Karan Singh
Amit K. Awasthi (Eds.)

LNICST 115

Quality, Reliability, Security and Robustness in Heterogeneous Networks

9th International Conference, QShine 2013
Greader Noida, India, January 2013
Revised Selected Papers

ICST

Springer

Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 115

Karan Singh Amit K. Awasthi (Eds.)

Quality, Reliability, Security and Robustness in Heterogeneous Networks

9th International Conference, QShine 2013
Greader Noida, India, January 11-12, 2013
Revised Selected Papers

Springer

Volume Editors

Karan Singh
Amit K. Awasthi
Gautam Buddha University
Greater Noida 201310, India
E-mail: {karan; amitkawasthi}@gbu.ac.in

ISSN 1867-8211 e-ISSN 1867-822X
ISBN 978-3-642-37948-2 e-ISBN 978-3-642-37949-9
DOI 10.1007/978-3-642-37949-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013937636

CR Subject Classification (1998): C.2, K.6.5, C.4, F.2, I.4, I.6

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

Printed on acid-free paper

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

Preface

This volume of LNICST is a collection of the papers presented at the 9th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE 2013). This technical event took place in the National Capital Region (NCR) of India, during January 11–12, 2013.

The main objective of QSHINE 2013 was to emphasize the recent technological developments in broadband high-speed networks, peer-to-peer networks, wireless and mobile networks, which have led to new challenging problems, such as providing QoS support to the emerging multimedia applications across both wired and wireless networks. It served as a forum for researchers from electronics engineering, computer science and engineering, and applied mathematics.

The proceedings include 87 papers covering a broad range of vital issues in areas related to heterogeneous networking for quality, reliability, security and robustness. These were presented in 14 technical sessions.

An invited session on the "Heterogeneous Network" included 24 papers (selected from 49). The technical sessions on "Reliability" included 9 papers (selected from 15). The following 35 papers (selected out of 61 submissions) were presented in regular sessions on "Security." The "Robustness" sessions enriched the conference with 18 papers (selected from 35). The remaining two papers (selected out of 9 submissions) are dedicated to QoS. For each paper, there were at least two independent reviews, most of which were offered by members of the TPC.

There were 14 keynote speeches at the conference by the following eminent academicians: Magda El Zark, University of California, USA; Anura P. Jayasumana, Colorado State University, USA; D.K. Lobiyal, JNU, New Delhi, India; M. Sufyan Beg, Jamia Millia Islamia University, New Delhi, India; Rama Shankar Yadav, MNIT, India; M. Salim Beg, AMU, India; Subhas Mukhopadhyay, Massey University, New Zealand; Shang-Kuo Yang, National Chin Yi University of Technology, Taiwan; Bharat Bhargava, Purdue University, USA; Raj Kamal, Devi Ahilya University, India; Venkataram Pallapa, Indian Institute of Science, India; Ruay-Shiung Chang, National Dong Hwa University, Taiwan; Pascal Lorenz, University of Haute Alsace, France.

We are very grateful for the participation of these speakers in making this occasion a memorable event.

The success of this conference along with this proceedings volume is largely a result of the devoted work of the 66 TPC members whom we thank. Special thanks are due to the conference Steering Committee and Organizing Committee for their help that made our job much easier and enjoyable. Warm thanks go to Justina Senkus, Dina Shakirova, Elisa Medini, Erica Polini, Natasa Milosevic and Austin. We wish to thank our technical sponsors Create-net, IEEE. The organizing team extends sincere thanks to Shri S.R. Lakha, Vice-Chancellor of Gautam Buddha University for the support in hosting the event. Last but not least, we wish to thank EAI-ICST and Gautam Buddha University.

General Chairs, QSHINE

Organization

Qshine-2013 was organized by Gautam Buddha University, Greater Noida, Delhi-NCR, India, in association with the European Association of Innovation (EAI) and The Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST).

Program Commitee

Steering Committee

Xi Zhang	Texas A&M University, USA
Sherman Shen	University of Waterloo, Waterloo, Canada

General Chairs

Karan Singh	Gautam Buddha University, Greater Noida, India
Amit K Awasthi	G.B.U., Greater Noida, India
Rajesh Mishra	G.B.U., Greater Noida, India

Technical Program Committee Chairs

Sandeep Sharma	G.B.U., Greater Noida, India
Mohd. Ashraf Saifi	G.B.U., Greater Noida, India
Navaid Zafar Rizvi	G.B.U., Greater Noida, India

Organizing Chairs

M.A. Ansari	G.B.U., Greater Noida, India
Visdushi Sharma	G.B.U., Greater Noida, India
Neeta Singh	G.B.U., Greater Noida, India
Shabana Urooj	G.B.U., Greater Noida, India
Navras Jaat Aafridi	G.B.U., Greater Noida, India
Bhavnesh Kumar	G.B.U., Greater Noida, India
S.H. Abbas Mehdi	STPI Ministry of IT, Govt. of India, Noida, India
Manmohan Singh	G.B.U., Greater Noida, India
Mayank Singh	IHET, U.K., India
Amit Ujlayan	G.B.U., Greater Noida, India
Athar Hussain	G.B.U., Greater Noida, India
Siya Ram	G.B.U., Greater Noida, India
Gopal Sharma	Technical Education and Training, Govt. of M.P., India

