA, each receiver needs to have one authentic key of the one-way key chain as a commitment to the entire chain. Other requirements are that the sender and receiver be loosely time synchronized, and that the receiver knows the key disclosure schedule of the keys of the one-way key chain. Both the loose time synchronization and the authenticated key chain commitment can be established with a mechanism providing strong freshness and point-to-point authentication. A receiver R sends a nonce $N_R$ in the request message to the sender S. the sender S replies with a message containing its current time $T_S$, a key $K_i$ of the one-way key chain used in a past interval i (the commitment to the key chain), the starting time $T_i$ of interval i, the duration $T_{int}$ is synchronization interval, and the disclosure delay d (the last three values describe the key disclosure schedule) as in (10) and (11):

$$A\text{->}S: (NM \mid RA), MAC (Kmac, NM \mid RA) \quad (10)$$

$$S\text{->}A: (Ts \mid Ki \mid Ti \mid Tint \mid d) Kmac, MAC (Kmac, NM \mid Ts \mid Kt \mid Ti \mid Tint \mid d) \quad (11)$$

Since we do not need confidentiality, the sender does not need to encrypt the data. The MAC uses the secret key shared by the node and base station to authenticate the data, the nonce $N_M$ allows the node to verify freshness. Instead of using a digital signature scheme as in TESLA, we use the node-to-base-station authenticated channel to bootstrap the authenticated broadcast. Authenticating broadcast packets. When a receiver receives the packets with the MAC, it needs to ensure that the packet is not a spoof from an adversary. The adversary already knows the disclosed key of a time interval, so it could forge the packet since it knows the key used to compute the MAC. We say that the receiver needs to be sure that the packet is safe.i.e. that the sender did not yet disclose the key that was used to compute the MAC of an incoming packet. As stated above, the sender and receivers need to be loosely time synchronized and the receivers need to know the key disclosed schedule. If the incoming packet is safe, the receiver stores the packet (it can verify it only once the corresponding key is disclosed). If the incoming packet is not safe (the packet had an unusually long delay), the receiver needs to drop the packet, since an adversary might have altered it.

Nodes broadcasting authenticated data. New challenges arise if a node broadcasts authenticated data. Since the node is memory limited, it cannot store the keys of a one-way key chain. Moreover, re-computing each key from the initial generating key $K_n$ is computationally expensive. Also, the node might not share a key with each receiver, so sending out the authenticated commitment to the key chain would involve an expensive node-to-node key agreement.

## III. Improvement of the SPINS Protocol

The following will discuss the protocol from three aspects: firstly, the establishment of the network, secondly, the correspondence of network security, thirdly, changes of the security network system. What is more, we will discuss the key timing update mechanism.

*(1) The Establishment of the Network*

Prior to the establishment of the network, the sensor nodes and base stations share ID number for each node, share the encryption key $K_{enc}$, the authentication key $K_{mac}$ as well as a one-way hash function F. Besides, ID, $K_{enc}$ and $K_{mac}$ for each node are different from the other nodes. In addition, the base station saves key $K_t$, which changes with the correspondence periodic cyclical. The periodic change of $K_t$ through a one-way hash function F (x), when the node joins the secure system, the base station will send the current communication cycle $K_t$ to the node. The key $K_t$ has the following four functions:

At the beginning of each correspondence cycle, the encryption key of node $K_{enc}$ and the current cyclical key $K_t$ carries on the AND operation, the result updates the $K_{enc}$;

At the beginning of each correspondence cycle, the authentication key of node $K_{mac}$ and the current cyclical key $K_t$ carries on the AND operation, the result updates the $K_{mac}$;

When cluster head is elected, broadcasts its elected information to the entire network and states the current cycle $K_t$ as the encryption key.

Each cycle, the key $K_t$ is the broadcast encryption key of the base station.

In the update of the encryption key and authentication key, if we use the following methods as in (12) and (13):

$$Kenc = Kenc \text{ XOR } Kt \quad (12)$$

$$Kmac = Kmac \text{ XOR } Kt \quad (13)$$

Among them, $K_{enc}$ is the updated encryption key; $K_{mac}$ is the updated authentication key.

However, XOR has the following properties as in (14) (15) and (16):

$$A \text{ XOR } B = C \quad (14)$$

$$C \text{ XOR } A = B \quad (15)$$

$$C \text{ XOR } B = A \quad (16)$$

If the adversary takes advantage of this nature, they may capture $K_{enc}$ and $K_{mac}$ to speculate $K_t$. Thus the data which has not renewed may be imitated, so there is the possibility of DoS attacks. Therefore, we use the following method to update the key as in (17) and (18):

$$Kenc = Kenc \text{ AND } Kt \quad (17)$$

$$Kmac= Kmac \text{ AND } Kt \quad (18)$$

The AND operation has not such nature like XOR operation, therefore, can enhance the security to a certain extent. Thus, it reduces the possibility of the Dos attacks.

With the participation of $K_t$ and function F, encryption keys and authentication key of the node are not always the same, but change dynamically. The broadcasts of the base station can be transmitted cipher text with dynamic key. These measures will enhance the security throughout our network undoubtedly.

*1 The Phase which Nodes Join in the Network:*

Since the base station sends information to the entire network from time to time, the nodes need trust the site, so the nodes must be able to determine the source message is broadcasted from the base station effectively. So at the beginning of the establishment of secure system, when the node joins the network, it is necessary to be able to certify the base station broadcast packets.

Security initialization of the base station: Reference to the μTESLA protocol, the base station will generate the key pool and determine the synchronization clock of the key when initialized. Assume that the size of the key pool is N, the synchronization cycle of the key is T, and the initial key is $K_N$, the key $k_{i+1}$ in the pool can use $k_{i+1}=F(k_i)$, where F is s a one-way key function stored in the base station.

*Nodes join the network*:

Assume that the node A requests base station S to join the network.Reference to the μTESLA protocol, the specific procedures described below as in (19) and (20):

$$A->S: (NM \mid RA), MAC (Kmac, NM \mid RA) \quad (19)$$

$$S->A: (Ts| Ki |Ti| Tint| d) Kenc, MAC (Kmac, NM |Ts| | Kt | Ti| Tint | d) \quad (20)$$

Among them, $N_M$ is a random nonce that use a strong fresh authentication; $R_A$ is a requesting packet to join the network, which contains the ID number of the node A; $K_{mac}$ and $K_{enc}$ are respectively authentication key and encryption key between node A and the base station. $T_s$ is the current time, $K_i$ was a key the previous time interval uses, $K_t$ is the shared key of the current communication cycle; $T_i$ is the start time of the current synchronization interval; $T_{int}$ is the synchronization interval: d is the delay of key publishing dimensions, $T_{int}$ is the unit.

After such a certification process, the nodes will receive all the information about the authenticated broadcast, and all sensor nodes join to the secure system, thus the nodes can accept a variety of control and data information of the base station.

*2 The Phase of the Establishment of the topology:*

The security risk of LEACH algorithm [17] for topology generation process is that an illegal node may impersonate the cluster header, and publish the messages informing to other nodes, causing the other sensor nodes select to join the cluster with the impersonated cluster header. In the hierarchical topology network, the role of cluster header is self-evident. Any hazards involving cluster head is the most destructive.Therefore, add the identity of the cluster head is essential. Our design has improved the LEACH algorithm: when the base station receives the message that a node states itself as a cluster header, carrying on the authentication according to its ID number and the authentication key. Remove the illegal nodes and broadcast the legitimate information. We use the μTESLA protocol to realize the authenticated streaming broadcast. When we select the cluster head, the remaining energy of a node is also used as a reference value. If residual energy of the node doesn't meet the energy required, the node will not be elected. This makes it possible to avoid the situation that the cluster head suddenly failed because of lacking energy, the situation may lead the entire cluster does not work. The specific process of the topology establishment is as follows:

When a node is selected as cluster header under the principles, it will broadcast messages to announce his election information. The messages will be encrypted with the current traffic cycle key $K_t$, and use the current authentication key $K_{mac}$ to generate message authentication codes;

Other nodes will collect all the election messages of the cluster headers. Record and store the signal strength of each message, and choose the strongest signal (nearest) cluster head. The nodes will wait for the broadcast of base station certification;

The base station receives all the broadcast messages of the cluster heads, and use the shared authentication key $K_{mac}$ to verify its legitimacy. Then the base station will broadcast the legitimate cluster head node to the entire network;

When general nodes receive the broadcast messages, they will use μTESLA protocol to authenticate and verify that the selected clusters whether found in the legitimate cluster heads list. The illegal cluster head will be discarded until that we find a legitimate one. Then the nodes will send requests message to the cluster to join in;

After the clusters receive all the messages of the nodes within the specified time, they will broadcast timetable for uploading the data of the specified node member.The timetable specifies the communication cycle time of the nodes in the cluster. Thus we can avoid a conflict within the cluster member. The nodes which not upload the data come into a "sleeping" state, this way the nodes will save some energy. The cluster is elected on a regular cycle, so the topology will be updated regularly.

*(2) Correspondence of the Secure Network*

Once the topology is established, that is, the secure communication system is established. The network

begins to work properly. Generally, the communication phases of the network are the following two types:

*1 Correspondence between the members of the cluster and the cluster head node:*

Member nodes in the cluster calculate the current communication cycle encryption key $K_{enc}$ and authentication key $K_{mac}$. We encrypt the induction data with $K_{enc}$, and produce news authentication code MAC with $K_{mac}$. During the period of the specified time in the timetable, upload the messages to cluster together with the node ID number.we add a counter in the message packet (after sending a message, counter value will plus one).Thus we can suppress the replay attacks and support the weak freshness of the data communication.

*2 Correspondence between the cluster head node and the base station:*

The cluster head node receives the information from the members of the cluster nodes and makes a certain amount of data information into a message packet, together with the ID number; then it will upload the packet to the base station. Base stations and nodes share the ID number, encryption key $K_{enc}$ and authentication key $K_{mac..}$ The base station will use the ID number and $K_{mac}$ to verify the upload information, the illegal message will be discarded, and the legitimate message will determine the freshness of the data according to the value of the counter. The replay message will be discarded; we will use $K_{enc}$ to decrypt the legitimate data.

*(3).Changes of the Secure Network System*

*1 New nodes join the network***:**

Before a new node joins the network, firstly, the node will be distributed unique ID number, $K_{enc,}$ $K_{mac}$ and one-way hash function F. This information pre-stored in the base station and the new node for the communication. This method can realize the confidentiality, integrity and freshness of the communication and the certification of the base station to the node. When the request of new node is sent to the base station, the base station will confirm its identity and then send the authenticated information and the current cycle of shared key $K_t$.Besides, it will determine the synchronization clock.

*2 The Node is failed:*

We design each sensor node with a energy bottom line, when members of the cluster node energy is running out to reach the energy of the bottom line, uploading messages to inform the base station through the cluster head immediately. Base station will delete this node, thus, this node is removed from the network, and the base station no longer accepts the data from the node. This method can avoid the possibility that the malicious nodes pretend to participate the normal operation.Since the elected cluster head node has the required energy as the cluster head, it is generally not appear that the cluster head node failure during normal operation.

*(4). Key Timing Update Mechanism*

SPINS has not provide the mechanism to renew the key. When the nodes have been deployed, the essential encryption key and the MAC key do not change. If we use the key consultation to produce the new key, may to be able to have the massive correspondence expenses and delay. This is not suit in WSNS. So we use the key self-updating mechanism based on shared variables. In the SPINS protocol, the communication between node i and base station described as follows in (21):

$$i\text{->}B: \{data\}\ (Kenc, C), MAC\ (Kmac, C|\{data\}\ (Kenc, C) \quad (21)$$

The transmission data takes the counter pattern (CTR) encryption, $K_{enc}$ is the bilateral encryption key, $K_{mac}$ is the bilateral MAC key and C is the counter, serves as the encryption initial vector (after sending or receiving a message, C will plus one). While in the SPINS protocol, $K_{enc}$ and $K_{mac}$ will not change during the lifetime of the network. So we add key update mechanism to the SPINS to ensure that the keys are different. Due to limited WSNS resources, if we use key exchange in the network, it will consume a certain amount of energy, so it is not suitable for the application of WSNS. A good way to reduce the network load is that the two sides shared a variable value and the original key. We use the following method to update the key as in (22) and (23):

$$Kenc=F\ (1)\ (Kmaster|C) \quad (22)$$

$$Kmac=\ F\ (2)\ (Kmaster|C) \quad (23)$$

The initial value of C is 0, its value will plus one after a message, Therefore, different message will use different encryption and MAC authentication keys. This method increases the security of SPINS to a certain extent. Besides, the sender and the receiver share a same C to keep pace.

To ensure the C must not be repeated, we use 4-Byte (32 bit) as the length of C. We can prove that the 4 Byte C will not be duplicated.The 32-bit C can be expressed $2^{32}$ message packets.Assume that the length of each code is 30 bit and the transmission speed of the network is 250kbps (this is the present speed of WSNs).Then the node can send data in a continuous cycle at least (2 * 30)/ (250 *2 *60 *60 *24)=5.8d. This figure was enough for a wireless sensor network. During the lifetime of the network, we can ensure that C does not repeat.

## IV. Security Analysis

The wireless sensor network combines a variety of advanced technology; it has incomparable advantages to traditional network. However, due to many restrictions that exist in wireless sensor networks, the security problem is difficult to solve.

Reference to the SPINS secure system our article has designed the secure communication agreement based on the topology wireless sensor network. Regarding the establishment of topology, we utilize the cluster election method of LEACH algorithm and increase the safety mechanism. In the SPINS framework agreement, the encryption key and the authentication key are throughout invariable in the network course of communications. In our design, we join a one-way function and a shared key separately in the base station and the node. Thus we

strengthen the correspondence security. Our security analysis has two aspects:

*(1) Research of the Safety Communication Protocol*

The security requirements of wireless sensor networks are the same with the general network. It should solve the problem of confidentiality, integrity, freshness, and authentication. The distinctive characteristics of wireless sensor networks, such as limited communication skills, limited supply of energy, making their network secure problems solved with different idea to overcome compared with traditional network. It cannot directly use the traditional network secure mechanism, so it is necessary to design secure communications for wireless sensor network. SPINS secure system is popular in the proposed secure mechanisms. It fully considers the security requirements of network. We propose a design scheme of secure communication protocol for wireless sensor network based on the hierarchical topology referring to SPINS.

*(2)The Design of Secure Communication Protocol*

Our design utilizes the LEACH algorithm. We add the authentications of the cluster-heads through base station to ensure the validity of the cluster-heads; in addition, we add a one-way hash function and a shared key in the key management. In our design, the encryption key and authentication key change dynamically to enhance the security of network communications. So our design is more secure and available.

## V. PERFORMANCE ANALYSES

*(1).Performance Analyses of the Improved Protocol*

When a node joins the network, it needs to send request to the base station and receive the response message of the base station. This way adds the cost of sending and receiving one more message, but each node only need once to join the network, relative to the time of the entire system, the cost is negligible.

Obviously, each communication cycle will increase some cost:

1 At the beginning of each communication cycle, nodes will update their own encryption key and authentication key, so it increases the cost of calculations.

2 When a node becomes a cluster head, it will broadcast message to the entire network. But the message is encrypted and comes with a MAC code. Compared with normal non-encrypted message, it increases the length of the message and also introduces the encryption algorithms, which also increases the cost of calculations.

3 Each correspondence cycle, the ordinary node also needs to receive the package of new cluster form the base station and carries on the authentication for the news. This is also increase the energy consumption.

Assume that the length of each message without encryption or authentication is 36 bytes, the length of each message with encryption and authentication is 41 bytes, among them 29 bytes is the data portion. Assume that the Mica2 [15] nodes send and receive messages consume 16.25μJ/byte and 12.25μJ/byte energy. Encrypt, decrypt and certify 8 bytes consumed 15μJ energy [16].

Assume that there are 1010 nodes in the network, including 10 cluster head nodes. Each cluster there is 100 cluster nodes. Each cycle every node will send 10 induction messages.

The consumption estimates that do not use secure communication protocols of each communication cycle.

Members of the cluster nodes: cluster members receive the elected cluster head broadcast a message: 12.25*36*10=4410; a member of the cluster to the cluster head to send request to join: 16.25*36 =585; cluster member receives cluster head news broadcast schedule: 12.25*36=441; induction of cluster members to send a message to the cluster head: 16.25*36*10=5850; cluster member nodes of a communication cycle total energy consumption: 4410 +585 +441 +5850 =11286μJ.

The consumption estimates that use secure communication protocols of each communication cycle.

Members of the cluster nodes: the beginning of the communication cycle update key: 16*15/8=30; cluster members to receive the elected cluster head broadcast news: 12.25*41*10=5022.5; receiving legitimate message from base station broadcast cluster head: 12.25*4l= 502.25; use μTESLA agreement to certify legitimate decryption: 12.25*41+29*15/8+29*15/8+29*15/8=665.375; cluster members sent request to the cluster head to join: 16.25*36=585; cluster members to receive the news of the cluster head broadcast schedule: 12.25*41+29*15/8= 556.625; cluster members send a induction message to cluster head: (16.25*4l+29*15/8 +29*15/8)*10 =7750; cluster member nodes of a communication cycle total energy consumption: 30+5022.5+502.25+665.375+585+ 556.625+7750=15l11.75μJ.

Cluster member nodes in each communication cycle consume more than 33.9% energy. However, as the size of network increases, the increasing energy consumption of cluster head will show a downward trend. When the network is under attack, we are able to resist the attack that malicious nodes inject illegal data or even destroy the entire network. So we think our consumption is worthy.

*(2) Performance Analyses of the Key Update Mechanism*

The key update mechanism ensures that the keys are different. So it increases the security of the network to a certain extent.

| | The SPINS protocol | The improved protocol |
|---|---|---|
| update key | | 16*15/8=30 |
| receive the elected cluster head broadcast | 12.25*36*10=4410 | 12.25*41*10=5022.5 |
| receiving legitimate message | | 12.25*41= 502.25 |
| certify legitimate decryption | | 12.25*41+29*15/8+29*15/8 +29*15/8=665.375 |
| send request to join | 16.25*36 =585 | 16.25*36=585 |
| receives cluster head news broadcast | 12.25*36=441 | 12.25*41+29*15/8= 556.625 |
| send a message to the cluster head | 16.25*36*10=5850 | (16.25*41+29*15/8 +29*15/8)*10 =7750 |
| total consume energy | 11286μJ | 30+5022.5+502.25+665.375+585+556.625+7750=15111.75μJ |

Figure 3. The energy consumption comparison of cluster members

The performance analysis shows that our improvements on SPINS are reasonable and feasible.

## VI. Conclusion

SPINS is popular and practical at present. It takes into account the context of data in confidentiality, integrity, fresh, certification and other aspects. However, its essential encryption key and the MAC key do not change. So it does not solve Dos attack. Besides, its security is fully dependent on the base station. Our improved design is based on the hierarchical topology. Our design can ensure the key changes dynamically. Thus, it reduces the possibility of the Dos attacks. Although our design may use a little more energy compared to the SPINS, the security of our design is much better than the SPINS. And the improved protocol consumed slightly more energy compared to SPINS, but safety performance has improved a lot. All in all, our design is a more reliable and secure than the SPINS.

## Acknowledgment

Acknowledgments: The authors would like to thank anonymous reviewers that provide helpful suggestions for greatly improving the presentation of the paper. Part of the work in the paper is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 61073009, 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] I. Akyildiz, W. Su and Y.Sankarasubramaniam, “A survey on sensor networks”, IEEE Commun. Mag., vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] Z.M. Fadlullah, T.Taleb, “combating against attacks on encrypted protocols through traffic-feature analysis”, IEEE/ACM Transactions on Networking 18, pp. 1234–1247, 2010.

[3] Li, C.T, Lee, C.C, “A novel user authentication and privacy preserving scheme with smart cards for wireless communications”, MathComput. Modelling 2011.

[4] Nyang, D.H, Lee, M.K, “Improvement of Das’s Two-Factor Authentication Protocol in Wireless SensorNetworks”, at: http: //eprint.iacr.org/2009/631.pdf/ (accessed on 07 July 2010).

[5] Khan, M.K.; Alghathbar, K, “Cryptanalysis and security improvement of ‘two-factor’ user authentication in wireless sensor networks”, Sensors 2450-2459, vol. 10, 2010.

[6] He, D., Gao, Y., Chan, S., Chen, C., Bu, J, “An enhanced two-factor user authentication scheme in wireless sensor networks”, Int. J. Ad-Hoc Sensor Wirel. Netw. vol. 0, pp. 1-11. 2010.

[7] He, D.J., Mab, M.D., Zhang, Y., Chen, C., Bu, J.J, “ A strong user authentication scheme with smart cards for wireless communications”, Comput. Commun., vol. 34, pp. 367-374, 2011.

[8] Sullivan, B, “Preventing a Brute Force or Dictionary Attack: How to Keep the Brutes Away from your Loot”, at: http: //h71028.www7.hp.com/ERC/cache/568358-0-0-0-121.html/ (accessed February 2010).

[9] Vaidya, B., Park, J.H., Yeo, S.S., Rodrigues, J.J.P.C., “Robust one-time password authentication scheme using smart card for home network environment”, Comput. Commun, vol. 34, pp. 326-336. 2011.

[10] PERRIGA, SZEWCZYKR, WENV, et al, “SPINS: security protocols for sensor networks”, Wireless Networks, vol. 8, no. 5, pp. 512-534. 2002.

[11] ZigBee Alliance, ZigBee document 053474r17, version 1.0[S/OL]. (2008-01), http: //www.zigbee.org.

[12] Raghavendra V, Kulkarni, Senior Member, “Computational Intelligence in Wireless Sensor Networks: A Survey”, IEEE, Anna Förster, Member, IEEE and Ganesh Kumar Venayagamoorthy, Senior Member, IEEE Communications Surveys& Tutorials, vol. 13, no. 1, First Quarter 2011

[13] Perrig, R, Canetti, J. Tygar and D. Song, “Efficient authentication and signing of multicast streams over lossy channels”, in: IEEE Symposium on Security and Privacy (2000).

[14] WR Heinzelman, eta1, “Ener-efficient communication protocol for wireless microsensor networks”, Proc.of 33 rd Annual Hawaii Inter Conf on System Sciences. Hawaii, USA: IEEE Computer Society, pp. 20-30. 2000.

[15] Jason Hill, Robert Szewczyk, Alec woo, etal, System “Architecture direction for networked sensors”, In Proceedings of ACM ASPLO IX, pp. 93-104, 2000.

[16] Leonardo B, Oliveira, Hao CW, etal, LHA-SP, “secure protocols for hierarchical wireless sensor networks”, Integrated Network Management, 2005.IM 2005.2005 9th IEIP/IEEE International Symposium on 15-19, pp. 31-44. 2005.

# Semantic MMT Model based on Hierarchical Network of Concepts in Chinese-English MT

[1, 2] Wen Xiong
[1] Beijing Normal University, Institute of Chinese Information Processing,
[2] CPIC-BNU Joint Laboratory of Machine Translation, Beijing, China
Corresponding author Email: stevens7979@sina.com

[1, 2] Yaohong Jin
[1] Beijing Normal University, Institute of Chinese Information Processing,
[2] CPIC-BNU Joint Laboratory of Machine Translation, Beijing, China
Email: jinyaohong@bnu.edu.cn

***Abstract*—To study the generation of the semantic tree of Chinese sentence in Chinese-English Machine translation (MT), a new semantic-analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on the hierarchical network of concepts (HNC) is proposed. Supported by word and rule knowledge-base of HNC, the model executed the semantic analysis using static and dynamic labels as a complex feature of MMT instead of a single feature of phrase structure grammar, and generated a HNC-MMT semantic tree for deep understanding of the semantic of Chinese sentence. Based on the semantic tree generated, the model can realize structure conversion of semantic chunks, and utilize a hybrid strategy of the statistical and rule-based to translate. Experiment shows one of the most important tasks of semantic analysis of Chinese sentence, the global eigen-chunk recognition, achieves accuracy above 85%, verifying the effectiveness. The model has been applied to system development of MT based on HNC.**

***Index Terms*—machine translation (MT), semantic analysis, hierarchical network of concepts (HNC), multiple-branched and multiple-labeled tree (MMT), syntactic analysis**

## I. Introduction

As a sub-field of computational linguistics, machine translation (MT) is a research using computer software to translate text or speech from one natural language to another. Natural language understanding (NLU) and natural language processing (NLP) are two important means and technologies used by MT. The researches of MT can be divided into two categories [1], i.e., rule-based and corpus-based. The former includes transfer-based [2], interlingua-based [3] [4], constraint-based formalisms [5], principles-based [6], and lexicalist approaches [7], and the latter includes statistics-based [8], example-based [9], and connectionist approaches [10]. In recent years, there is a new hybrid machine translation (HMT) emerged [11], which combines the advantages of rule-based and statistics-based approaches, including deep integration and shallow integration. For example, there are methods of rule post-processing by statistics and statistics guided by rule.

The statistics-based MT has gained good effectiveness. However, there are still some cases, which do not meet the semantic constraints in the translations. Thus, it is necessary to understand source text deeply using the NLU technique to improve the quality of the translated texts, involving the semantic analysis of sentences of the source texts. In this case, the rule-based method is more conducive to merging the expert's knowledge to analyze the semantic info of source sentence.

Hierarchical network of concepts (HNC) [12] is an important theory for semantic analysis of Chinese, which expresses the knowledge of a linguist with word knowledge-base of HNC and rule knowledge-base of HNC. The semantic analysis in HNC is to capture the deep semantic structure and semantic expression of Chinese sentence using the HNC knowledge in the process of NLU.

In this paper, we propose a novel semantic-analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on HNC in Chinese to English machine translation, which parses the semantic of Chinese sentence and generates HNC-MMT of Chinese sentence using the semantic symbol and analysis means of HNC. The semantic HNC-MMT can be used in the generation of translated text. The rest of the paper is organized as follows. Section 2 surveys the related work on MMT, MT based on HNC, and HMT. Section 3 introduces the MMT model of Chinese information processing (CIP) and the label content of it. The proposed HNC-MMT is presented in detail in Section 4. Section 5 provides the experiment conducted. Finally, Section 6 concludes the conclusion of the paper.

## II. Related Work

The analytical capacity of phrase structure grammar is limited, and the generative capacity of it is too strong, resulting in an ungrammatical content generated. Feng *et al.* presented MMT model of Chinese information [13] to overcome the defects of the phrase structure grammar and


doi: 10.4304/jnw.8.1.237-244

to meet the requirements of the computer processing of Chinese in 1983. Linear and hierarchical structures are shown among linguistic symbols in the same time, which cannot be presented by a single feature. Thus, functional unification grammar [14] was proposed, which parses the sentence for the multiple characteristics of the linguistic symbols. The translation of Chinese sub-sentence is a difficulty in the MT of Chinese-English patent, thus, the model of degraded sentence was introduced, and related rules were proposed [15] to form the translation algorithm for the Chinese sub-sentence. For the solid expressions in MT of Chinese-English patent, especially for the expression of the claim sentence of Chinese, the rule-based method [16] can improve the quality of translations effectively.

The selection of the main verb in the case of multiple verbs used together, and the decision of the boundary of the long noun-phrase (NP) are still two difficulties of the syntactic analysis of Chinese. However, the semantic analysis based on HNC theory and the use of principle of language-logic dynamic-representation (lv) [17] can handle the two situations effectively. Integrated with the system of the syntax trees, the research formed a hybrid-strategy method, which can help to improve the performance of the MT of Chinese-English patent. Passive voice is often appeared in patent document. Thus, it is an important step of MT to recognize the corresponding Chinese sentence having a passive expression, and then to handle the generation of translated text using a rule-based method [18]. Element sub-sentence is widely existed in Chinese patent documents. Most of the problem of Chinese-English MT about the element sub-sentence can be perfectly resolved on an online system of patent MT in SIPO by the analysis of semantic structure of element sub-sentence in three types, and the proposed rules of Chinese-English MT [19].

The analysis of long Chinese sentence is the fourth difficulty. Thus, the method can improve the performance of MT of Chinese-English patent by the segmentation of long Chinese sentence using features coming from HNC theory [20]. The annotated method of corpus of sentence-level semantic [21] annotates the semantic info in three categories: sentence category, semantic chunk, and sub-sentence included in the semantic chunk, which is a degradation of sentence or a chunk extension. The corpus annotated by the method has been an important knowledge resource for Chinese information processing and foreign language study.

There are different directions in hybrid methods. For example, a hybrid approach is to integrate information from a rule-based machine translation system into a statistical machine translation framework [22]. The techniques of hybridization are grouped into three parts: the morphological, lexical, and system level. As an opposite direction, another method is to add statistical bilingual components to a rule-based system [23], which has a higher degree of grammaticality than a phrase-based statistical MT system, where grammaticality is calculated according to correct verb-argument implemented and translation of long-distance dependency.

## III. MMT Model of CIP

One of the most important features of the MMT model is the adoption of *multiple-labels* to describe the Chinese sentence, which can be seen as a *complex-feature* to describe the Chinese sentences. In the MMT model, the concept of "*multi-value label-function*" was proposed. Many MT systems have adopted the Chomsky's phrase structure grammar as a theoretical basis of system design. According to the phrase structure grammar, each node of the syntax-tree obtained by parsing only has one corresponding label, which can be expressed by *single-value label-function* L as: L (*node*) = *label*, where the *node* expresses a node of the syntax-tree, and the label represents a corresponding *label* of the node. The relation between the node and the label is a many-to-one relationship. Because the language feature represented by the single-value label-function is quite limited, a large number of ambiguous structures are generated for long sentence. Thus, there will be many different syntax-trees for the same long sentence, which has brought great difficulty to disambiguation and a significant increase in the overhead of parsing. Unlike the phrase structure grammar, MMT model uses *a multi-value label-function* to replace the single-value label-function, which can be expressed as: M (*node*) = {$label_1$, $label_2$,..., $label_n$}, where the label of a node is no longer one, but corresponds to multiple labels {$label_1$, $label_2$,..., $label_n$}. Thus, the method of multi-value label-function improves the ability in knowledge representation and knowledge description. Each node can record as much as possible grammatical and semantic features. These features compose a complex feature to overcome the weaknesses using phrase structure grammar to describe Chinese sentence fundamentally and to become a new kind of syntactic analysis model of Chinese.

Since the model construction of the syntactic analysis is an important composition of knowledge representation from the viewpoint of artificial intelligence, the resolving of this problem can serve knowledge reasoning, and to form finally intelligent decisions.

### A. MMT Definition

The MMT [24] is a knowledge-representation model of syntactic analysis defined as follows:

1) An MMT has and only has one root;

2) If the root has a child node, then, every child node is an MMT;

3) Any node in the MMT has child nodes with the number from *0* to *n*. If a node has not a child node, the node is a terminal node (a leaf); otherwise it is a non-terminal node;

4) The label in each of the nodes of MMT is a set of multiple labels.

### B. Label Content of MMT of CIP

There are two kinds of label in MMT. The label kind is static if it can be given in a lexical independently, or if it is inherent in the word itself, and the label kind is dynamic if it is generated when the relation between words is occurred. Therefore, the dynamic label is added

into the MMT along with the procedure of syntactic analysis gradually.

For automatic analysis of Chinese sentences and automatic generation of syntax tree, MMT adopts a combination method using the static label of word knowledge and the dynamic label of syntax, semantic, and logical function. The content of static labels of MMT is described as follows.

- **The part of speech (POS) tag**: nouns, premises words, position words, time words, difference words, numerals, quantifiers, body sexual pronouns, predicate pronouns, verbs, adjectives, adverbs, prepositions, conjunctions, particles, modal particles, onomatopoeia, and interjection.
- **The phrase type tag**: the verb phrase, noun phrase, adjective phrase, and the number and quantity phrases.
- **The inherent semantic label of words**: images, materials, phenomena, time and space, measure, abstract, attributes, and actions.
- **The inherent grammar label of words**: a different noun required a different quantifier, the different valence of verb, and substance properties.

The content of dynamic labels of MMT is described as follows.

- **The labels of syntactic function**: subject, predicate, object, attribute, adverbial, complement, adnex, the center language.
- **The semantic relationship tags**: agent, patient, dative, involved, moment, period, start time, spatial points, space segment, starting point of space, end point of space, initial state, final state, causes, results, tools, way, purpose, conditions, roles, content, scope, modification, comparison, accompanying, judgments, statements, attached and so on.
- **Logical relationship labels**: argument 0 (deep subject of sentence), argument 1 (deep direct object of sentence), and argument 2 (deep indirect object of sentence).

## IV. HNC-MMT

The HNC theory is founded by Z.Y. Huang [12], which is a theoretic framework of NLU oriented. Based on the researches of traditional sinology and modern linguistics, HNC starts with the semantics of the language, and uses conceptualization, hierarchy, and network as a basic means, getting rid of the shackles of the syntactic analysis using formal language theory. With foundation of language concept space and methods of the formal, HNC links the surface structure and deep semantics of natural language using the sentence category (SC), and is a unified theory of the syntactic, the semantic, and the pragmatic.

HNC theory simulates the cognitive mechanisms of human brain, and divides the cognitive structure of the human brain into local and global types of the associative skeleton. It thinks of the expression of the associative skeleton is the deep fundamental problem of the language. The local association is of a lexical level, and the global association is of a sentence level and a chapter level. The concept representation-system of HNC theory focuses on the expression of abstract concepts, which uses a five-tuple for the diversity of the abstract concept, and uses a network-hierarchical symbol for the connotation of it. Three semantic networks were proposed, such as semantic network of primitive concept, that of basic concepts, and that of logic concept.

HNC thinks of that there is a conceptual space existed in the human brain, which is a base for human to recognize the world and to think. The existence of the conceptual space is a basic assumption of the HNC, and is one of the axioms of the HNC. The language concept-space is a subspace of the whole concept-space, which is used for human to understand and to apply the natural language, and corresponds to the natural language space, so human beings have a common language concept-space and a variety of natural language spaces. Therefore, there is a one-to-many mapping between the language concept-space and the natural language space.

The concepts of HNC are divided into abstract concept and concrete concept. The former refers to the concepts expressed by object do not have physical attributes, and the latter refers to the concepts expressed by the object have a physical property. Compared with the concrete concepts, abstract concepts have primitive and systematic characteristics, so HNC semantic network is designed for abstract concepts, but it expresses the concrete concepts by linking it to the abstract concept. The HNC semantic network is a hierarchical structure; each layer has some nodes, called conceptual nodes, which are labeled with the numbers 0-13, wherein the numbers 10-13 expressed as hexadecimal lowercase letters: a, b, c, and d.