Jalaj Sharma	G.B. Pantnagar University, Pantnagar, UK, India
Manisha Manjul	G.B.U., Greater Noida, India
Navneet Singh	G.B.U., Greater Noida, India
Nidhi Gulati	G.B.U., Greater Noida, India

Advisory Committee

Anuradha Mishra	G.B.U., Greater Noida, India
Ela Kumar	G.B.U., Greater Noida, India
Neeti Rana	G.B.U., Greater Noida, India
M.N. Hoda	Bharati Vidyapeeth's Institute of CAM, New Delhi, India

Additional Reviewers

Abhay Kumar, India
Abhijeet Maitra, India
Ajay Kumar Sharma, India
Akhilesh R. Upadhyay, India
Alain Bretto, France
Aleem Ali, India
Al-Sakib Khan Pathan, Malaysia
Amitabh Naag, India
Anil Kumar Sagar, India
Anish Mathuria, UK
Aparajita Ojha, India
Ashok G. Ambekar, India
Atif Azad, Germany
B. Ramadoss, India
B.B. Amberker, India
Bharat Bhargawa, USA
Bhoopendra Dwivedi, India
Bhuperdra Gupta, India
Brahmjit Singh, India
Carlos Becker Westphall, Brazil
Challa Rama Krishna, India
Cheng-Chi Lee, Taiwan
D.A. Ashoko, India
D.K. Mishra, India
D.K. Yadav, India
D.K. Lobiyal, India
Debendra K. Panda, India
Dharm Raj, India
Dimitrios A. Karras, Greece
Edwin Vijay Kumar, India
Eisuke Kita, Japan
Gautam Tiwari, USA
Golden Raymond B., India
Guoyi Peng, Japan
H.J. Kamaruzaman Jusoff, Malaysia
Hasmukh Morarji, Australia
Imran Bashir Bhatti, Germany
Indranil Sen Gupta, India
Irfan Ahmad Khan, India
J.P. Saini, India
Jay Singh, Korea
Jayshree Vajpai, India
Jyoti Prakash Singh, India
K. Mustafa, India
K.B. Mishra, India
M.M. Sufyan Beg, India
M.P. Sebastian, India
Madan Singh, Roma
Madhu Jain, India
Mahesh Chandra, India
Maode Ma, Singapore
Maodema, Singapore
Mayank Mishra, India
Michael Coyle, USA
Mohammad Hashim, Germany
Mohd Kamil, India
Mostafa Fouda, Japan
Murlidhar Kulkarni, India

N.S. Murthy, India
Nalin Sarda, Australia
Neha Gupta, USA
Nika Raatikainen, Finland
Nitin Sharma, India
P. Venkataram, India
P.C. Jain, India
P.C. Jha Du, India
P.K. Singh, India
Pratiksha Saxena, India
Praveen Kumar, India
Preetam Kumar, India
R.B. Mishra, India
R.P. Ojha, India
Raaziyah Shamim, India
Rajeev Kumar, India
Rajendra K. Asthana, India
Rajesh Khanna, India
Rakesh Maheshwar, India
Rama Murthy G., India
Rama Shankar Yadav, India
Ramanujan, India
Ramesh B. Ghodgaonkar, India
Ranjit Biswas, India
Ranvijay Singh, India
Rashidali, UAE
Rishikesh Patankar, India
S. Kazim NAQVI, India
S.B. Sabbarwal, India
S.K. Chaturvedi, India
S.S. Patnayak, India
S.S. Rajput, India
S.S. Sridhar, India
Saba Akhtar, India
Saibal K. Pal, India
Saibal Kumar Pal, India
Saikishopr Elangovan, Australia
Sandeep Lohani, USA
Sanjay Agarwal, Taiwan
Sanjay Jasola, India
Sanjay Singh, Germany
Sanjeev Kumar, India
Sanjeev Tokekar, India
Sanjib Ansekhar Roy, India
Sanoj Kumar, India
Sarsij Tripathi, India
Sathish Babu B., India
Shahid Akram, Germany
Shazil Kamal Shah, Australia
Shirshu Varma, India
Shyam Lal, India
Subhamay Maitra, India
Sudheer Kumar Sharma, India
Sukumar Nandi, India
Tarun Jagani, Germany
Thtikan Boonkang, Thailand
U.B.S Chandrawat, India
Udaishanker, India
V. Lakshmi Narasimhan, USA
Vinod Pandey, India
Virendra C. Bhavsar, Canada
Vrinda Tokekar, India
Winfried E. Kuehnhauser, Germany
Ziyad Al-Khinalie, Malaysia

Technical Sponsoring Institutions

IEEE Delhi Section
Create-Net

Keynote Speakers

Magda El Zark	Professor, CSE, University of California, USA
Anura P. Jayasumana	Professor, Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA

Subhas Mukhopadhyay	Professor, School of Engineering and Advanced Technology, Massey University, Palmerston North, New Zealand
Shang-Kuo Yang	National Chin Yi University of Technology, Taiwan
Bharat Bhargava	Professor, Department of Computer Science, Purdue University, USA
M. Sufyan Beg	Professor, CSE, Jamia Millia Islamia University, New Delhi, India
Rama Shankar Yadav	Professor, CSE, Motilal Nehru National Institute of Technology, Allahabad, India
Mahesh Chandra	Professor, CSE, BITS, Mesra, Ranchi, India
D.K. Lobiyal	Associate Professor, School of Computer and Systems Sciences at Jawaharlal Nehru University, Delhi, India
Prakash Kumar	Country Head, CISCO Systems, India

Table of Contents

Network

Robustness

Security

Reliability

Quality of Service

DSG-PC: Dynamic Social Grouping Based Routing for Non-uniform Buffer Capacities in DTN Supported with Periodic Carriers

Rahul Johari[1], Neelima Gupta[2], and Sandhya Aneja[3]

[1] USICT, GGSIP University, Delhi-110078
[2] Department of Computer Science, University of Delhi, Delhi-110007
[3] Institute of Informatics and Communication, University of Delhi, Delhi

Abstract. Routing a message in networks that are dynamic in nature with time varying partially connected topology has been a challenge. The heterogeneity of the types of contacts available in such a network also adds to complexity. In this paper we present an approach to transfer messages in disruption/delay tolerant network when there is only intermittent connectivity between the nodes. Most of the existing approaches exploit either the opportunistic contacts and transfer messages using the probabilities of delivering a message or use periodic contacts. In addition to opportunistic contacts, we also have scheduled carriers that are available periodically. Scheduled carriers guarantee delivery of the messages to the base station, however, it may have an added delay. If a message can tolerate the delay through the scheduled carrier, it waits else it may be forwarded to an opportunistic contact. We define a utility function for a node to decide whether to forward the message to an opportunistic contact or to a scheduled contact.

This is an improvement over routing through opportunistic contacts that exploits social behavior of the nodes as in [2]. We compare performance of our approach with [2] on message delivery ratio, message delay and message traffic ratio (number of messages forwarded / number of messages delivered) and found that our algorithm outperforms [2] on all the three metrics. We also studied the impact of initializing the probabilities of the nodes proportional to the varying buffer size in [2]. It was found that delivery ratio increased significantly without increasing the message traffic ratio and delay.

1 Introduction

MANETs (Mobile Ad hoc Networks) address challenges related to mobility and low battery life when there is no pre-existing communication infrastructure. In most of routing techniques for MANETs there is an underlying assumption that there always exists a connected path from source to destination. However, there are number of applications in which nodes need to exchange messages over partitioned MANETs called Delay Tolerant Networks. Research in the area of Delay Tolerant Networks (DTNs) has received considerable attention in the last few

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 1–15, 2013.

years owing to their widespread occurrence in a variety of applications such as– (a) InterPlaNetary(IPN) Internet project [3], (b) Wizzy digital courier service which provides asynchronous (disconnected) Internet access to schools in remote villages of South Africa [8], (c) A scenario where a hypothetical village is being served by digital courier service, a wired dial-up Internet connection and a store and forward LEO satellite. Route selection through any one of them depends upon the variety of factors including message source and destination, size, time of request, available connects or other factors like cost and delay etc. [8], and d) Transmission of information/ message during mission critical operations like natural disasters or battle zones.

Delay and Faults are usually tolerable in such aforementioned applications. In this work, we look at the problem of routing in a DTN [6]. A Delay-Tolerant Network is an occasionally connected network that may experience frequent, long-duration partitioning and may never have an end-to-end contemporaneous path. Thus, in contrast to the traditional routing techniques of MANET's where the aim is fast delivery of a message, here the aim is the delivery itself. Since the links are available intermittently, messages must be stored and forwarded later, delays are inevitable in such networks.

One of the earliest approaches proposed for routing in partitioned networks is *epidemic routing* [17]. However, it is very expensive in terms of message traffic and buffer space, which reduces life of the network. To overcome the problems of heavy traffic and high buffer requirement, probabilistic routing [11,19] and social routing [5,7,4] schemes were proposed. Both the approaches provide considerable improvement in delivery ratio under low buffer requirement. The probabilistic routing schemes are based on heuristic that if a node has interacted with a group of nodes in the past it will do so again in the future. On the other hand, the social routing assumes that nodes that are a part of same social network will interact more frequently with the members of that social group. Taking advantage of both the schemes, a dynamic social grouping (henceforth referred to as DSG) method was proposed by Cabaniss et al. in [2] that forms social groups based on contact patterns and use consistent routes to base stations based on delivery patterns to deliver the messages with high probability. However their work considers only the opportunistic contacts to deliver the messages. In practical scenarios, many times one or more scheduled carrier to the destination is also available. Work on scheduled contacts [8,18,16,21] relies on periodic/scheduled contacts to deliver the message. One advantage of the scheduled contacts is higher probability of delivery where as the delay may be more. Another advantage is that since the scheduled contacts are special nodes, they are equipped with better resources, say more buffer capacity.