The semantic network of the basic concepts has a total of nine one-level nodes; that is, *j0* expresses sequence and generalized space; *j1* is for time; *j2* is for space; *j3* is for the number; *j4* is for quantity and range; *j5* is for quality and class; *j6* is for the degree; *j7* is for the basic properties of objective; *j8* is for the basic properties of subjective.

The semantic network of primitive concepts has 14 one-level nodes, divided into two categories. The numbers of *0-5* belong to one category, called principally primitive concept, and the numbers of 6-d belong to another category, called primitive concept of extension. The principally primitive concepts consist of six one-level nodes; that is, *0* is for action; *1* is for process; *2* is for transfer; *3* is for effect; *4* is for relation; *5* is for state. The six nodes are the six basic viewpoints for natural language to represent everything as a whole, and the six basic chains for everything happening, developing, and disappearing, called action-effect chain in HNC.

The primitive concepts of extension describe the content including activity of physiological instinct, the mental, thinking, the intellectual, the professional, the social, pursuit, and stipulation. The primitive concepts of extension are designed for the natural language to describe human activity, which are a compound of principally primitive concept and basic concept.

Logic concepts are divided into two categories, such as language-logic concept and basic-logic concept. The former has 12 one-level nodes, including single identifiers of main semantic block, whose purpose is to build a variety of semantic signs and symbols that serve the analysis of SC and apperceiving of semantic-chunk. The basic-logic concept has two one-level nodes: comparison, and basic judgment.

The semantic network of HNC has the basic characteristics of conceptualization, the primitive, hierarchy, and network. Any node in the semantic network corresponds to a concept, but one natural language has not necessarily the word having the meaning corresponding to the concept. The nodes of semantic network must be no ambiguity, but the word may be ambiguous, since a word can have several different meanings.

The concepts represented by the nodes of semantic network are primitive concepts, whose number is finite, but the combinations of them are infinite. The structure of concepts is hierarchical with high level and low level. For example, "*deserved thing*" is high-level concept, and *wealth*, *experience*, *lesson*, and *knowledge* are underlying concepts.

The hierarchy of semantic network embodies the gradual expression approach to the concept in HNC, from high-level to low-level, which helps the establishment of the associative skeleton of concept, and expresses the natural connection between the high-level concept and underlying concept. Since the characteristic of network is inherent in the concept system, there is longitudinal and transverse correlation between the nodes of concept, so the characteristic of network is an important part of semantic network.

HNC expresses the global associative skeleton using semantic chunk and SC, and proposes the axiom of the action-effect chain, which reflects the largest common of everything and describes the basic rule of existence and development of everything in the universe by using six chains: action, process, transfer, effect, relation, and state.

To illustrate the basic content of the human mind, HNC theory introduces the judgment as a supplement of the action-effect chain, which is a response of subjective against objectivity and emotion, therefore, the action-effect chain and judgment fully express the relationship between subjective and objective, the rational and the emotional, and their relations. They constitute the generalized action-effect chain, which are the basis for the classification of principally primitive concept and the basic for the partition of semantic category of sentence (SC).

In HNC, the SC includes 57 basic SCs, 3192 mixed SCs, and more than 10 million compound SCs. 57 basic SCs can be divided into seven kinds, such as the action sentence, the effect sentence, the process sentence, the transfer sentence, the relation sentence, the state sentence, and the judgment sentence, which describes one of the seven parts of the generalized action-effect chain.

The SC mixed by basic SC and describing two or more parts of the chain of the generalized action-effect is called hybrid SC, and the SC mixed by basic SC or hybrid SC is called compound SC in the HNC. The sentence of basic SC or mixed SC has at most one eigen semantic-chunk (that is similar to the role of the predicate of a sentence), and the sentence of compound SC has two or even more eigen chunks.

HNC theory proposed a complete theoretical framework for the engineering realization, including the sentence processing, the sentence-group processing, and the chapter processing. According to the viewpoint of HNC, language translation is a mapping from one natural language space to another, including the analysis and understanding of source language, the generation process of target language, as well as the intermediate process between the analysis and generation. The analysis of source language in HNC analyzes the constitution of the semantic chunks of sentence using the techniques of SC analysis.

The generation of the target language relies on the mapping word knowledge-base from HNC concept symbols to the target language words, and the rule knowledge-base from words to semantic chunk and from semantic chunk to sentence. The intermediate processing includes six sub-processes, such as SC conversion and sentence format conversion, primary and auxiliary transformation of semantic chunk and constitution transformation of semantic chunk, and the reorder of auxiliary chunk and clause, which is a necessary condition for producing high-quality translation.

The machine translation based on HNC theory uses the techniques of SC analysis to activate, to extend, to enrich, to convert, and to store the associative skeleton of concepts in the space of language concept, and completes the mapping from source language to the space of language concept to achieve the understanding of the source language. According to the analysis result of source-language SC, the MT can determine the semantic-chunk type, semantic-chunk number, and semantic-chunk order of target language, and analyzes the internal composition of semantic-chunk according to the expected knowledge and constitution information of semantic-chunk to be translated.

The sentence handled by HNC-MMT is terminated according to Chinese punctuation, such as full stop, question label, and exclamatory. Since there are different situations of single sentence and complex sentence in the unit of sentence, HNC-MMT model is different from MMT of CIP on the content of label and its value owing to HNC-MMT is semantic analysis oriented. Therefore, it adopts more semantic-labels, such as conceptual categories and sentence categories for the sentence structure, which serving for automatic knowledge reasoning, named concept associative skeleton.

### *A. Label Content of HNC-MMT*

There are also two kinds of label in HNC-MMT like in MMT. There are about 261 static labels of HNC, distributed in six categories, which are listed as follows.

- **Class of generalized concept**: dynamic concept, concept of abstract noun, concept of specific objects, people, property, and logical concept.

- **Conceptual categories**: specific concept (associated with persons or things), abstract concept, and ambident concept.
- **Lv attribute**: flag of main chunk, flag of auxiliary chunk, the back flag of auxiliary chunk, connection specifier within chunk, coreference specifier, and logic specifier within sentence.
- **Morpheme**: adjective prefix, person prefix, verb prefix, noun suffix, verb suffix, people suffix, and substance suffix.
- **Pure dynamic-representation verb**.
- **Sentence categories**: generalized role sentence, the number of main chunk, chunk-extension sentence, prototype sentence-degradation, passive voice, lead of sentence-category conversion, concise state sentence, the first main chunk, sentence category applicable to language-logic-0 (l0), symbol of HNC, effect category, basic concept category, composite structure of eigen chunk (EK), SC code, type of concept category, and word form of translation.

There are about 487 dynamic labels of HNC, distributed in four categories.

- **Features of word form**: Chinese character, string, starting Chinese characters of string, end Chinese characters of string, English string.
- **Chunk features**: concept category of chunk, semantic interpretation of chunk, sentence category of chunk, attribute of sentence category, weighted value of EK, semantic expression, result of semantic analysis, the number of child node, level symbol, semantic relation between sentences, share of clause, grade of adjective, tense, style, voice, and deformation of verb.
- **Position Features Used by HNC**: Such as relative position, parent node, child nodes, sibling nodes, a comma at the beginning, a period at the beginning, a comma at the end, a period at the end, right search and locating within a sentence, left search and locating within a sentence, and taking word string.
- **Functions Used by HNC**: Such as function of node Generation, assignment function, function of value transfer, selection function, condition function, equivalent judgment function, consistent judgment function of high-level semantic, function of node deletion, function of eigen-value deletion, function of adding nodes, function of replacing node, and function of moving node.

*B. Implementation*

Directed by HNC theory, the linguist constructed word knowledge-base and rule knowledge-base to express the static word knowledge and dynamic rule knowledge. We used a production rule to express our HNC rules; that is, left of which is one or multiple conditions composed by static or dynamic labels, word-form features, and position features to match one or multiple static knowledge points; right of which is the action needed to be completed after the match of the rule is successful, which adds dynamic labels using HNC functions and static labels attained by the analysis of rules into the HNC-MMT, wherein the new static labels are not given by word knowledge-base initially.

Rules of semantic analysis of HNC-MMT are divided into two categories, which are listed as:

- **Analysis rules**: verb-object (vo) and object-verb (ov) processing, auxiliary chunk generation, recognition of global eigen, generation and weighting of upside load and down load, rule of compound constitutes of eigen, rule of eigen excluding, rule of eigen queuing, rule of eigen generation, rule of format conversion, rule of clause analysis, rule of clause segmentation, and generation of sentence structure.
- **Transformation rules**: conversion of compound constitute of eigen, format conversion, basic format conversion, clause transformation, adjustment of chunk sequence, long distance collocation of chunks, apposition of chunks, and processing of repeat structure.

The flow of processing is always to judge whether the word pointed currently is matched to the content of position *0* in the rule. When matched, the processing matches the left nodes of the position *0* one by one, if succeed, it matches the right nodes of the position *0* one by one. We use nine representative rules to describe the characteristics of the rule knowledge-base as follows.

$$(0)CHN[Ke]+(1)LC\_CC[l01,l02,l1]=> \\ \text{LC\_TREE}(QE,0,0)\$. \quad (1)$$

Where *(0)CHN[Ke]* represents the word *0* pointed by processing currently is a special Chinese character *ke*; *(1)LC_CC[l01, l02, l1]* expresses the word *1* after the word *0* has a special semantic category, such as *l01*, *l02*, and *l1*; *LC_TREE (QE, 0, 0)* indicates that the processing will add a label of chunk before the eigen chunk (*QE*) into the position *0*, which is the position of word *ke* in this case. For example, this rule will be activated by the following three sentences:

1) Chuliqi *ke genju* celiang jieguo lai xuanze.

2) Beiyong dianci *ke jiang* dianyuan wending di tigong gei bianxieshi zhongduan.

3) Qizhong renhe huan *ke bei* renyi qudai.

When the position pointer points the Chinese word *Ke*, after it is a Chinese word "*genju*" ("*jiang*" in 2) and "*bei*" in 3)) which has a word concept-category *l1* (*l02* in 2) and *l01* in 3)), (1) is activated by the match between the three sentences and the left of it, and the position of Chinese word "*Ke*" will be added a *QE* label.

$$(-1)LC\_CC[uv,uu]+(0)LC\_CC[v]=> \\ \text{LC\_TREE}(EU,-1,-1)\$. \quad (2)$$

Where *(0)LC_CC[v]* represents the concept-category of word *0* pointed by processing currently is a dynamic-representation concept *v*; *(-1)LC_CC[uv, uu]* expresses the word *-1* before the word *0* is an adverb (*uv*) or pure adverb (*uu*) concept; *LC_TREE (EU, -1, -1)* indicates that the processing will add a *modification of E* chunk (*EU*)

into the position *-1*, which is the position of word *-1* in this case. For example, this rule will be activated by the following sentence:

4) *Fenbie shezhi* cheng sudong he lengcang liangzhong zhileng fangshi.

Where "*fenbie*" is the word belonging to *uv* concept-category, and "*shezhi*" is the word belonging to *v* concept-category. Equation (2) is activated by the match between the sentence 4) and the left of (2), and the position of word "*fenbie*" will be added an EU label.

$$(0)LC\_CC[v]+(1)LC\_QH[h\$g] => \text{LC\_TREE}(E,0,0)\& \text{PUT}(fp,LC\_E\_SCORE,VOOV)\$. \quad (3)$$

Where *(1)LC_QH[h$g]* expresses the word *1* before the word *0* is a *suffix of noun* (*h$g*); E is an E chunk; *PUT (fp, LC_E_SCORE, VOOV)* represents a father-parent node (*fp*) will be assigned an E score (*LC_E_SCORE*) with verb-object (*vo*) and subject-predicate (*ov*) appeared sequentially (*VOOV*). For example, this rule will be activated by the following sentence:

5) Zai *zhaoshe dian* shi gai gongjian de bei zhaoshe qu ronghua.

Where "*zhaoshe*" is the word belonging to *v* concept-category, and "*dian*" is the word belonging to *h$g* concept-category, (3) is activated by the match between the sentence 5) and the left of (3), and the position of word "*zhaoshe*" will be added an *E-score VOOV*, showing the verb "*zhaoshe*" is dissimilated into a noun.

$$(0)CHN[bi]+(\text{f})\{(1)LC\_CC[u]\ \&\ END\%\} => \text{LC\_TREE}(E,1,1)\$. \quad (4)$$

Where *(f){ (1)LC_CC[u]&END%}* expresses the processing searches for the concept-category adjective (*u*) from the current word rightward, and the word searched must be an end of sub-sentence (*END%*). For example, this rule will be activated by the following two sentences:

6) Juhewu de hanliang bi toushe quyu nei *di*.

7) Lingyi ceng de zheshelv bi jidi *di*.

Where "*di*" is the word belonging to *u* concept-category, and the position of "*di*" is *END%*, (4) is activated, and the position of the word with *u* concept-category will be added an *E* label.

$$(0)\{CHN[bei]\ \&!BEGIN\%\}+(1)LC\_CC[v]=> \text{LC\_TREE}(QE,0,0)\& \text{PUT}(fp,VOI,P)\$. \quad (5)$$

Where *(0){CHN[bei]&!BEGIN%}* expresses the word *0* is a Chinese character "*bei*", which does not locate at the start of sub-sentence; *LC_TREE (QE, 0, 0)& PUT (fp, VOI, P)* represents *QE* label will be added to the position of "*bei*", and the *fp* of it will have a passive voice (*VOI, P*). For example, this rule will be activated by the following sentence:

8) Zhege quduan *bei xiugai*.

Where "*xiugai*" has a concept-category *v* and word "*bei*" is within the sub-sentence.

$$(0)CHN[yongyu]+(1)LC\_CC[l01,l02,l1] => \text{LC\_TREE}(EQ,0,0)\& \text{PUT}(fp,LC\_EXP,CENTER\_EQ)+ \text{PUT}(1,LC\_E\_SCORE,V\_COMP)\$. \quad (6)$$

Where *PUT (fp, LC_EXP, CENTER_EQ)* expresses the *fp* will have a semantic label (*LC_EXP*) with a value *center of EQ* (*CENTER_EQ*); *PUT (1, LC_E_SCORE, V_COMP)* represents the word *1* will have a score; that is, *complex structure of E chunk* (*V_COMP*). For example, this rule will be activated by the following sentence:

9) Tiaozhiqi *yongyu dui* yinshua de tuxiang shixian hang zhhe.

"*Dui*" is a Chinese character with concept-category *l02* in the above sentence.

$$(0)CHN[Yizhong]+(\text{f})\{(1)END\%\& CHN[fangfa,zhuangzhi,shebei]\}=> \text{LC\_TREE}(BK,0,1)\& \text{PUT}(fp,LC\_EXP,AN+N)\$. \quad (7)$$

Where *(f){ (1)END%&CHN[fangfa, zhuangzhi, shebei]}* represents the processing finds a Chinese word "*fangfa*" or "*zhuangzhi*" or "*shebei*" at the end of sub-sentence, and appoints the position of the word with *1*; *LC_TREE (BK, 0, 1)& PUT (fp, LC_EXP, AN+N)* expresses the processing will add a *block or chunk* (*BK*) label into the position of "*Yizhong*" and the position of "*fangfa*" (or "*zhuangzhi*", or "*shebei*"), and put *fp* a noun-phrase (*AN+N*) label. For example, this rule will be activated by the following sentence:

10) *Yizhong* jiguangshu jiagong *fangfa*.

$$(0)\{LC\_CHK[L1]\&CHN[yi]\}+(\text{f})\{(1)CHN[fanwei]\}=> \text{LC\_TREE}(ABK,0,1)\$. \quad (8)$$

Where *(0){LC_CHK[L1]&CHN[yi]}* expresses the word *0* is "*yi*" having a chunk label *L1*; *ABK* represents an auxiliary chunk label. For example, this rule will be activated by the following sentence:

11) Jiguangshu *yi* bi suoshu yanmo de guang tousheqi da de *fanwei* zhaoshe dao suoshu yanmo shang.

"*Yi*" has a chunk label *L1* in the above sentence.

$$(0)\{LC\_CHK[L1]\&CHN[Genju]\}+(\text{f})\{(1)CHN[fangfa]\}=> \text{LC\_TREE}(ABK,0,1)\$. \quad (9)$$

Where *(0){LC_CHK[L1]&CHN[Genju]}* expresses the word *0* is "*genju*" having a chunk label *L1*. For example, this rule will be activated by the following sentence:

12) *Genju* qianshu renyi quanli yaoqiu suoshu de *fangfa*.

The total processing flow in HNC-MMT for semantic analysis of Chinese is illustrated as follows.

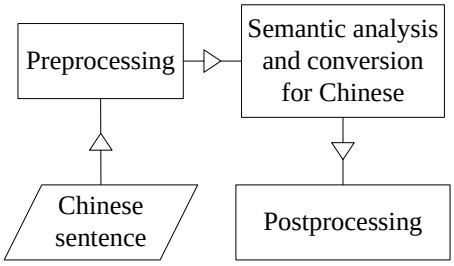



Figure 1. Processing flow of HNC-MMT for semantic analysis of Chinese.

In the above Figure 1, after pre-processing, the Chinese sentence is sent to the module of semantic analysis and conversion. Finally, there is a post-processing, and the results of analysis and translation are given, wherein the pre-processing is illustrated as follows.

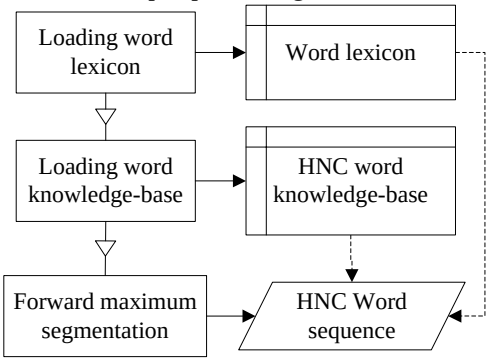



Figure 2. Pre-processing of semantic analysis of HNC-MMT.

In the above Figure 2, the module of pre-processing loads the word lexicon and word knowledge-base, and then segments the Chinese sentence using the forward maximum segmentation to acquire the sequence of HNC words, which includes word form and knowledge-feature of words, that is, static label of HNC. The semantic analysis and conversion for Chinese in the Figure 1 are illustrated in Figure 3.

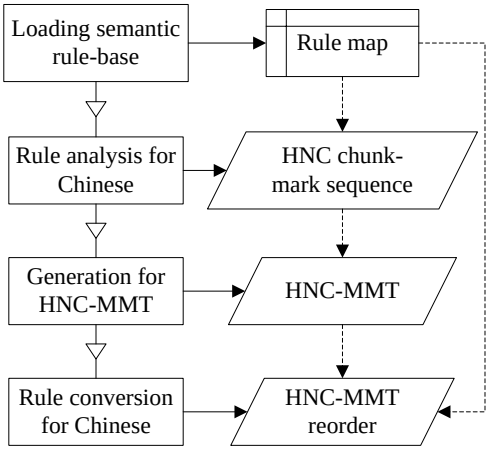



Figure 3. Semantic analysis and conversion of HNC-MMT for Chinese.

In the Figure 3, the module of semantic analysis and conversion loads the semantic rule-base, and executes the rule analysis on the Chinese sentence to generate the HNC-MMT. Finally, a rule conversion for Chinese is performed, which realizes a reorder on the HNC-MMT. After that, a post-processing is performed, illustrated as follows.

In the Figure 4, the module of post-processing executes the English word selection to generate E chunk of English corresponding to the E chunk of Chinese, and performs the transformation of English verb to generate the final English sentence.

## V. EXPERIMENTS

The experimental corpus was 10 bilingual documents of Chinese-English from China Patent Information Center (CPIC). We extracted 1000 Chinese sentences from these documents sequentially as the data set of experiment. The main content of the experiment was the recognition of the global E chunk of Chinese sentence since global E chunk plays an important role in the semantic analysis of Chinese, and we took the accuracy as our measure for the evaluation of experiment. The accuracy is defined as follows:

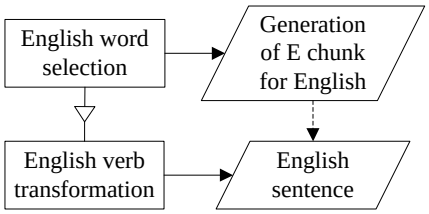



Figure 4. Semantic analysis of post-processing of HNC-MMT.

$$Accuracy = (\#\ recognized\ correctly) / (\#\ needed\ to\ be\ recognized). \quad (10)$$

The experimental result of open test showed the accuracy of global E chunk was *85.1%*, and the throughput of the processing was about *455* Chinese characters per second.

## VI. CONCLUSION

In this paper, we proposed a semantic analysis model of Chinese multiple-branched and multiple-labeled tree (MMT) based on HNC in Chinese-English machine translation, achieved a HNC-MMT of semantic analysis for Chinese using static labels and dynamic labels of HNC, utilized word knowledge-base and rule-base. The experimental result showed the accuracy of E chunk of Chinese sentence was above *85%*. The results generated by the HNC-MMT can be used in the generation of translation, and hybrid methods of MT can be used based on the results. For example, we can send the Chinese chunk into a statistical MT engine, and then integrate the output of the engine, and execute a special processing on the predicate verb to generate the translation. Therefore, the method of multiple strategies hybrid using HNC-MMT will be our future works.

## ACKNOWLEDGMENT

The paper was supported by "the National High Technology Research and Development Program of

China (No. 2012AA011104)" and "the Fundamental Research Funds for the Center Universities."

## REFERENCES

[1] J. Hutchins, “Latest developments in machine translation technology beginning a new era in MT research”, in MT Summit IV, Kobe Japan, 1993.
[2] M. C. McCord, “Design of LMT: a Prolog-based machine translation system”, in Computational Linguistics, 15, pp. 33–52, 1989.
[3] A. Okumura, K. Muraki, and S. Akamine, “Multi-lingual sentence generation from the PIVOT interlingua”, in MT Summit 3, pp. 67–71, 1991.
[4] K. Goodman and S. Nirenburg, “The KBMT project: a case study in knowledge-based machine translation”, San Mateo, Ca.: Morgan Kaufmann, 1991.
[5] K. Martin, “Functional unification grammar: a formalism for machine translation”, in 10th International conference on computational linguistics, Proceedings of Coling84, pp. 75–78, 1984.
[6] B. Dorr, “Parameterization of the interlingua in machine translation”, in Coling 92 (2), pp. 624–630, 1992.
[7] A. Sanfilippo *et al*., “Translation equivalence and lexicalization in the ACQUILEX LKB”, in TMI-92, pp. 1–11, 1992.
[8] P. Brown *et al*., “A statistical approach to language translation”, in Coling 88, pp. 71–76, 1988.
[9] M. Nagao, “A framework of a mechanical translation between Japanese and English by analogy principle”, in *Artificial and Human Intelligence*, Elithorn A and Banerji R (eds.) North-Holland, 1984, pp. 173–180.
[10] A. N. Jain, A. E. McNair, A. Waibel, H. Saito, A. G. Hauptmann, and J. Tebelskis, “Connectionist and symbolic processing in speech-to-speech translation: the JANUS system”, in MT Summit 3, pp. 113–117, 1991.
[11] A. Eisele *et al.*, “Hybrid machine translation architectures within and beyond the EuroMatrix project”, in 12th EAMT conference, Hamburg, Germany, 2008.
[12] Z. Y. Huang, *The HNC (Hierarchical Network of Concepts) Theory-a New Approach to Computer Understanding of Natural Languages*. Tsinghua University Press, Beijing, 1998. (In Chinese)
[13] Z. W. Feng, “Multi-label and multi-branch tree analysis of Chinese sentences”, in Proceedings of ICCIP’83, 1983.
[14] M. Kay, “Parsing in functional unification grammar”, in *Natural Language Parsing: Psychological*. Computational and Theoretical Perspectives, 1985.
[15] Y. H. Jin and W, Xiong, “A sentence degeneration model and its application in Chinese-English patent machine translation”, in 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 421–424, 2011.
[16] W. Xiong and Y. H. Jin, “A new Chinese-English machine translation method based on rule for claims sentence of Chinese patent”, in 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 378–381, 2011.
[17] Y. H. Jin, “A hybrid-strategy method combining semantic analysis with rule-based MT for patent machine translation”, in 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 1–4, 2010.
[18] Z. Y. Liu and Y. H. Jin, “The research of passive voice in Chinese-English patent machine translation”, in 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 300–303, 2011.
[19] Z. Y. Liu, Y. H. Jin, and Y. H. Chi, “Research on element sub-sentence in Chinese-English patent machine translation”, in 2011 International Conference on Asian Language Processing (IALP), pp. 193–196, 2011.
[20] Y. H. Jin and Z. Y. Liu, “Improving Chinese-English patent machine translation using sentence segmentation”, in 2010 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 1–6, 2010.
[21] Z. Y. Liu, Y. H. Jin, and C. J. Miao, “A sentence-level semantic annotated corpus based on HNC theory”, in 2011 International Conference on Asian Language Processing (IALP), pp. 59–62, 2011.
[22] R. Zbib *et al.*, “Methods for integrating rule-based and statistical systems for Arabic to English machine translation”, *Machine Translation*, 26 (1-2), pp. 67-83, 2012.
[23] N. Habash, B. Dorr, and C. Monz, “Symbolic-to-statistical hybridization: extending generation-heavy machine translation”, *Machine Translation*, 23, pp.23–63, 2009.
[24] Z. W. Feng, *Natural Language Processing by Computer*. Foreign Language Education Press, Shanghai, pp. 316–322, 1996.

# Optimized Information Transmission Scheduling Strategy Oriented to Advanced Metering Infrastructure

Weiming Tong
School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, China
Email: dianqi@hit.edu.cn

Xianji Jin, Lei Lu
School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, China
Email: mrking2001@163.com, lulei@hit.edu.cn

***Abstract*—Advanced metering infrastructure (AMI) is considered to be the first step in constructing smart grid. AMI allows customers to make real-time choices about power utilization and enables power utilities to increase the effectiveness of the regional power grids by managing demand load during peak times and reducing unneeded power generation. These initiatives rely heavily on the prompt information transmission inside AMI. Aiming at the information transmission problem, this paper researches the communication scheduling strategy in AMI at a macroscopic view. First, the information flow of AMI is analyzed, and the power users are classified into several grades by their importance. Then, the defect of conventional information transmission scheduling strategy is analyzed. On this basis, two optimized scheduling strategies are proposed. In the wide area, an optimized scheduling strategy based on user importance and time critical is proposed to guarantee the important power users' information transmission being handled promptly. In the local area, an optimized scheduling strategy based on device and information importance and time critical is proposed to guarantee the important devices and information in AMI user end system being handled promptly. At last, the two optimized scheduling strategies are simulated. The simulation results show that they can effectively improve the real-time performance and reliability of AMI information transmission.**

***Index Terms*—smart grid, advanced metering infrastructure (AMI), information transmission, scheduling strategy**

## I. INTRODUCTION

Possessing the features of high reliability, efficiency, scalability and self-healing, smart grid has been rapidly developed during the recent years [1, 2, 3]. AMI (advanced metering infrastructure) is considered to be the first step in constructing smart gird [4]. AMI is a two-way network processing system that used to meter, retrieve, store, analyze, apply the electricity consumption information and remotely control the intelligent devices at the user end. AMI can completely change the status quo of one-way power flow and information flow. It can provide information platform and technical support for the two-way interaction between end users and smart grid [5]. Besides, it also provides fundamental facilities for advanced applications in smart grid.

The goal of AMI is to allow customers to make real-time choices about power utilization and for the utilities to be able to mitigate demand load during peak times, reducing unneeded power generation and increasing the effective capacity of the regional grids. AMI is not singular but integrated technology. It can provide power users and utilities with required information for decision making, required abilities for decision execution and a series of optional functions. Through AMI, power enterprises can precisely maintain the power operation and assets management businesses, thus to provide better service for end users.

The realization of all the above functions relies heavily on the prompt and reliable information transmission inside AMI, especially the real time performance. For example, in order to monitor load shedding and proceed with demand response, the power utilities need to reliably collect time-based metering data and obtain the total amount of power being consumed in a subdivision in real time. In previous automatic meter reading systems (AMR), the data updating interval is usually 15 minutes or more. In AMI, the demand response analysis and control system (DRAACS) need to gather the real time power measurements from meter data management system (MDMS) to help with customer saving and efficient management of utility resources [6, 7]. More quickly data update means more promptly demand response.

To improve the real time performance of information transmission, we can upgrade the communication network to obtain higher communication rate. Another effective way is to adopt reliable scheduling strategies. The work in [8] proposes a performance analysis method of multiuser scheduling system based on block

This work is supported by the National Natural Science Foundation of China (No. 51077015 and No. 50907014).
Corresponding author: Xianji Jin.


doi: 10.4304/jnw.8.1.245-252

diagonalization zero forcing transmission strategy. The work in [11] describes a QoE-driven power scheduling architecture and strategy in smart grid. Some authors research the scheduling strategies based on specific network, like CAN [9, 10], wireless network [12], and fiber optic network [13]. Some authors research the general scheduling strategies for broadcast network [14, 15, 16, 17, 18, 19]. The existing research work mainly focuses on the scheduling strategy in a specific network system at a microcosmic view. However, the communication terminals in AMI are geographically widely distributed in a country or region, and the communication network in AMI always contains several scopes and grades, including wide area network (WAN), local area network (LAN) and home area network (HAN). It is necessary to research the information transmission scheduling strategy in AMI at a macroscopic view.

A reality is that the importance of various power users is different. For instance, an AMI subsystem deployed at a large hospital need to be given more consideration than a residential quarter, thus the power consumption information of the hospital should be transmitted more often. Especially, for some emergent conditions (power loss, tampering and loss of signal from an end device), the alarms information should be transmitted and handled more promptly, or that may cause damages to the patients' health, even cause some serious personal casualties. Furthermore, in an AMI subsystem at the user end, the importance of various intelligent devices is different, and the importance of the information from various devices is also different. For instance, smart meters deployed at the feeder lines and transformers are more important than common electricity meters, electricity meters and load control devices are more important than other field devices (non-electricity meters, field displays, etc.). Besides, among all the information flow between user end devices and AMI head ends, the load shed commands and meter event alarms information possess more importance and value than meter read data and event logs information.

The conventional information transmission scheduling strategy in AMI is periodical polling. It ignores the difference of various users' importance and allocates the same network load to each user, which may cause that the important power users' information cannot be handled in time. It also ignores the difference of the importance of various field devices and information, which may cause that the important information cannot be handled in time. Hence, this paper mainly researches the AMI information transmission scheduling strategy giving consideration to power user/device/information importance and time critical.

This paper is organized as follows. Section II analyzes the information flow in AMI and gives a brief power users classification method according to the importance. Section III states the conventional information transmission scheduling strategy. Section IV proposes an optimized information transmission scheduling strategy based on user's importance and time critical between MDMS and AMI head end in the wide area. Section V proposes an optimized information transmission scheduling strategy based on device and information importance and time critical between AMI head end and intelligent devices in the local area. Section VI demonstrates the two proposed scheduling strategies with examples. Section VII presents the final conclusion of this paper.

## II. AMI INFORMATION FLOW ANALYSIS AND POWER USERS CLASSIFICATION

### *A. AMI Information Flow Analysis*

Generally, AMI is composed of user end intelligent devices, communication network, AMI head end and AMI administration center [20, 21]. The description of each AMI component is shown as follow:

- **Intelligent devices**: deployed at the user end, include smart meters (electricity meters, water meters, gas meters, etc.), DERs (distributed energy resources), load control devices (thermostats, air-conditioning devices) and electric vehicle charging and discharging facilities, etc. Smart meter is the core components. It measures, records, displays, and transmits data such as energy usage, generation, text messages, and event logs to authorized systems. It also serves as a gateway between the utility, customer site, and customer's HAN devices and/or load controllers.
- **AMI communication network**: include various types of network, like HAN used to connect intelligent devices at the user end, LAN used to connect user end systems with AMI head ends, WAN used to connect AMI head ends with MDMS. AMI communication network cover multiple networks and protocols, e.g., optical fiber, 230MHz private network, wireless network, broadband over power line, etc.
- **AMI head end**: usually deployed in local AMI subsystem. It is responsible for the two-way communication between AMI administration center and user end devices to retrieve data and execute commands. It also balances load on the communication network resulting from scheduled meter reads and remotely manages firmware updates, configuration changes, provisioning functions, control and diagnostics.
- **AMI administration center**: contains several business process systems, like MDMS used to aggregate, validate, estimate and permit editing of meter data such as energy usage, generation and meter logs, DRAACS used to send demand response event notifications to meters and load control devices through the AMI system, UIMS used to manage user information, MAMS used to manage meter assets information. MDMS and DRAACS are the core components in AMI administration center.

As shown in Figure 1, AMI information transmission is two-way and interactive. MDMS exchanges information with AMI head ends through WAN (e.g.,

optical fiber, 230MHz private network, GPRS wireless public network, etc.). AMI head end exchanges information with user end systems through LAN (e.g., BPL, PLC, etc.). Intelligent devices at the user end communicate with each other through HAN (e.g., Zigbee, Homeplug, etc.).

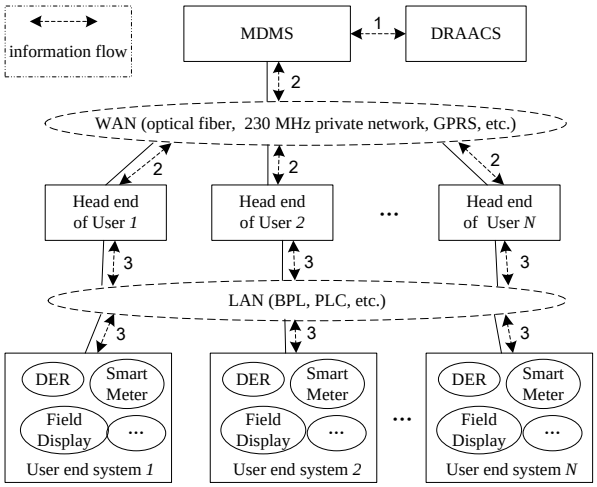



Figure 1. AMI information flow

AMI information consists mainly of meter read data, load shed commands and meter event alarms. Meter read data consist of basic power parameters (e.g. voltage, current, power, energy measurements, etc.), power quality parameters (e.g. harmonic, voltage unbalance, voltage sag, etc.) and non-electricity parameters (e.g., gas and water measurements, etc.). Load shed commands include meter turn on/off, load shed start/end, etc. Meter event alarms include power tampering, outage, restoration and loss of signal, etc. Besides, there are a number of events and error logs transmitted in AMI system. By reference to Figure 1, the information flow in AMI is summarized in Table I.