We propose an approach that combines the advantages associated with social routing, probabilistic routing and scheduled contacts for routing a message. For routing we assume social groups among opportunistic nodes are formed in the network in a similar way as that in DSG. Periodic carriers do not participate in group formation and mergers. We update the individual probabilities even when two nodes meet and there are no messages to be exchanged. In contrast to DSG where the initial probabilities of the nodes are uniform, we assign the initial probabilities proportional to the buffer capacity of the nodes. We first show the impact of doing this on DSG and then extend the work of [2] to include scheduled carriers whose probability to deliver a message to the base station is very high as compared to the probability of other nodes. We define a utility function for a node to choose between an opportunistic carrier and a scheduled carrier. The buffer capacity of the scheduled carrier is higher than that of the other nodes. To be able to use groups to forward messages we make use of contact strength to define joint individual probabilities that are used to make routing decisions. We show through simulation that the message delivery ratio, message delay and the traffic ratio improve considerably over DSG when the time period of the carriers is not too big. We also show the impact of time period of the carrier nodes on the performance. It was observed that delivery ratio increased significantly without increase in the message traffic ratio and delay.

The paper is organized as follows: Section 2 presents the related work done in the area of routing in delay tolerant network. Section 3 describes the problem. Section 4 presents the proposed algorithm. Section 5 presents experimental setup and analysis of results generated after the implementation of the algorithm on ONE Simulator [10]. Section 6 presents conclusion and future work.

2 Related Work

One of the earliest approaches proposed for routing in partitioned networks is *epidemic routing* [17]. The goal of epidemic routing was to maximize the message delivery ratio while minimizing the message delivery time. It relies on replicating messages through buffer contents synchronization when two nodes come in contact until all nodes have a copy of every message. It operates without knowledge of the communication pattern and is well-suited for networks where contacts between the nodes are unpredictable. However, it is very expensive in terms of message traffic and buffer space, which reduces the life of the network. That is, the approach does not seem to scale as the number of messages in the network grows. An intelligent buffer management scheme can improve the delivery ratio over the simple FIFO scheme. Epidemic routing is sometimes useful in transferring control messages in a part of a network.

In [11,12,13,20] authors claim that mobility is mostly not random and there is a pattern in encounters. The Probabilistic Routing Schema is based on individual probabilities of nodes of successful delivery of a message. Delivery probabilities are computed using the history of encounters and transitivity to reflect that if a node has been encountered in the past, it is likely to be encountered again. When two nodes come in contact, the one with lower probability of delivery forwards its messages to the one with higher probability updating its own probability upward as it does so. If a node drops a message its probability is reduced to reflect the nodes inability to transfer. This algorithm shows a marked improvement in terms of message transmission rate while maintaining a high delivery probability under the resource (buffer capacity and bandwidth) constraints. The approach introduced by Wang et al. in [19] is also based on probabilities of nodes of successful delivery of a message to a base station. However the delivery probabilities are computed based on successful delivery of messages rather than regular contacts between the nodes.

The SimBet routing algorithm [5] borrows ideas from social networking and contact patterns to predict paths to destinations to improve delivery ratio and time. The algorithm assumes that the groups of frequently encountered nodes have been formed and calculates the Betweenness and Centrality metrics to make routing decisions. Bubble Rap [7] extends their work by allocating nodes into social groups based on direct and indirect contacts. Based on the global knowledge of the nodes contact, they construct k-cliques/clusters to define the groups. However, grouping is static in their approach. In [1,2] Cabaniss et al presents an approach that combines advantages of social grouping and message forwarding based on probabilities of delivery of message of a node. Groups are formed dynamically as nodes come in contact with each other.The message is either forwarded using individual delivery probability of the node, which is computed on the basis of message forwarded by the node, or using group probability of groups of which node is the member. In [2] the authors present an application of social grouping on network with single base station and in [1] approach for network with n message sources and sinks is presented.

Jain et al. [8] proposed an approach for routing in DTN using modified Dijkstra. They assumed that DTN nodes posses knowledge about time and duration of contact. On the basis of this knowledge a node may create number of routing metrics. Their results show that the efficiency and performance increases with the amount of information used for the metric. In DTN determining futuristic knowledge of opportunistic contacts is extremely difficult and therefore the approach presented in [8] may only exploit the scheduled contacts for message transfer and opportunistic contact may not be used.

Jones et al. in [9] presented an approach that is using link state routing protocol. The link state packets are exchanged using epidemic approach that leads to overhead. Whenever a node has a message to transfer, it forwards it to the node that is closest to the message destination. The approach is good in the sense that it requires very small buffer space and is therefore scalable. But the approach may not be practical in the scenario wherein the nodes have sparse connectivity.

The concept of using carrier nodes to transfer data to nearest access point in sensor networks is used in [16]. The carrier nodes acting as data mules follow a random walk and come in contact with sensor nodes. The movement of the data mules generates opportunistic contacts that are exploited to deliver the message. The concept of using a ferry for carrying the messages to the destination was introduced by Zhao et al [21]. A ferry following a fixed schedule moves along a fixed trajectory. Two different approaches have been presented. Firstly, a node may adjust its own path so that it comes in contact with ferry node, when it has a message to be delivered. In the other approach a node calls up a ferry on demand. The ferry adjusts its path so as to come in contact with the node requiring its services. The approach relies on direct contact between the ferry and the sender/receiver nodes. It fails to exploit the mobility and the contacts between the nodes to deliver or receive messages to/from the ferry.