TABLE I
AMI INFORMATION FLOW DESCRIPTION

| Line # | Direction | Description |
|---|---|---|
| 1 | to DRAACS | Load shed events |
| | to MDMS | Load shed control notifications |
| 2 | to Head end | Meter read requests, load shed commands, planned outage information, HAN equipment commands |
| | to MDMS | Meter read data, various meter events and confirmations, HAN equipment responses, outage and restoration notifications, event logs. |
| 3 | to End user system | Meter read requests, turn on/off commands, pricing data, provisioning requests, firmware updates, prepayment information |
| | to Head end | Meter read data, various meter events (e.g., tampering, outage, and restoration), various confirmations (e.g., meter turn on/off, load shed start/end, and meter provisioning), error logs. |

*B. Power Users Classification*

Power system users can be classified from different angles. According to the voltage level, that can be classified into residential power users (less than 1/10 kV) and large-scale industrial power users. According to the electricity price, that can be classified into industrial users, agricultural users, commercial users, residential users [23, 24]. According to the merchandising locations and channels, that can be classified into direct supply, wholesale, urban and rural users.

To distinguish the importance of various power users, this paper classifies the power users macroscopically into important users and common users according to the power reliability requirement, mainly refers to the politics/ economy/ environment losses and affects that power failure causes. The important users are of great significance and value in the social, politics and economic fields in a country or region. Their power interruptions may cause large personal casualties, environment pollutions, large political influences, large economic losses and serious confusion of social public order.

Further, the important users can be classified into 3 grades by their significance, as shown in Table II.

TABLE II
CLASSIFICATION OF IMPORTANT POWER USERS

| Grade | Description |
|---|---|
| Grade 1 | 1) Of great significance to the country and public security, e.g., governments, large enterprises, hospitals, etc.<br>2) Power failure may cause serious public accidents or personal casualties. |
| Grade 2 | 1) Of large politics or economic significance, e.g., transportation hubs, communication hubs, etc.<br>2) Power failure may cause great economic losses or public confusion. |
| Grade 3 | 1) Of significance to people's social life, e.g., schools, supermarkets, cinemas, etc.<br>2) Power failure may cause economic losses. |

## III. CONVENTIONAL INFORMATION TRANSMISSION SCHEDULING STRATEGY

The conventional information transmission scheduling strategy in AMI is periodical polling, which means that MDMS polls the AMI head ends in area user system at a fixed time in a predetermined sequence. AMI head end polls the user end devices periodically in the same way.

Each area user system need some time to process the communication tasks inside it. Thus, MDMS need to transmit the polling request message to the first user system after its internal communication is completed and the data is updated. Then MDMS polls the next user system in the same way until all the user systems are polled completely. The start time of the next polling cycle should be scheduled as appropriate. If the internal communication in each user system is completed and the data is updated already, MDMS can start polling right away. Else, MDMS need to wait a moment $\Delta t$ to ensure

that the internal communication is completed. The conventional scheduling strategy is shown in Figure 2.

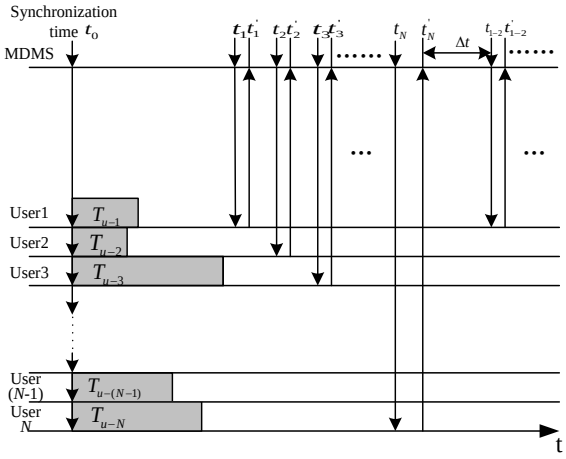



Figure 2. Conventional information scheduling strategy

The communication duration time satisfies the constraints:

$$\sum_{i=1}^{N}(T_i + T_i^{'}) + \Delta t \geq \max[T_{u\text{-}i}] \quad (\Delta t \geq 0) \qquad (1)$$

where $T_i$ is the polling request message transmission time of user *i*, $T_i^{'}$ is the response message transmission time, $T_{u\text{-}i}$ is the duration time of the internal communication inside user system *i*.

Due to the difference of the network structure and information quantity among the area user systems, the duration time of the internal communication in each user system is different. The conventional communication scheduling strategy can only guarantee the basic information transmission and data updating in AMI, but it may bring about the low efficiency and real time performance of the relatively important area users communication. Besides, the time critical of the important information are not given consideration. Hence, this paper proposes two optimized scheduling strategies of information transmission oriented to AMI, which can be used to guarantee the prior information transmitting and updating of important power user in the wide area and important information in the local area.

## IV. Optimized Strategy Based on User Importance and Time Critical

In section II we classify the AMI system users into important users and common users. Further, we classify the important users into three grades. Here we propose an optimized scheduling strategy based on user importance and time critical.

As shown in Figure 3, the important users will be polled several times in one polling cycle. Meanwhile, the common users will be polled only one time.

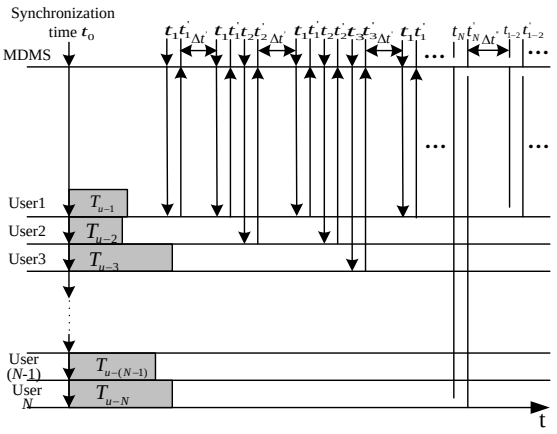



Figure 3. Optimized information transmission scheduling strategy based on user importance and time critical

According to the classification of the important power users, the prescaling factors are defined as follow: $k_1$ (grade 1), $k_2$ (grade 2), $k_3$ (grade 3), here $k_i$ means the polling times of the grade *i* important users in a total polling cycle. The objective functions are

$$\begin{cases} k_1 \geq k_2 \\ k_2 \geq k_3 \\ k_3 \geq 2 \end{cases} \qquad (2)$$

where $k_1$, $k_2$, $k_3$ are all integers.

Suppose that in an AMI system, user 1 is grade 1 important user, user 2 is grade 2 important user, user 3 is grade 3 important user.

The constraints of the communication duration time are

$$\begin{cases} k_j(T_j + T_j^{'}) + \Delta t_j^{'} \geq T_{\text{u-}j}\,(j = 1,2,3) \\ \sum_{j=1}^{3}\left[k_j(T_j + T_j^{'}) + \Delta t_j^{'}\right] + \sum_{j=4}^{N}(T_j + T_j^{'}) + \Delta t^{''} \geq \max\ T_{\text{u-}j}] \\ (\Delta t_i^{'}, \Delta t^{''} \geq 0) \\ T_{\text{op}} \leq (1+\alpha)T_{\text{co}} \end{cases} \qquad (3)$$

where $T_{\text{co}}$ is the total polling duration time under the conventional scheduling strategy, $T_{\text{op}}$ is the total polling duration time under the optimized scheduling strategy, $\alpha$ is the relaxation coefficient (the value is usually between 10%~50%).

The optimized scheduling strategy is dispatched based on user importance and time critical, that allocates more network load to the important users than the common ones. Thus the important users' information can be transmitted several times in a total polling cycle, that improves the scheduling problem between MDMS and area user system, and guarantees the real time performance and reliability of important users' information transmission.

## V. Optimized Strategy Based on Device and Information Importance

The optimized scheduling strategy based on user importance and time critical focuses on the scheduling problem between MDMS and AMI head ends macroscopically in the wide area. In the local area user system, AMI head end communicates with intelligent devices at the user end. The importance of various intelligent devices is different, and the importance of various devices information is also different. For example, smart meters and load control devices are more important than other user end devices, and the load shed commands and meter event alarms possess more importance and value than meter read data event logs information. Thus the important devices (smart meters and load control devices) and important information (meter event alarms and load shed commands) should be allocated more network load during the polling cycle.

Giving consideration to the device and information importance and time critical, this paper proposes an optimized scheduling strategy to resolve the scheduling problem between AMI head end and user end intelligent devices in the local area. The basic principle of this scheduling strategy is stated as follow:

1) First, we classify the intelligent devices at the user end according to the importance and time critical of each device. E.g., the intelligent devices are classified into 3 grades: grade A (e.g., load control devices), grade B (e.g., electricity meters), grade C (e.g., water meters and gas meters). The importance and time critical satisfy that A>B>C.

2) Then, we classify the information of intelligent devices according to importance and time critical of the information itself. For example, the information of device A is classified into 1 grade: $M_A$ (e.g., load shed commands), the information of device B is classified into 2 grades: $M_{B1}$ (e.g., meter event alarms), $M_{B2}$ (e.g., meter read data), the information of device C is classified into 2 grades: $M_{C1}$ (e.g., water and gas meter read data), $M_{C2}$ (e.g., event logs). The importance and time critical is $M_A$>$M_{B1}$>$M_{B2}$>$M_{C1}$>$M_{C2}$. $M_A$ is called the most important information, $M_{B1}$ and $M_{B2}$ are called relatively important information, $M_{C1}$ and $M_{C2}$ are called common information.

3) We make that one total polling cycle contains several sub-polling procedures. That means not all information is polled during one time. The most important information is transmitted in every sub-polling procedure. The relatively important information is transmitted several times in a total polling cycle. The common information is transmitted only once in a total polling cycle.

For example, one possible scheduling strategy is shown in Figure 4. In this figure, one total polling cycle contains 6 sub-polling procedures. The information of device A is transmitted 6 times in a polling cycle. The information of device B is transmitted 2 times in a polling cycle. The information of device C is transmitted only 1 time in a polling cycle.

Assuming that one total polling process contains $m$ ($m$ is a positive integer) sub-polling procedures. The objective function is

$$\begin{cases} i = m \\ k \geq p \end{cases} \quad (i, k, p \text{ are all integers}) \tag{4}$$

where $i$ is the polling times of most important information in a total polling cycle, $k$ is the polling times of relatively important information in a total polling cycle, $p$ is the polling times of common information in a total polling cycle.

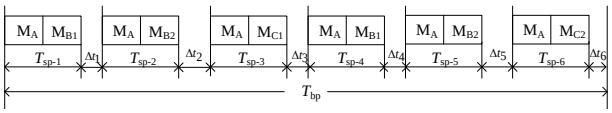



Figure 4. Optimized information transmission scheduling strategy based on device and information importance and time critical

The constraints of the communication duration time are

$$\begin{cases} p = 1 \\ T_{\text{bp}} = \sum_{q=1}^{m} T_{\text{sp}-q} \\ T_{\text{bp}} \leq T_a \end{cases} \tag{5}$$

where $T_{bp}$ is the total polling duration time. $T_{\text{sp}-q}$ is the duration time of sub-polling $q$ ($q = 1,2,\cdots,m$), $T_a$ is the maximum time that common information requires.

Normally $T_{\text{sp}-q}$ is not equal to each other. In some cases, it is required that every sub-polling duration time is equal (the master station proceeds each polling procedure at a fixed time). That requires

$$T_{\text{sp}-1} + \Delta t_1 = T_{\text{sp}-2} + \Delta t_2 = \cdots = T_{\text{sp}-m} + \Delta t_m \tag{6}$$

where $\Delta t_i$ ($i = 1,2,\cdots,m$) is the addition time to make each sub-polling duration time equal.

In this occasion, the constraints are

$$\begin{cases} p = 1 \\ T_{\text{bp}} = \sum_{q=1}^{m} (T_{\text{sp}-q} + \Delta t_q) = m(T_{\text{sp}-1} + \Delta t_1) \\ T_{\text{bp}} \leq T_a \end{cases} \tag{7}$$

In a concrete implementation, to satisfy Formula (6), we can find out the maximum $T_{\text{sp}-q}$ ($q = 1,2,\cdots,m$), e.g. $T_{\text{sp}-r}$. Set $\Delta t_r = 0$, then the other $\Delta t_e$ ($1 \leq e \leq m$, $e \neq r$) can be easily determined. In this case, we can get

$$\begin{cases} \Delta t_e = T_{\text{sp}-r} - T_{\text{sp}-e} \\ T_{\text{bp}} = mT_{\text{sp}-r} \end{cases} \tag{8}$$

## VI. Simulation Results

### A. Wide area AMI System Simulation

In this section, the performance of the information transmission strategy based on user importance and time critical is simulated and evaluated.

The simulation model simulates an AMI system in the wide area, which consists of several components: MDMS, AMI head ends, user end systems, WAN used to connect MDMS with head ends, LAN used to connect head end and user end systems. Here we assume that WAN is TD-CDMA wireless network and the communication rate is 128 kbps, LAN is power line communication network and the communication rate is 19.2 kbps, and there are 20 end users in the AMI system. The simulation settings are stated in Table III.

TABLE III
SIMULATION SETTINGS

| Parameter | Value |
|---|---|
| Number of users | 20 |
| WAN | TD-CDMA (128 kbps) |
| LAN | PLC (19.2 kbps) |
| Duration time | 10 minutes |

Assuming that among the 20 end users, user 1 is grade 1 important user, user 2 is grade 2 important user, user 3 is grade 3 important user, others are common users. The meter data amounts of grade 1/2/3 important users are all 1024 bytes, and the meter data amounts of common users are all 4096 bytes. To simplify the simulation, we set $k_1=k_2=k_3$ in Formula (3), that means we ignore the different grades of the important users and allocate the same network load to grade 1, grade 2 and grade 3 important user.

The simulation runs for 10 minutes to show the performance of conventional scheduling strategy and proposed scheduling strategy based on user importance and time critical. The simulation results are shown in terms of polling times in Figure 5 and network load usage rate in Figure 6.

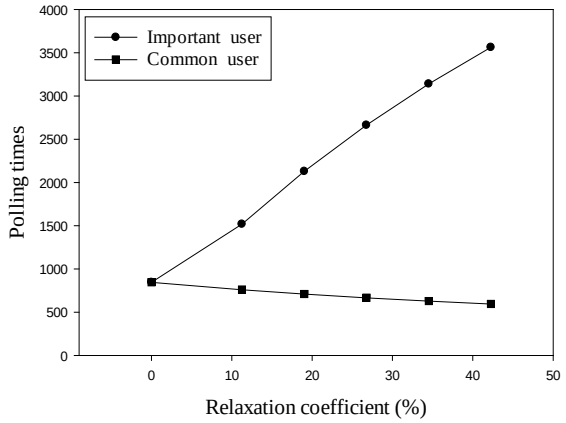



Figure 5. Polling times of important user and common user

From Figure 5 and Figure 6, we can see that the polling times and network load usage rate of important user increase with the increase of relaxation coefficient. Correspondingly, the polling times and network load usage rate of common user decrease with the increase of relaxation coefficient. Compared with the results of conventional scheduling strategy (while the relaxation coefficient is equal to zero), the scheduling strategy based on user importance and time critical can efficiently improve the real time performance of important users' information transmission. Although it causes a certain sacrifice of the real time performance of common users' information transmission, it is still worthwhile.

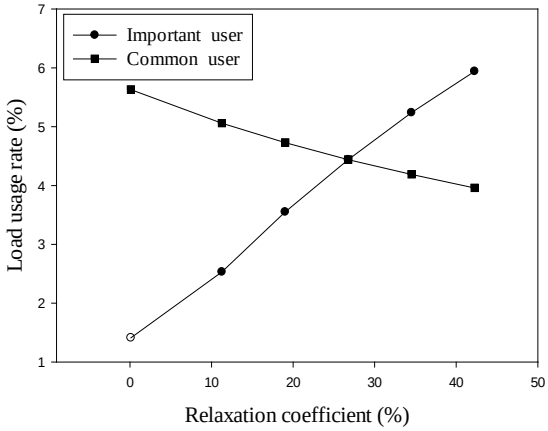



Figure 6. Network load usage rate of important user and common user

### *B. Local Area AMI System Simulation*

In this section, the performance of the information transmission strategy based on device and information importance and time critical will be simulated and evaluated.

The simulation model simulates an AMI system in the local area, which consists of an AMI head end, LAN and user end system. Here we assume that LAN is power line communication network and the communication rate is 19.2 kbps. Field devices in the user end system include 20 load control devices, 20 electricity meters, 20 non-electricity meters. The importance of load control device is bigger than smart meter, and the importance of electricity meter is bigger than non-electricity meter. The simulation settings are stated in Table IV.

TABLE IV
SIMULATION SETTINGS

| Parameter | Value |
|---|---|
| Number of devices | 60 |
| LAN | PLC (19.2 kbps) |
| Duration time | 10 minutes |

Assuming that the meter data amounts of an electricity meter are 64 bytes, the meter data amounts of a non-electricity meter are 32 bytes, the load shed commands amounts of a load control device are 16 bytes. Suppose that all the 16 bytes data of a load control device are the most important information. 16 bytes data of electricity meter are the most important information, the other 48 bytes data are relatively important information. All the 32 bytes data of non-electricity meter are common information.

The simulation runs for 10 minutes to show the performance of conventional scheduling strategy and proposed scheduling strategy based on device and

information importance and time critical. The simulation results are shown in terms of polling times in Figure 7 and network load usage rate in Figure 8.

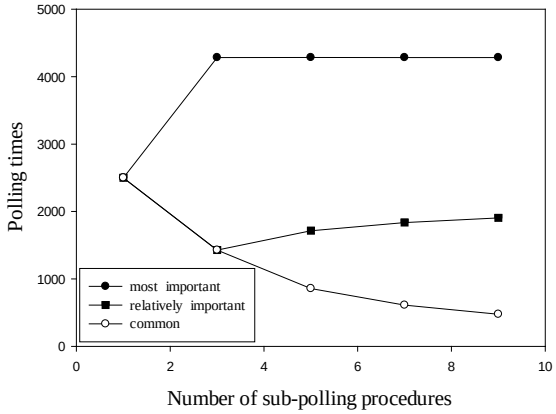



Figure 7. Polling times of most important information, relatively important information and common information

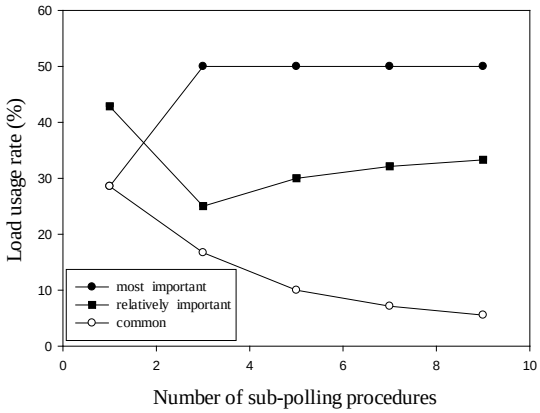



Figure 8. Network load usage rate of most important information, relatively important information and common information

From Figure 7 and Figure 8, we can see that the polling times and network load usage rate of the most important information approach a constant number with the increase of the number of sub-polling procedures. The polling times and network load usage rate of relatively important information increases with the increase of the number of sub-polling procedures. Correspondingly, the polling times and network load usage rate of common information decrease with the increase of the number of sub-polling procedures. Compared with the results of conventional scheduling strategy (while number of sub-polling procedures is equal to 1), the scheduling strategy based on device and information importance and time critical can efficiently improve the real time performance of important information transmission.

## VII. CONCLUSION

This paper focuses on the information transmission problem in AMI in smart gird. The functions of each AMI component are described, the information flow in AMI is analyzed, and the AMI end users are classified by their importance. To improve the real time performance and avoid the defect of conventional information transmission scheduling strategy in AMI, this paper proposes two optimized scheduling strategies: the scheduling strategy based on user importance and time critical used in wide area AMI system, the scheduling strategy based on device and information importance and time critical used in local area AMI system. The objective functions and constraints of the two scheduling strategies are given. The simulation results show that the two optimized scheduling strategies can efficiently improve the real time performance and reliability of information transmission of important power users, devices and information. The research work in this paper can provide reference for other power transmission and distribution systems in smart grid.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 51077015 and No. 50907014).

## REFERENCES

[1] A. Ipakchi and F. Albuyeh, “Grid of the future”, *IEEE Power & Energy Machine*, vol.7, pp. 52-62, July 2009.

[2] .S. Xiao, “Consideration of technology for constructing Chinese smart grid”, *Automation of Electric Power Systems*, vol.33, pp. 1-4, May 2009.

[3] Y. X. Yi and W. P. Luan, “Smart grid and its implementations”, *Proceedings of the CSEE*, vol. 29, pp. 1-8, December 2009.

[4] J. C. Zhang and Z. Y. Chen, “The impact of AMI on the future power system”, Automation of Electric Power Systems, vol. 34, pp. 20-23, January 2010.

[5] W. P. Luan, “Advanced metering infrastructure”, *Southern Power System Technology*, vol.3, pp. 6-10, February 2009.

[6] S. M. Tian and R. W. Xu, “Key technology research of China advanced metering infrastructure”, on *2010 International Conference on Power System Technology*, Hangzhou, China, October 2010.

[7] A. Mehdi, S. Vahid, and A. Behzad, “Advanced metering infrastructure system architecture”, in *Proceedings of 2011 Asia-Pacific Power and Energy Engineering Conference*, pp. 1140-1146, March 2011.

[8] R. P. Yuan, T. Y. Zhang, and J. C. Huang, “Performance analysis of multiuser scheduling system based on block diagonalization zero forcing transmission strategy”, *Information Technology Journal*, vol.10, pp. 863-869, 2011.

[9] W. K. Cao, T. X. Zhang, S. W. Zhou, and Q. M. Chen, “Automotive hierarchical control network based on TTCAN and CAN for entire cars and its scheduling strategy”, *Journal of Northeastern University*, vol. 28, pp. 1640-1643, November 2007.

[10] T. Guesmi, S. Hasnaoui, H. Rezig, and O. Korbaa, “Design and implementation of a global scheduling strategy for data distribution over can-based networks”, *International Journal of Computers and Applications*, vol.33, pp. 271-283, 2011.

[11] L. Zhou, J. P. Rodriques, and L. M. Oliveira, “QoE-driven power scheduling in smart grid: Architecture, strategy, and methodology”, *IEEE Communications Magazine*, vol.50, pp. 136-141, 2012.

[12] X. M. Mao. And P. L. Qiu, “Research on energy efficient scheduling strategy in wireless sensor networks”, *Journal on Communications*, vol.29, pp. 56-61, November 2008.

[13] B. Mukherjee, "Performance of a dual-bus fiber optic network operating under a probabilistic scheduling strategy", *Performance Evaluation*, vol.12, pp. 127-139, April 1991.
[14] S. Natti and M. Kezunovic, "A risk-based decision approach for maintenance scheduling strategies for transmission system equipment", in *Proceedings of the 10th International Conference on Probabilistic Methods Applied to Power Systems*, pp. 590-595, 2008
[15] W. Xu and X. D. Dong, "Enhanced multi-mode transmission by user scheduling in MISO broadcast channels with finite-rate feedback", *Wireless Personal Communications*, vol. 65, pp. 103-123, July 2012.
[16] Y. P. Zhang and C. Li, "Research of the scheduling method for fieldbus network real-time information", in *Proceedings of the 6th International Forum on Strategic Technology*, vol. 2, pp. 1125-1128, 2011.
[17] J. Y. Zhang and P., Q. Fan, "Optimal scheduling for network coding: Delay v.s. efficiency", in *Proceedings of IEEE Global Telecommunications Conference*, 2010.
[18] L. X. Peng, L. M. Sheng, G. X. Guang, and D. Z. Yu, "A scheduling strategy by trade-offing between power and delay in broadcast channel", *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, vol. 38, pp. 95-98, July 2010.
[19] H. Wu, Z. Tang, and R. F. Li, "A priority constrained scheduling strategy of multiple workflows for cloud computing", in *Proceedings of 14th International Conference on Advanced Communication Technology: Smart Society Innovation through Mobile*, pp.1086-1089, 2012.
[20] The Advanced Security Acceleration Project (ASAP-SG), "Security profile for advanced metering infrastructure", pp. 9-17, 2010.
[21] Y. Q. Yuan, X. Y. Ge, and L. L. Lu, "A new information security model in smart grid", in *Proceedings of Communications in Computer and Information Science*, vol. 267, pp. 39-44, 2012.
[22] Y. Huang, A. Brocco, P. Kuonen and M. Courant, "Smart GRID: A fully decentralized grid scheduling framework supported by swarm intelligence", in *Proceedings of 7th International Conference on Grid and Cooperative Computing*, pp. 160-168, 2008.
[23] H. C. Han, H. P. Xu, and Z. Q. Yuan, "Research of interactive charging strategy for electrical vehicles in smart grids", in *Proceedings of 2011 International Conference on Electrical Machines and Systems*, 2011.
[24] Z. L. Piao, and D. M. Tan, "Intelligent control strategies and devices of line reactive power", in *Proceedings of 2010 China International Conference on Electricity Distribution*, 2010.

**Weiming Tong** was born in 1964, in Heilongjiang Province, China. He got his PhD title from Harbin Institute of Technology, China, in 2003. He is currently a professor at Harbin Institute of Technology. His research interests include electrical intelligent technology, distribution automation and substation automation, and EMC technology. He has published more than 200 peer reviewed research papers.

**Xianji Jin** was born in 1982, in Liaoning Province, China. He is currently a PHD candidate at Harbin Institute of Technology in China. He obtained his master degree in electrical engineering at the same university in 2007. He obtained his bachelor degree in electrical engineering and automation at the same university in 2005. His research topic is information and communication technology of power system.

**Lei Lu** was born in 1987, in Henan Province, China. He is currently a PHD candidate at Harbin Institute of Technology in China. He obtained his master degree in electrical engineering at the same university in 2009. He obtained his bachelor degree in electrical engineering and automation at the same university in 2007. His research topic is information security of power system.

# Credit Scoring Model Hybridizing Artificial Intelligence with Logistic Regression

Han Lu
School of Economics and Management, Beihang University, Beijing 100191, China
Email: hanluivy@126.com

Han Liyan, Zhao Hongwei
School of Economics and Management, Beihang University, Beijing 100191, China
Email: hanly@buaa.edu.cn, hongwei_zhao@yeah.net

***Abstract*— Today the most commonly used techniques for credit scoring are artificial intelligence and statistics. In this paper, we started a new way to use these two kinds of models. Through logistic regression filters the variables with a high degree of correlation, artificial intelligence models reduce complexity and accelerate convergence, while these models hybridizing logistic regression have better explanations in statistically significance, thus improve the effect of artificial intelligence models. With experiments on German data set, we find an interesting phenomenon defined as 'Dimensional interference' with support vector machine and from cross validation it can be seen that the new method gives a lot of help with credit scoring.**

***Index Terms*—Credit Scoring, Neural Networks, Support Vector Machine, Logistic Regression, Artificial Intelligence**

## I. INTRODECTION

The credit scoring model [1] is developed on the basis of historical data about the performance of previously made loans with some quantitative techniques, such as statistics, mathematical programming, artificial intelligence and data mining.

The most popular methods adopted in credit scoring are logistic regression and its variations. The statistical principle discriminating different groups in a population can be traced back to 1936 in Fisher's publication [2] which used a linear model to calculate the distance between two classes as the judgment. It is known as the Fisher's discrimination. In 1980, Martin [3] first introduced logistic regression method to the bank crisis early warning classification, he analyzed the bankruptcy probability interval distribution, two types of errors and the relationship between the split point, then found that size, capital structure, and performance were key indexes for the judgment, the accuracy rate of overall classification can reach at 96.12%. Wigintion [4] was one of the first researchers to report credit scoring results with logistic regression model. Although the result was not very impressive, the model was simple and can be illustrated easily. Then with the work of Hosmer and Lemeshow [5], it had become the main approach in the practical credit scoring application. Nonetheless, as known, there are quite a few limitations associated with its applications in credit scoring. First of all, it entails intensive data pre-processing effort through interactive variable selection analysis. This usually requires domain expert knowledge and in-depth understanding of the data. In addition, these methods need many conditions for hypothesis, in real world application, assumptions regarding the data may not be held, such as being linear separable and the data should follow certain distributions. Most importantly, based on these algorithms, it is difficult to automate the modeling process and lacks of robustness. When environment or population changes occur, the static models usually fail to adapt and may need to be rebuilt again.

In response to the concern for classification accuracy in retail loans applications, researchers then found the use of neural network classification models. Neural networks are specialized hard-wares or soft-wares that emulate the processing patterns of the biological brain. Desai, Crook, Overstreet [6] compared the classification accuracy of two neural networks models- multilayer perceptrons and modular neural networks with some traditional techniques, such as linear discriminant analysis and logistic regression for credit scoring in the credit union environment. The results indicated that customized neural network offered a very promising avenue if the measure of performance was percentage of bad loans correctly classified. Tam and King [7] compared the artificial neural system approach with a linear classifier, the logistic regression model, KNN model, and ID 3 model to predict bank failures. They concluded that neural networks were more accurate, adaptive, and robust in comparison with other methods. West [8] investigated the credit scoring accuracy of five neural networks models: multilayer perceptron (MLP), mixture of experts (MOE), radial basis function (RBF), learning vector quantization (LVQ), and fuzzy adaptive resonance (FAR). In terms of the results, the difference of performance among these five neural networks models was marginal. However, the author suggested that although the MLP was the most commonly used neural network model, the mixed of experts and RBF should be considered for credit evaluation applications, there still have many other studies reported different conclusions. For example, Yobas, Crook, Ross [9] made a comparison of the predictive ability of linear discriminant analysis, neural networks, genetic algorithms and decision trees in the classification of credit card applicants. The outcome from neural network was found to be marginally less successful than linear discriminant analysis in classifying


doi: 10.4304/jnw.8.1.253-261

the bads and considerably less successful at predicting the goods. Though this method can win others in the accuracy sometimes and has the feature of robustness, it has a fatal flaw — it cannot be readily interpreted. Just because of this, the results made by the neural networks are not very popular in the evaluation of credit risk.

Over the last few years, the application of a new classification technique — the support vector machine (SVM) has been investigated by researchers and generated promising results in credit scoring and financial risk predictions. The support vector machines (SVM) approach is first proposed by Vapnik (1995) [10]. The main idea of SVM is to minimize the upper bound of the generalization error rather than empirical error. After the invention of SVM, some researchers have introduced SVM to the credit scoring problem. Suykens, Gestel, Brabanter, et al. [11] used least squares SVM (LS-SVM) for credit rating of banks and reported the experimental results compared with ordinary least squares (OLS), ordinary logistic regression (OLR) and the multilayer perceptron (MLP). The result showed that the accuracy of LS-SVM classifier was better than the other three methods. Schebesch and Stecking [12, 13] used a kind of standard SVM proposed by Vanik with linear and RBF kernel for dividing credit applicants into subsets of 'typical' and 'critical' patterns which can be used for rejecting applicants. And they concluded these kinds of SVM should be widely used because of their performance. Gestel, Baesens, Garcia, et al. [14] discussed the benchmarking study of 17 different classification techniques on eight different real-life credit datasets. They used SVM and LS-SVM with linear and RBF kernels and adopt a grid search mechanism to tune the hyper parameters in their study. The experimental results indicated that six different methods were the best in terms of classification accuracy among the eight datasets — linear regression, logistic regression, linear programming, classification tree, neural networks and support vector machines. In addition, the experiments showed that the SVM classifiers can yield the overall best performance. Huang, Chen, Jiau [15] demonstrated that the SVM-based model was very competitive to back-propagation neural network (BPN), genetic programming (GP) and decision tree in terms of classification accuracy.But there still are two obvious drawbacks with this method. One is when the variables are not 'meaningful' and 'huge', SVM requires long time to train. Another is just as neural networks, though this method has good robustness and can always get better accuracy; it lacks of explanation capability for their results. That is, the results obtained from SVM classifiers are not intuitive to humans and hard to be illustrated.

This paper focuses on the accuracy of credit scoring and tries the way to find virtual helpful variables in the model in order to reduce the data dimension and at the same time improve interpretability. The structure of the rest paper is as follows: the next section describes the methods proposed in this study and their main principles in detail. Section 3 is about experiment design, including data and variable description, data preprocessing, criterion of effectiveness, evolutionary learning, paramers selection and cross-validation. Experimental studies using the original methods and the methods hybridizing logistic regression are presented in section 4. Finally, section 5 discusses the interesting results and gives some remarks.

## II. Models and Methodology

Let $X = (x_1, x_2, \cdots x_m)^T$ be a set of m random variables which describe the information from a customer's application form and credit reference bureau. The actual value of the variables for a particular applicant k is denoted by $X_k = (x_{1k}, x_{2k}, \cdots x_{mk})^T$ . All samples denoted by $S = \{(X_k, y_k)\}, k = 1, 2, \cdots, N$ , where N is the number of samples, $X_k$ is the attribute vector of the kth customer, and $y_k$ is its corresponding observed result of timely repayment. If the customer is good, $y_k = 1$ , else $y_k = -1$ . Let $I = \{i \mid y_i = 1, i \in N, (x_i, y_i) \in S\}$ is on behalf of good customers, $J = \{i \mid y_i = -1, i \in N, (x_i, y_i) \in S\}$ is on behalf of bad ones.

Credit scoring problem can be described simply as making a classification of good or bad for a certain customer using the attribute characteristics of a certain customer. In order to make a more accurate judgment, a lot of quantitative techniques have been used to develop credit scoring models. Some of typical credit scoring techniques are briefly described below.