3 Problem Definition

Consider a scenario where people (showing social behavior) living in remote villages have to transfer mails/ messages to a central office (fixed base station) located at an urban area far away from the village. Periodically a postman visits the villages. The people in the village are well aware about the day and the time of visit of the postman. Occasionally people from the village may also visit the urban area and may deliver the message to the central office. A villager, wanting to send a message to the central office, needs to take a decision as to whether he should give his message to another villager, planning to visit the city, whose probability of visiting the central office is low or to wait for the postman to arrive. Thus, we require a routing approach that exploits availability of the postman (acting as periodic carrier) and the movement of village people (acting as opportunistic contact) to maximize the number of messages delivered without much delay. In this paper we present an approach that maximizes the message delivery in a DTN having opportunistic as well as periodic contacts.

4 Preliminaries

In this section we briefly explain the DSG based algorithm of Cabaniss et al. [2]; The nodes having common interest tend to meet frequently. Groups are formed in DSG so that good contact strength between nodes of the common interest can be exploited to route messages. A group consists of at least two nodes. First time a group is formed when the contact strength between two nodes exceeds the threshold (ψ). The contact strength $\lambda_{A,B}$ between two nodes $Node_A$ and $Node_B$ is computed as in [2] i.e. $\lambda_{A,B}$ is initialized to zero and is updated as follows:

$$\lambda_{A,B} = (1-\alpha)\lambda_{A,B} + \alpha \frac{time_{contact}}{time_{contact} + time_{nocontact}} \tag{1}$$

where α is a control parameter. The nodes common to two groups may initiate the process of enlarging the group by mergers. Mergers are initiated when the ratio of overlapping members of the group to the unique members of the union is above threshold (τ). The updated information regarding the group is maintained by the cluster-heads or the group heads. Any node which does not want to be a part of the group may resign from it and the group dynamism is maintained. The nodes of same group coming in contact updates the group information on the basis of group versions. Group formation, group merger, node resignation and group updates are done in the same manner as in [1].

Besides contact strength with other nodes in the network, each node maintains a probability of delivery σ_A to the base station. Initially all the nodes except the base station are assigned uniform probabilities with base station having probability 1. Probabilities are updated as the messages are forwarded. Each cluster head also maintains group's probability of delivering a message to the base station. Group's probability is the average of individual probabilities of its members. Besides individual probabilities, each node maintains a list β_A of the group probabilities of all the groups $Node_A$ is a member.

Let $\Delta_A = max\{\beta_A, \sigma_A\}$. When two nodes $Node_A$ and $Node_B$ come in contact, they compare their probabilities of delivering a message to the base station individually or through one of its groups. $Node_A$ forwards its messages to$Node_B$ if $\Delta_B > \Delta_A$. $Node_A$ also increments its probability as per the following formula:

$$\sigma_A = (1 - \phi)\sigma_A + \phi\Delta_B \tag{2}$$

where ϕ defines the weight factor.

In [1], 'joint individual probability' (γ_A) instead of individual probability is used to make the routing decisions. Joint individual probability signifies the node's capability to deliver the message to the destination through itself or through any of its groups. The joint individual probability of $Node_A$ is computed as follows:

$$\gamma_A = 1 - \{\{1 - \sigma_A\} \times \Pi_{i=1}^{|GroupsofA|}(1 - \beta_{Gp_i})\} \tag{3}$$

where β_{Gp_i} is the group probability of $Group_i$. When two nodes say $Node_A$ and $Node_B$ meet they exchange their joint individual probability and $Node_A$ forwards its messages to $Node_B$ if $\gamma_B > \gamma_A$. $Node_A$ with lower joint individual probability updates its individual probability as follows:

$$\sigma_A = (1 - \phi)\sigma_A + \phi \cdot \gamma_B \tag{4}$$

When a node drops a message due to ttl expiry or buffer overflow its delivery probability is decreased as follows: $\sigma_A = (1 - \phi)\sigma_A$ to indicate node's inability to forward it.

5 Improvements Proposed in DSG for DSG Supported with Periodic Carriers (DSG-PC)

As mentioned in previous section, DSG uses social group formation and probability of delivery to take routing decision for a message in a network of nodes with uniform buffer capacities. In this section, we propose improvements in procedures of initialization, evaluation and updation of individual probabilities by considering node's bufferspace as a part of their capability to deliver message.

5.1 Individual Probability and Initialization of Individual Probabilities

In contrast to DSG where the individual probabilities are initially assigned uniformly to all the nodes, we assign the individual probabilities proportional to buffer space of a node based on the theory that a node with larger buffer space is a better candidate to deliver the message, by virtue of retaining the message for longer duration, than the one having less buffer space. In our approach we have assigned the initial delivery probability to all the nodes on the basis of their buffer space which is different from the one presented in DSG wherein all nodes are assigned equal probability initially. We compute initial probability of $Node_A$ as follows:

$$\sigma_A = \frac{BufferSpace_A}{MaxBufferSpace} \tag{5}$$

where $MaxBufferSpace$ is a network wide parameter. Probabilities are updated from time to time as explained in next section.