### A. Logistic Regression

In the credit scoring problems, the most commonly used technique is logistic regression. Just as linear regression, logistic regression assumes that the sum of the weighted input variables is linearly correlated to the natural log of the odds that the outcome event will happen. It can be described as (1):

$$\log(p/1-p) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e = \beta^T X_k + e$$

Where $\beta = (\beta_1, \beta_2, \ldots\ldots, \beta_k)$ is the vector of the coefficients of the model, the maximum likelihood method can be applied to compute the estimate of $\beta_i$ {i=1, 2, ……, k}. We refer to p/ (1-p) as odds-ratio and assume the regression model in (1) is obtained; the estimated probability of no default is as follows:

$$p = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \tag{2}$$

Logistic regression can overcome the flaw of linear regression, which is that the right side of the model could take any value from -∞ to +∞ but the left side can only take values between 0 and 1. And just like linear regression, it has good interpretations in statistical sense. Nevertheless the method is intrinsically linear, which cannot well deal with non-linear effects in practice, now researchers usually use the variable combinations and trial-and-error process to deal with non-linear effects. In addition, the method is sensitive to redundancy or collinearity in the input variables, which can give bad estimates of the coefficients. Furthermore, to give the results of better explanations also calls for the residual error to obey the logic normal distribution.

### B. Neural Networks

Neural networks approach is an interconnected assembly of simple processing elements, i.e. nodes, whose functionality is loosely based on the animal neuron. The processing ability of the neural network is stored in the inter-unit connection strength, or weight, obtained by a process of adaptation to, or learning from a set of training patterns (Gurney 1997) [23]. Neural networks are a class of powerful data modeling tools and have been widely used in practical problems such as classification, evaluation, and forecasting problem. The two of most popular neural networks for classification are Multilayer Perceptron (MLP) and Radial Basis Function network (RBF).

Common framework for both MLP and RBF can be constructed below, which is typically composed of an input layer, one or more hidden layers and an output layer, each consists of several neurons. An illustrative structure with one hidden layer is shown in Figure 1: an input layer with k neurons, a neuron for every input variable, a hidden layer with m neurons and an output layer with one neuron. There is no connection between the input and output layers either. The activation function of the j-th neuron in the hidden layer is given by $B_j(X, w_j)$ which can be different functions in MLP and RBF. The activation function of the output neuron is given by:

$$f(X,\theta) = \beta_0 + \sum_{j=1}^{m} \beta_j B_j(X, w_j) \tag{3}$$

where $\beta_j$ is the weight of the connection between the hidden neuron j and the output neuron and $\theta$ is the vector of parameters of the neural net. The activation function of all hidden and output neurons is the identity function.

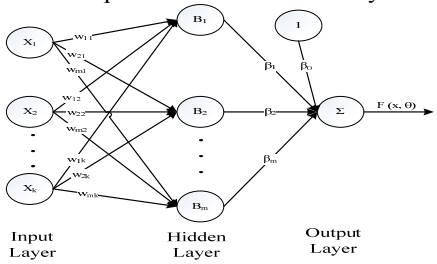



Figure 1. Structure of common framework for both MLP and RBF

The difference between MLP and RBF is related to the activation function considered in the hidden layer. In this way, MLP considers (4) as activation function.

$$B(X, w_j) = \prod_{i=1}^{m} x_i^{w_{ji}} \tag{4}$$

Where $w_{ji}$ is the weight of the connection between input neuron i and hidden neuron j and $w_j = (w_{j1}, w_{j2}, \ldots\ldots w_{jk})$ is the weight vector. In RBF, the hidden activation functions are radial basis function, the mostly used is the Gause function (5):

$$B(X, w_j) = \exp\left(-\frac{1}{r_j^2}\left\|x - C_j\right\|^2\right) \tag{5}$$

Where $w_j = (c_j, r_j), c_j = (c_{j1}, c_{j2}, \ldots\ldots c_{jk})$ is the centre or average of the j-th Gaussian RBF transformation $r_j$ is the corresponding radius or standard deviation.

### C. Support Vector Machine

The main idea of support vector machine is to minimize the upper bound of the generalization error not the empirical error. Usually it maps the input vectors into high-dimensional feature space through some nonlinear mapping. In this space, an optimal separating hyperplane which is one that separates the data with the maximal margin is constructed by solving a constrained quadratic optimization problem whose solution has an expansion in terms of a subset of training patterns that lie closet to the boundary.

Suppose $z_i = \varphi(x_i)$ where $\varphi(\cdot)$ is a nonlinear function that can map the input space into a higher dimensional feature space (z-space). If the training set is liear separable in the feature space, after normalization, the classifier should be constructed as follow:

$$\begin{cases} w \cdot z_i + b \geq 1 & \forall i \in I \\ w \cdot z_i + b \leq -1 & \forall i \in J \end{cases} \tag{6}$$

To classify all the samples correctly, we need $y_i(wx_i + b) - 1 \geq 0$, if there exists a (w, b) pair for which the constraints above are satisfied, the distance between the two boundary hyperplanes is 2/||w|| so the maximal distance is the minimal of $\|w\|^2$. The main step in SVM is to construct the optimal hyperplane. Based on the description above, this can be found by solving the following quadratic programming problem:

$$\begin{cases} \min J(w,b) = \|w\|^2 / 2 \\ s.t. \quad w \cdot z_i + b \geq 1 \\ \qquad w \cdot z_i + b \leq -1 \end{cases} \tag{7}$$

In many practical situation, the training samples cannot be linear separable in z-space. There is a need to use soft margin and C penalty parameters, this is formulized as the following constraint optimization problem:

$$\begin{cases} \min J(w,b,\xi_k) = \frac{1}{2}\|w\|^2 + C\sum_k \xi_k \\ w \cdot z_i + b \geq 1, \forall i \in I \\ w \cdot z_i + b \leq -1, \forall i \in J \\ \xi_k \geq 0, \forall k \in I \cup J \end{cases} \tag{8}$$

where C is the corresponding penalty parameters indicating a tradeoff between large margin and a small number of margin failures. The solution to this optimization problem can be given by the saddle point of the Lagrange function with Lagrange multipliers $\alpha_i$, and then the problem can be transformed into its dual form:

$$\begin{cases} \max J(\alpha) = -\frac{1}{2}\sum_k \sum_l \alpha_k \alpha_l y_k y_l z_k \cdot z_l + \sum_k \alpha_k \\ 0 \le \alpha_k \le C, \forall k \\ \sum_k \alpha_k y_k = 0, y_k = 1, k \in I, and, y_k = -1, k \in J \end{cases} \quad (9)$$

And then with the optimal solution $\alpha_i$, we can get the support vectors that are these ones with $\alpha_i \neq 0$, the optimal hyper plane's coefficient is $W^* = \sum_{i=1}^{k} \alpha_i^* y_i x_i$ and with $\alpha_i[y_i(wx_i + b) - 1] = 0$, we can easily construct the optimal hyperplane.

The most common method for mapping function $Z_i = \phi(x_i)$ is to use the kernel function $K(x_k, x_l)$, which is the inner product in the feature space, to perform the mapping. One main merit of SVM is that by a kernel function it can make the training data linearly separable in the z-space as possible as it can do. Usually, the commonly used kernel functions are listed below.

(1) Linear kernel function: $K(x_k, x) = x_k^T x$

(2) Polynomial function: $K(x_k, x) = (x_k^T x + 1)^d$

(3) Gaussian function:

$K(x_k, x) = \exp(-\|x_k - x\|^2 / \sigma^2)$

## III. Experiments Design

This section is structured in six subsections. Firstly, we have a brief description with the dataset used in the experiments. And then we discuss the process of data preprocessing for modeling. In the third part, we define the evaluation criteria. The methods for training the networks and SVM are introduced in the next two subsections. Finally, to test the robustness of models, we design two cross-validation methodologies and they are illustrated in the sixth subsection.

### A. Dataset Description

The credit dataset used in these experiments is German credit dataset, which is provided by Professor Dr. Hans Hofmann of the University of Hamburg and is obtained from UCI Machine Learning Repository (http: //www.ics.uci.edu/~mlearn/databases/statlog/german/). The total number of instances is 1000 including 700 creditworthy cases and 300 default cases. There is no missing data.

For each applicant, 20 kinds of attribute are available; the variable names of these attributes used in the models are listed below with short names in brackets. There are 13 categorical attributes including status of existing checking account (checking), credit history (history), purpose for the credit (purpose), saving account (savings), present employment since (employed), personal status and sex (marital), other debtors/guarantors (coapp), property style (property), other installment plans (other), housing situation (housing), job status (job), telephone status (telephone), foreign worker or not (foreign) and 7 numerical attributes including duration in month (duration), credit amount (amount), installment rate in percentage of disposable income (installp), present residence since (resident), age in years (age), number of existing credits at this bank (existcr), number of people being liable to provide maintenance for (depends).Table 1 and table 2 below show the basic statistics and information of these attributes.

TABLE I.
STATISTICS OF CATEGORICAL ATTRIBUTES

| Variable | Level | Missing | Mode | Mode% |
|---|---|---|---|---|
| foreign | 2 | 0 | 1 | 96.3 |
| coapp | 3 | 0 | 1 | 90.7 |
| other | 3 | 0 | 3 | 81.4 |
| housing | 3 | 0 | 2 | 71.3 |
| job | 4 | 0 | 3 | 63 |
| savings | 5 | 0 | 1 | 60.3 |
| telephon | 2 | 0 | 1 | 59.6 |
| marital | 4 | 0 | 3 | 54.8 |
| history | 5 | 0 | 2 | 53 |
| checking | 4 | 0 | 4 | 39.4 |
| employed | 5 | 0 | 3 | 33.9 |
| property | 4 | 0 | 3 | 33.2 |
| purpose | 10 | 0 | 3 | 28 |

TABLE II.
STATISTICS OF NUMERICAL ATTRIBUTES

| Variable | Means | STD | Missing | Min | Median | Max |
|---|---|---|---|---|---|---|
| age | 35.55 | 11.38 | 0 | 19 | 33 | 75 |
| amount | 3271.26 | 2822.74 | 0 | 250 | 2319 | 18424 |
| depends | 1.16 | 0.36 | 0 | 1 | 1 | 2 |
| duration | 20.9 | 12.06 | 0 | 4 | 18 | 72 |
| existcr | 1.41 | 0.58 | 0 | 1 | 1 | 4 |
| installp | 2.97 | 1.12 | 0 | 1 | 3 | 4 |
| resident | 2.85 | 1.1 | 0 | 1 | 3 | 4 |

From these statistics and information of these attributes, we can see some attributes have relatively concentrated distribution, for example foreign and coapp, the modes get more than 90%. With the numerical attributes, amount is more 'big' than others in the amount level. So there may need some preprocessing before modeling.

### B. Data Preprocessing

To deal with the categorical attributes, the typical method is to code them with their levels as dummy variables. All of the categorical attributes are used with their levels. For example, foreign has two levels, so there will be two variables in the model — one is foreigh 0, and another is foreign 1. Other variables are the same. It can be seen that the way to deal with categorical attributes will course the number of total variables much increasing.

To deal with the numerical attributes, there is no need to do normalization with the logistic regression model. Though the variable coefficient estimates $\beta_i$ vary because of the unit, the correlation coefficient estimate $R^2$ and the model's results are not influenced, therefore it makes no effect on the choice of variables. Since there is no proof about the necessity of normalization with networks and support vector machines, we cannot give a logistic reason for normalization, but in the practice, we always normalize the numerical attributes in the models so that the results cannot be affected by the unit. In order to better demonstrate we also try the experiments with dimensionless numerical attributes and these ones which are not normalized respectively to compare, the results will be illustrated in the fifth section. The method we use

for normalization is $\frac{x_i - mean}{std}$, this method can transform any distribution of the variable into standard normal distribution, which maybe well meets the variable distribution of the model's requires.

### C. The Evaluation Criteria

Let the number of creditworthy cases classified as good be GG and classified as bad with GB, denote the number of default cases classified as good with BG and as bad with BB. Then the evaluation criteria measure the accuracy of the classification, which is defined as follows:

$$\text{Good credit accuracy (GCA)} = \frac{GG}{GG+GB} \times 100\%$$

$$\text{Bad credit accuracy (BCA)} = \frac{BB}{BG+BB} \times 100\% \quad (10)$$

$$\text{Overall accuracy (OA)} = \frac{GG+BB}{GG+GB+BG+BB} \times 100\%$$

Defined by these three indicators, one can see GCA is the specificity, which determines the ability to identify good clients; BCA is the sensitivity, for the model it shows the ability to identify bad customers. At the same time, OA gives the total efficiency of the model. It reflects prediction accuracy of the model and can compare with others.

In our study, Type I error occurs when a bad credit is classified as good credit, which equals 1-BCA. And Type II error occurs when a good credit is classified as a bad credit, which equals 1-GCA. For credit scoring, Type I error is more critical than Type II error. Note that all these measures with Type I error and Type II error are mostly obtained using a 0.5 probability threshold for the classification. However, the use of arbitrary cut-off probabilities makes the computed error rates difficult to interpret [16] and the use of a relevant pay-off function and prior probabilities to determine the optional model could lead to some kinds of bias on the results.

ROC curve can come over this problem. Receiver operating characteristics (ROC) graph is useful for organizing classifiers and visualizing their performance. ROC graph is two-dimensional graphs in which BCA (true positives) rate is plotted on the Y axis and 1-GCA (false positives) rate is plotted on the X axis. And ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positive). As the production process of ROC, it can be seen that the diagonal line y=x represents the strategy of randomly guessing a class. To compare classifiers we want to reduce ROC performance into a single scalar value representing expected performance. A common method is to calculate the area under the ROC curve, abbreviated as AUC. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. Bradley [17] has compared popular machine learning algorithms using AUC, and found that AUC is a very effective way to measure the results of models, and it can exhibit several desirable properties compared to accuracy. Moreover, in the study of Fawcett [18], AUC measure is more sensitive to the errors on the positive class, since it has an important statistical meaning: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance, considering also all possible thresholds. It is very meaningful because in the credit scoring problem, it will bring more cost if one judges a bad credit as a worthy one.

Because of these reasons, the AUC has been selected as the main evaluation criteria in this paper which can be seen in the ROC pictures, and the BCA and GCA are also be used for checking the accuracy of these models as a reference.

### D. Training the Networks

Though there are many kinds of neural networks, in our study we only experiment on the most widely used ones: MLP and RBF, which are introduced briefly above.

Firstly, to train the network adequately, the training sample should be a good representative of the population under study, since the neural network models are good at generalization, but cannot extrapolate well. Thus, the training data should cover the entire expected input data space. In accordance with these guidelines, the network was trained with a sample of 750 observations. The training set is an unbiased sample with 70% vectors from the good loan class and 30% from the bad loan class. Further, to ensure that the network is not trained with vectors from one class or one single credit union, the observations were intermingled randomly.

Moreover, for a network to be efficiently trained, besides a good training sample, sufficient number of hidden neurons is also essential. Although many rules of thumb can be applied to select the size of the hidden layer, for example, number of hidden neurons = training facts× error tolerance or number of hidden neurons = 5-10% of training facts, in most of these cases trial and error is the best guide. If the network fails to converge to a solution, one might try fewer hidden nodes, and settle on a size based on the performance of the system.

In addition, to get an optimal number of hidden neurons, neural network should also be trained with optimal number of training cycles. If a network is undertrained, results will be underfitting, on the other hand if a network is overtrained, results will be overfitting. Thus, we experimented with neural network models with relative large number of hidden neurons- 20 neurons in the hidden layer, all the input variables with their levels as input neurons and one neuron in the output layer. By varying the number of hidden neurons, the network was trained with the training sample to determine the optimal design.

### E. Select Parameters for SVM

In order to compare with the networks, we choose the linear kernel function with support vector machine (LSVM) for test. Though SVM is a powerful learning method for classification problem, its performance is not only sensitive to the algorithm that solves the quadratic programming problem, but also to the parameters setting in the SVM.

In the process of using SVM, the first issue is how to search the best parameters of SVM for a specified problem. Most widely used learning method for parameters selection is grid search (GS). In the grid search method, each point in the grid is defined by the grid range

[ (C min, sigma min), (C max, sigma max) ] and a unit grid size ($\Delta$_C, $\Delta$_sigma) is evaluated by the objective function F. The point with the smallest F value corresponds to the optimal parameters. Just as neural network, we used 750 samples and gird search to select the parameter C of LSVM.

### *F. Cross-validation*

As the best model is tailored to fit one sub-sample, the model often estimates the true error rate over optimistically. Therefore, to get a true estimate of the error rate, we applied two kinds of cross-validation methodologies which was suggested by Zhang, Hu, Patuwo, et al. [19]: Firstly, to test the robustness of the models, one should apply a simple validation technique by dividing the data set into a training sample and a validating sample with a small scale which valuates the predictive effectiveness of the fitted model. Secondly, to study the overall predictive capability of the classification models for unknown population, one should use the whole data set as large test set if the data set for unknown population is not available.

To implement the first cross-validation methodology, we divide the data sample into four mutually exclusive equal sub-samples. Each sub-sample has the same rate for the bad customers and the good ones. We train the neural networks and the SVM with three sub-samples, and validate the models with the fourth remaining sub-sample. Therefore, out-of-train prediction of validation gives us a relative true classification rate of all the observations in the data set with its averages. Secondly, to test the overall predictive capability of the unknown population comprehensively, we use the entire data sample. By using the entire data set as the test sample, we can reduce the sampling variation in the test design. Finally, we use pair-t test to test the difference between the means of the "original" method and the method hybridizing logistic regression in section 4.

## IV. EXPERIMENTS RESULTS

The experimental results presented in this paper are structured in three subsections. The first subsection describes the performance of original method. The process of hybridizing logistic regression is shown in the second subsection. Finally, the comparison of these methods is shown in the third subsection.

### *A. Experiments with the Original Methods*

In this section, we focus on the effects of original methods - logistic regression (LG), MLP, RBF, and LSVM.

Table 3 lists the results of these models with normalization and the results without normalization; we use the whole data set (1000 observations and 20 variables) for this experiment.

From the results, we can see the models with normalization have the same results as the models without normalization in LG, MLP and RBF. But there is a very interesting founding with LSVM. Without normalization the results using LSVM is very bad and the main error occurs in the prediction of the good. We also do some similar experiments to test the results and these results are the same. SVM cannot do right job without normalization in the good prediction. We believe that this phenomenon may be caused by kernel function, but no reference has been introduced for this, so we define it as 'Dimensional Interference'. In the following experiments, we only use models with normalization.

TABLE III.
CLASSIFICATION OF DIFFERENT MODELS WITH/WITHOUT NORMALIZATION

| | | With normalization | | | |
|---|---|---|---|---|---|
| goal | result | LG | MLP | RBF | LSVM |
| bad | bad | 160 | 184 | 151 | 161 |
| bad | good | 140 | 116 | 149 | 139 |
| good | bad | 76 | 63 | 86 | 77 |
| good | good | 624 | 637 | 614 | 623 |
| | OA | 78.40% | 82.10% | 76.50% | 78.40% |
| | | Without normalization | | | |
| goal | result | LG | MLP | RBF | LSVM |
| bad | bad | 160 | 116 | 149 | 176 |
| bad | good | 140 | 184 | 151 | 124 |
| good | bad | 76 | 519 | 467 | 483 |
| good | good | 624 | 181 | 233 | 217 |
| | OA | 78.40% | 29.70% | 38.20% | 39.30% |

The cross-validation results of these original models with normalization are summarized in table 4. The train data set result is the average of four times experiments with 750 observations and 20 variables, and the validation data set result is also the average with 250 observations, they both use equal proportion of the bad ones. And the entire data set is used as test data set.

From the table 4, we can see all of these models do not get satisfactory accuracy with the prediction for the bad ones, though OA is really high. Only the LSVM model is a little better and with the expense of good customer forecast. Though MLP and RBF both do quite well in the prediction of good customers, but they still cannot beat LG.

TABLE IV.
CROSS-VALIDATION RESULTS OF THESE MODELS

| | goal | B | B | G | G | BCA | GCA | OA |
|---|---|---|---|---|---|---|---|---|
| | result | B | G | B | G | | | |
| LG | train | 113 | 112 | 56 | 469 | 50% | 89% | 78% |
| | valiadation | 38 | 37 | 18 | 157 | 51% | 90% | 78% |
| | test | 150 | 150 | 71 | 629 | 50% | 90% | 78% |
| MLP | train | 116 | 109 | 40 | 485 | 52% | 92% | 80% |
| | valiadation | 38 | 37 | 15 | 160 | 51% | 91% | 79% |
| | test | 186 | 114 | 56 | 644 | 62% | 92% | 83% |
| RBF | train | 113 | 112 | 66 | 459 | 50% | 87% | 76% |
| | valiadation | 38 | 37 | 20 | 155 | 51% | 89% | 77% |
| | test | 153 | 147 | 86 | 614 | 51% | 88% | 77% |
| L-SVM | train | 119 | 106 | 63 | 462 | 53% | 88% | 77% |
| | valiadation | 44 | 31 | 22 | 153 | 59% | 87% | 79% |
| | test | 163 | 137 | 85 | 615 | 54% | 88% | 78% |

### *B. Hybridizing Logistic Regression*

Within MLP and RBF procedures, they both give the importance of the variables. For MLP, the importance order is

$$duration \succ checking \succ amount \succ purpose \succ history \succ employed$$

and which are that variables have the importance over 50%. And in RBF, the order is very different, and this is $duration \succ age \succ amount \succ checking \succ property$ Further, the importance has no exact interpretation in statistics. Unluckily, there is no way to find out which variables are good enough ones only through neural network. The similarly, SVM works on the support vectors which are the observations with $\alpha_k$ is not zero. So it cannot give the importance of variables. At the same time, it can be seen that if the input variables are too much, it also cannot do well with the kernel function transformation to construct the separable hyperplane.

So, we turn to find help with statistics methods, and try to use logistic regression to deal with this problem. Just because logistic regression has good statistical interpretation with the variables, and may give the variable selection problem a reasonable pass. Logistic regression can choose helpful variables for three reasons. Firstly, logistic regression selects the most relevant variables with the target variables into the model. Secondly, in statistics it can use the Wald test to decide whether adding variables improves the unconstrained model. The last one is that by doing variables selection with logistic regression can exclude the multiple correlations among variables. The steps of selecting variables using the forward logistic regression are summarized in table 5.

TABLE V.
THE STEPS OF VARIABLES SELECTION

| | improvement | | | model | | | OA | variables |
|---|---|---|---|---|---|---|---|---|
| | $x^2$ | df | Sig. | $x^2$ | df | Sig. | | |
| 1 | 131.3 | 3 | 0 | 131.3 | 3 | 0 | 70.0% | checking |
| 2 | 38.5 | 1 | 0 | 169.8 | 4 | 0 | 73.4% | duration |
| 3 | 29.3 | 4 | 0 | 199.1 | 8 | 0 | 74.8% | history |
| 4 | 33.5 | 9 | 0 | 232.7 | 17 | 0 | 76.2% | purpose |
| 5 | 18.8 | 4 | 0 | 251.4 | 21 | 0 | 76.2% | savings |
| 6 | 11.1 | 2 | 0 | 262.5 | 23 | 0 | 76.6% | coapp |
| 7 | 6.5 | 1 | 0 | 269.0 | 24 | 0 | 77.4% | installp |
| 8 | 7.0 | 1 | 0.01 | 286.9 | 28 | 0 | 77.5% | amount |
| 9 | 8.6 | 2 | 0.01 | 295.5 | 30 | 0 | 78.0% | other |

Through comparison with the progress of selecting variables in neural networks, advantages can be seen more clearly. For example, "employed" is a very important variable in the MLP, but it is not the same useful in the RBF. By checking in logistic regression with the p-value (shown in the table 5), we know this variable is not important enough and may be excluded out of the models. Another good example is the variable "job". We can see it is an important variable in two networks, but with the p-value we know it has little meaning for target.

### C. Comparison of These Methods

Finally, using the important variables selected by logistic regression as inputs, we experiment with MLP, RBF and LSVM again and the way of cross-validation is just as the original methods. The result is summarized in table 6.

Compare with table 4, we can see MLP hybridizing logistic regression can do much better in prediction of the bad loans, so OA has a little improvement. It is the same with RBF, OA has a little lift mainly because of the lift in the accuracy of BCA. But it is opposite with LSVM, though OA has also got a quite lift, this lift is mainly because of the accuracy for good loans, and at the same time it reduces the degree of differentiation of bad customers in general as a trade-off.

TABLE VI.
CROSS-VALIDATION RESULTS OF HYBRIDIZING MODELS

| | goal | B | B | G | G | BCA | GCA | OA |
|---|---|---|---|---|---|---|---|---|
| | result | B | G | B | G | | | |
| MLP | train | 124 | 101 | 49 | 476 | 55% | 91% | 80% |
| | valiadation | 42 | 33 | 20 | 155 | 56% | 89% | 79% |
| | test | 166 | 134 | 69 | 631 | 55% | 90% | 80% |
| RBF | train | 120 | 105 | 86 | 439 | 53% | 84% | 75% |
| | valiadation | 42 | 33 | 20 | 155 | 56% | 89% | 79% |
| | test | 166 | 134 | 63 | 637 | 55% | 91% | 80% |
| L-SVM | train | 113 | 112 | 56 | 469 | 50% | 89% | 78% |
| | valiadation | 40 | 35 | 15 | 160 | 53% | 91% | 80% |
| | test | 157 | 143 | 71 | 629 | 52% | 90% | 79% |

To test whether the means of accuracy between the hybridizing methods and the original ones are different or not, we use the pair-t test. The result is shown in table 7 below.

TABLE VII.
PAIRWISE COMPARISON BETWEEN ORIGINAL MODELS AND HYBRIDIZING MODELS

| | | Bad Loan | | Good Loan | | Overall | |
|---|---|---|---|---|---|---|---|
| | | O | H | O | H | O | H |
| mean | | 0.52 | 0.54 | 0.89 | 0.89 | 0.78 | 0.79 |
| t-value | pooled | -2.34 | | -0.43 | | -1.47 | |
| | satterthwaite | -2.34 | | -0.43 | | -1.47 | |
| | cochran | -2.34 | | -0.43 | | -1.47 | |
| pr>\|t\| | pooled | 0.0325 | | 0.6733 | | 0.1598 | |
| | satterthwaite | 0.0349 | | 0.6734 | | 0.1604 | |
| | cochran | 0.0474 | | 0.6789 | | 0.1786 | |

As show in table 7, on an average, the overall performance of hybridizing models for prediction with bad loan is better than the original models and the difference is statistically significant at 5% rejection level. However, the means of GCA and OA are not statistically significant. Just because of this, although it is difficult to say that the hybridizing model is significantly improved, the predictive results in bad customers have been greatly enhanced, which is the most important in retail credit risk, after all Type I error is more serious with credit evaluation.

This lift in BCA mainly dues to reduction of redundant variables and getting the major character which makes the model increasing the predictive power; while because of reducing variables, the amount of information about customers is also reduced, so it leads some difficulty to upgrade the good ones. So there must be a trade-off between a lift in BCA and a decline in GCA that the business man must face with. But after all with less variables model's complexity and convergence speed have been greatly improved. In our experiments, MLP and RBF have been sped up as twice as before. Further more, with

hybridizing logistic regression, artificial intelligence models will have statistical significance in some degrees which can help interpretations.

To further compare the accuracy of model with any distribution, we give these ROC graphs with original models and hybridizing models in figure 2-4 below.

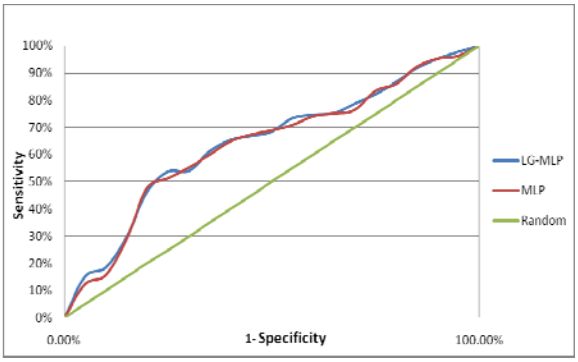



Figure 2. The ROC graph of MLP and hybridizeing MLP

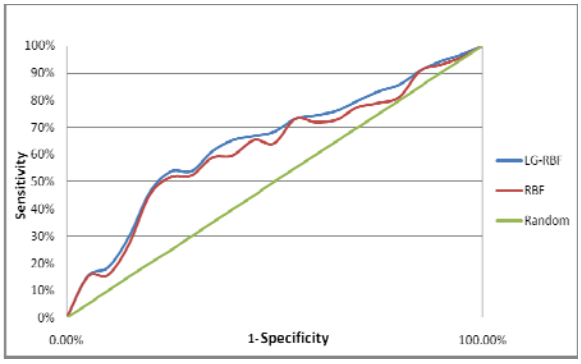



Figure 3. The ROC graph of RBF and hybridizing RBF

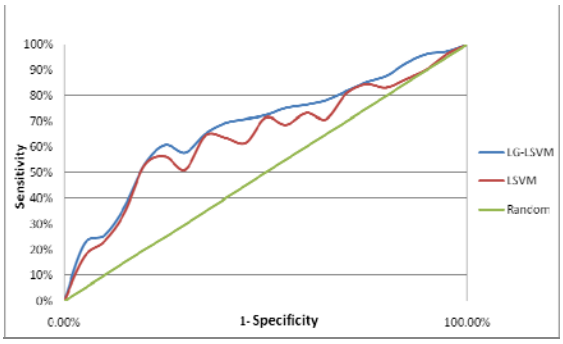



Figure 4. The ROC graph of LSVM and hybridizing LSVM

With these graphs of ROC, we can see no model can be completely superior to other models in any case. The AUGs between the MLP models are almost the same, and a little bigger in RBF and LSVM with hybridizing logistic regression models than the original models. So in this way it can be said that the hybridizing models is more effective with credit scoring problem.

## V. Conclusions

Most artificial intelligence techniques are black-box methods. They are lack of rules for selecting good inputs, and unfortunately they all face with the trouble "garbage in garbage out". Thus, they usually suffer from dimension curse. Focus on this, we introduce hybridizing logistic regression with artificial intelligence techniques, which has better interpretability for the input variables, reduces the dimension and speeds up the rapid of convergence.

There is an interesting discovery in our experiments. The SVM cannot do well with the numerical variables which are not be normalized. We define this phenomenon as 'Dimensional interference'. That may be caused by the distribution of variable and the kernel function cannot effectively transform it to the high dimension. So in the high dimension space, it still cannot be separated by line. But there are few references which have discussed about this. It still needs more theoretical study.

Based on experiments, there is an interesting founding: that is there must be a tradeoff between a lift in BCA and a decline in GCA. It is understandable because of the reduction of the variables, if a model has the power of better distinguish core features, which may help to judge the bad ones, it will lose some information and lower the specificity for a degree, which decreases the accuracy for some good cases. Fortunately, GCAs of all the models are quite impressive, so a small decline is acceptable.

Finally, to sum up, this study provides a new way to do credit scoring, which is quite different from these methods based on the techniques improvement. And in this work, we test a number of statistical approaches to better solve and evaluate. Moreover, in our opinion, for other applications such as pattern recognition, the method of selecting useful variables hybridizing with statistical methods can also be taken in, though its results maybe need further discussion, we still believe it will give a lot of help beyond thought.

## Acknowledgment

R.B.G. thanks Project supported by the National Natural Science Foundation of China (Grant No.70831001 and No.70821061).

## References

[1] L.C. Thomas, D.B. Edelman, J.N. Crook, "Credit Scoring and its Applications", Philadelphia: Society of Industrial and Applied Mathematics: Philadelphia, 2002.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 1936, vol. 7, pp. 179-188.

[3] D. Martin, "Early warning of bank failure: A logit regression approach", Journal of Banking & Finance, 1977, vol. 1 (3), pp. 249-276.

[4] J. C. Wiginton, "A note on the camarison of logit and discriminant models of customer credit behavior", Journal of Financial and Quantitative Analysis, 1980, vol. 15, pp. 757-770.

[5] D. W. Hosmer, S. Lemeshow, "Applied Logistic Regression", New York: Wiley & Sons, 1989.

[6] V. S. Desai, J. N. Crook, G.A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment",European Journal of Operational Research, 1996, vol. 95 (1), pp. 24-37.

[7] K. Y. Tam, M. King, "Managerial applications of neural networks: the case of bank failure predictions", Management Science, 2001, vol. 38 (7), pp. 926-947.

[8] D. West, "Neural network credit scoring models",Computers & Operations Research, 2000, vol. 27, pp. 1131-1152.

[9] M. B. Yobas, J. N. Crook, P. Ross, "Credit scoring using neural and evolutionary techniques", IMA Journal of

Mathematics Applied in Business and Industry, 2000, vol. 11, pp. 111-125.

[10] C. Cortes, V. Vapnik, "Support vector networks", Machine Learning, 1995, vol. 20, pp. 273-297

[11] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, et al., "Least squares Support vector machines", Singapore: World Scientific, 2002.

[12] K. Schebesch, R. Stecking, "Support vector machines for classifying and describing credit applicants: detecting typical and critical regions", Jounal of Operational Research Society, 2005, vol. 56 (9), pp. 1082-1088.

[13] K. Schebesch, R. Stecking, "Support vector machines for credit scoring: Extension to non standard cases", In: Proceeding of Innovations in Classification, Data Science, and Information Systems, 2005, pp. 498-505.

[14] T. V. Gestel, B. Baesens, J. Garcia, et al., "A support vector machine approach to credit scoring", Banke en Financiewezen, 2003, vol. 2, pp. 73-82.

[15] Y. M. Huang, C. M. Chen, H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem", Nonlinear Analysis: Real World Applications, 2006, vol. 7 (4), pp. 720-747.

[16] K. G. Palepu, "Predicting takeover targets: a methodological and empirical analysis", Journal of Accounting and Economics, 1986, vol. 8 (1), pp. 3-35.

[17] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", Pattern Recognition, 1997, vol. 30, pp. 1145-1159.

[18] T. Fawcett, "An introduce to ROC analysis", Pattern Recognition, 2006, vol. 27, pp. 861-874.

[19] G. Zhang, M. Hu, B. Patuwo, et al., "Atificial neural networks in bankruptcy prediction: general framework and cross-validation analysis", European Journal of Operations Research, 1999, vol. 116 (1), pp. 16-32

# 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)*

*(Contents Continued from Back Cover)*



REGULAR PAPERS