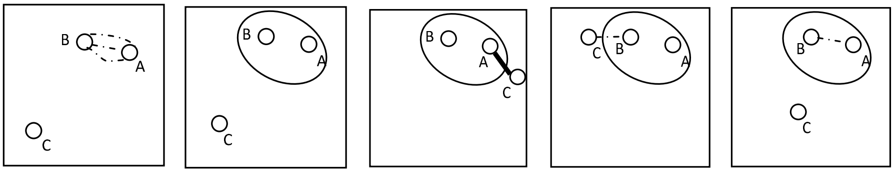
5.2 Delivery Probabilities

Joint individual probability as defined in [1] allows a node with low individual probability and weak contact, with other group members, to accept messages which it is less likely to deliver. Consider the situation when node's average contact strength with its group members is low say 0.2 but above the threshold of resigning. A message should not be forwarded to such a node just because its group probability is high. For example, consider $Node_A$ with $\sigma_A = .1$ member of two groups G1 and G2 with group probability as P(G1)=.7 and P(G2)=.8. Using the approach used in [1], the $\gamma_A = 0.946$. This means $Node_A$ will receive the messages but is less likely to forward it to other members of group due to low average contact strength. We have used the measure of average contact strength in determining the joint individual probability of a node. We propose to compute the joint individual probability of $Node_A$ as follows:

$$\gamma_A = 1 - \{\{1 - \sigma_A\} \times \Pi_{i=1}^{|GroupsofA|} 1 - \lambda_{A,Gp_i} \cdot \beta_{Gp_i}\} \tag{6}$$

where λ_{A,Gp_i} is the average contact strength of $Node_A$ in $Group_i$ and β_{Gp_i} is the group probability of $Group_i$. Now the joint individual probability of $Node_A$ with weak average contact strength i.e (0.2) with this group member is 0.35. Clearly λ_A as defined above is a better measure to capture the node's capability to deliver the messages.

DSG updates the delivery probability only when the message is delivered; however even when two nodes come in contact but they have no message to exchange, their individual probabilities should be updated. Consider an example wherein there are three nodes in a network, $Node_A$, $Node_B$ and $Node_C$ (see figure 1). Suppose $Node_A$ is a base station with probability P(A)=1.0. Initially the delivery probability of other nodes is 0.0 i.e. P(B)=0.0 and P(C)=0.0. Suppose $Node_B$ comes in contact with $Node_A$ frequently thereby increasing their contact strength but have no messages to exchange. So B and A form a group Gp_1 with probability of B as P(B)=0.0, and group probability say $P(Gp_1)$=0.50. Now C meets A and delivers 10 messages thereby increasing probability of C to say P(C)=0.55. Next, C moves close to B and meets B. C has a message for A. Now C will not forward the message to B. Next B meets A (they are in the same group). Message of C could have been delivered to A through B if we had



(a) Base station ($Node_A$), and $Node_B$ meet frequently. $Node_B$ has no messages to transfer

(b) $Node_B$ and $Node_A$ form a group

(c) $Node_A$ and $Node_C$ meet and $Node_C$ delivers messages to $Node_A$ and increases its probability

(d) $Node_C$ meets $Node_B$, a good node for delivering the message, but $Node_C$ does not forwards the message to $Node_B$

(e) $Node_B$ meets $Node_A$. $Node_C$'s message could have been delivered

Fig. 1. A scenario of message exchange and updation of probabilities in DSG

forwarded the message to B. We update the probabilities when the connection between two nodes terminates, even though they do not have a message to exchange. However, we do so at the time of connection removal and only when the contact duration is sufficient to transfer at least one message. In the above example the contact between $Node_A$ and $Node_B$ would lead to increase in probability of $Node_B$ to say P(B)=0.40 and then the group probability will become P(Gp_1)=0.70. In this scenario $Node_C$ would forward the message to $Node_B$ and the message would be delivered.

6 Proposed Algorithm

The proposed algorithm, DSG-PC, uses scheduled carriers to transfer messages in an intermittently connected network, running delay tolerant applications, equipped with periodic carriers and has advantages over probabilistic routing and social grouping behavior. The model of periodic carrier is somewhat similar to the DAKNet [14] but DAKNet does not take advantage of the opportunistic contacts and regular movement patterns to deliver messages. DSG-PC exploits availability of opportunistic contacts as well as periodic carriers.

We propose a periodic carrier node that is an especially designed drop- in node following a fixed trajectory over fixed time interval and has large buffer capacity. The schedule (time and place where periodic carrier would meet the DTN nodes) of the periodic carrier is known in advance to all the DTN nodes in the network. For example at time say 9'o clock a periodic carrier starts its schedule from its office and meets a node say $Node_A$ between 10.00 to 10.30. It meets the $Node_B$ between 12.00 to 12.30 and then delivers its messages to base station between 1.00 to 1.30. Next day again it follows the same schedule to collect and deliver messages. $Node_A$ and $Node_B$ knows in advance that they can meet periodic carrier at particular fixed schedules and also for fixed duration. In

DSG-PC, periodic carriers do not generate any messages and are used to improve timely delivery of messages in the scenario like delay tolerant network with opportunistic nodes. A periodic carrier does not participate in group formation and mergers.

We define a utility function that helps a node to choose among opportunistic contacts or periodic carriers. The decision of forwarding a message to an opportunistic contact or to a scheduled carrier depends upon the delay tolerance capabilities of the message. Each node at the time of contact with any other node checks delay tolerance capabilities of all the messages in the buffer using their creation time (m_{toc}) and time to live (TTL_m) parameters. The messages which can tolerate the delay of at least one scheduled carrier are kept in buffer. On the other hand for the messages which cannot tolerate the delay the utility value of the encountering node say i for the message m ($U_i(m)$) is computed. The node in contact forwards the messages only if its utility is less than the encountering node. The utility value of an encountered node is computed as follows:

$$U_i(m) = jointIndProbability_i \times \begin{cases} f(m)\ when\ Node_i\ is\ opportunistic\ carrier \\ !f(m)\ when\ Node_i\ is\ scheduled\ carrier \end{cases} \tag{7}$$

where

$$f(m) = \begin{cases} 1\ when\ (m_{toc} + TTL_m - CurrTime) \leq min(CD_j) \\ 0\ when\ (m_{toc} + TTL_m - CurrTime) > min(CD_j) \end{cases} \tag{8}$$

$CurrTime-$ Current Time and CD_j-Delay introduced byj^{th} carrier for $j = 1 \dots n$ and $n =$number of carriers. Intuitively $f(m) = 1$ means that the message cannot tolerate the delay by one of the scheduled carriers and hence the utility value of any $Node_B$ for m is 0. $f(m) = 0$ means that the message can tolerate the delay by any of the scheduled carriers and hence the utility value of any $Node_B$ for m is nothing but its joint probability of delivering a message to the base station.

Thus if message can tolerate delay introduced by a periodic carrier it waits for the periodic carrier to arrive since it will guarantee delivery of the message, else it uses DSG like approach to forward message to an opportunistic contact i.e. the message is forwarded to an opportunistic contact with higher probability. For example, consider a node $Node_A$ having message m comes in contact with $Node_B$ at current time say 100 sec. Now suppose message m can tolerate the delay according to its TTL requirement say for 50 sec. Now, if any of schedule carrier delivers before 50 sec (i.e. message can tolerate the delay) then f(m) evaluates to 0 and $U_B(m)$ also evaluates to 0. Hence $Node_A$ does not forward the message to opportunistic contact $Node_B$. In case message is not able to tolerate the delay then $U_B(m)$ evaluates equal to $Node_B$'s joint individual probability. Now like DSG, if $U_B(m) > U_A(m)$, $Node_A$ forwards message to $Node_B$ otherwise not.

We assume that groups along with the individual and group probabilities of the nodes have already been computed using an algorithm like DSG.

When $Node_A$ comes in contact with $Node_B$, it makes a decision whether to transmit a message to $Node_B$ or not. The decision of the node is based on the following factors :- a) If $Node_B$ is a destination node the node A transfers all its messages. b) If $Node_B$ is any of the carrier nodes, then $Node_A$ transfers only those messages to $Node_B$ which can sustain the delay of $Node_B$ reaching the destination . c) If $Node_B$ is an opportunistic contact, it determines if any carrier is scheduled shortly whose delay it can sustain. If so, it waits for the carrier to arrive as it guarantees the delivery of message where as the opportunistic contact would deliver the message only with some probability. Otherwise, the two nodes exchange the joint individual probability and if the probability of $Node_B$ is higher than $Node_A$ it forwards the message to opportunistic contact.

Algorithm 1. DSG-PC Routing Algorithm

```
Notation
$jointIndProbability_A -$ Joint individual Probability of A
$jointIndProbability_B -$ Joint individual Probability of B
$message_i -$ Message in DTN
$Dm_i -$Delay sustained by message
$S_i -$Carrier Node
$Dest -$ Message Destination

Trigger$-$ $Node_A$ contacts $Node_B$
In $Node_A$
if Node$_B = Dest$ then     Deliver message to $Node_B$
else if $Node_B = S_i$ then
    for all $message_i$
        $CarrierDelay = Compute\_Carrier\_Delay()$
        if $Dm_i > CarrierDelay$ then
            Deliver Message to $Node_B$
        end if
    end for
else
Transmit $jointIndProbability_A$ to $Node_B$
Receive $jointIndProbability_B$ from $Node_B$
if $jointIndProbability_B > jointIndProbability_A$ then
    $CarrierDelay = Compute\_Carrier\_Delay()$
    for all $message_i$
        if $Dm_i > CarrierDelay$ then
            Wait for the Carrier
        else
            Transmit messages to $Node_B$
        end if
    end for
else
```

```
    Receive messages from $Node_B$
end if
Compute_Carrier_Delay()
Notation
$C-$ set of carrier nodes
$T_i-$ Delay in reaching to $Node_A$ in $i^{th}$ cycle for $j_{th}$ carrier
$T_{Dest}-$ Delay in reaching destination from $Node_A$ for $j_{th}$ carrier
$Delay_j-$ Delay for $j_{th}$ carrier
In $Node_A$
    $Delay_j = T_i + T_{Dest}$
    $CarrierDelay = min(Delay_j)$ where $j = 1$ to $j \leq |C|$
return $CarrierDelay$
```

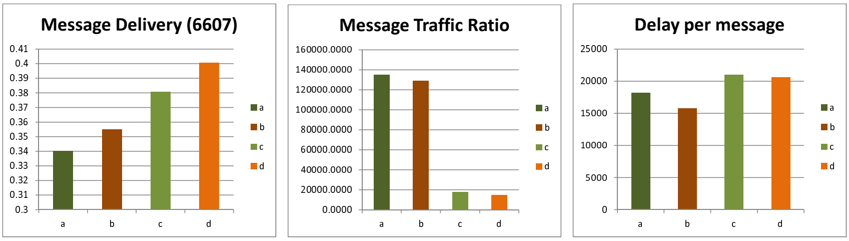
7 Analysis

7.1 Experimental Setup

The Opportunistic Network Environment(ONE) simulator [10] implemented in Java and available as open source has been used to evaluate the algorithms. The ONE simulating environment is capable of simulating the mobility pattern of the nodes and the message exchange between them. Many of the routing algorithms applicable to DTN environment are pre-implemented in the simulator. We implemented the routing as used in our algorithm and the one used in DSG by extending the functionalities available in ONE. Three metrics viz message delivery ratio, message traffic ratio and delay per message were used to compare the performance of DSG and DSG-PC. The simulator generated the message delivery ratio and the message traffic but the functionality for generating delay per delivered message was added to the simulator. The experimental setup was also slightly modified to study the impact of assigning initial delivery probabilities proportional to the buffer space of the DTN nodes on three metrics. The simulation was repeated for both replicated and non-replicated message forwarding.

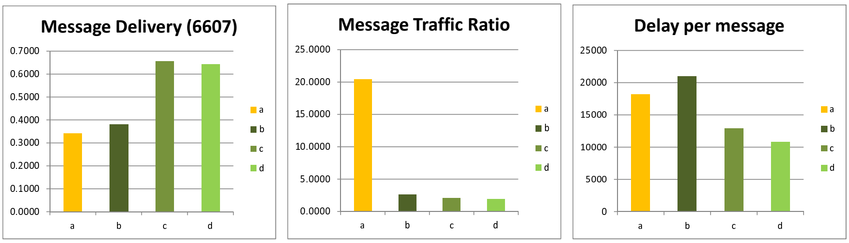
7.2 Data Source and Simulation Parameter

In order to objectively compare the results of DSG-PC with DSG extensive simulations were carried out on the data obtained from an experiment conducted at University of Cambridge at the 2005 Grand Hyatt Miami IEEE Infocom conference as used in DSG [15]. We also added the contact pattern of two periodic carrier nodes, following a fixed trajectory and a fixed time period, with other nodes in the data. For simulation purpose one of the node, as mentioned in the data, was considered to be a sink/base station. The simulation was run on the whole data set. Total number of messages generated were 6607. Message size was varied between 512 KB to 1MB. The transmission speed of nodes was 256kBps. The buffer space for carrier node was 2000MB and for DTN nodes it



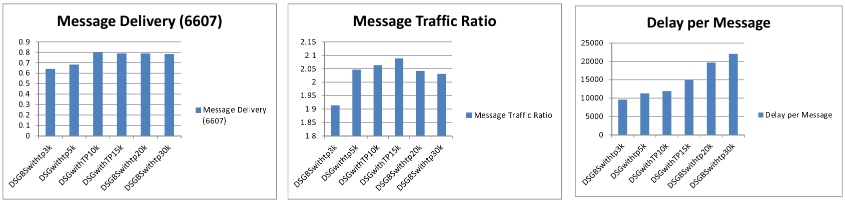
(a) Message delivery ratio (b) Message traffic (c) Delay per message

Fig. 2. Comparison of Impact of initial equal individual probability and initial individual probability proportional to buffer space. where 'a'is DSG Replicate, 'b 'is DSG Replicate with probability proportional to buffer space, 'c'is DSG Non Replicate and 'd'is DSG Non Replicate with probability proportional to buffer space



(a) Message delivery ratio (b) Message traffic (c) Delay per message

Fig. 3. A comparison of DSG based Routing in DTN with Opportunistic contacts and Periodic contact where 'a'is DSG Replicate, 'b'is DSG Non Replicate, 'c'is DSG-PC with Periodic Carrier(10k Time Period), and 'd'is DSG-PC with Periodic Carrier(3k Time Period)



(a) Message delivery ratio (b) Message traffic (c) Delay per message

Fig. 4. Impact of Variation in Periodicity of Carriers

was varied from 100MB to 30MB. The message TTL(time to live) was set as one day(1440 min). The threshold used for group formation (ψ) is 0.004 and for group merger(τ) is 0.300 as used in DSG. The probability decay rate (ϕ) is 0.075 and that of contact decay ratio (α) is 0.300.

7.3 Comparison of Results

We conducted two sets of experiment. The results were compared for replicated as well as non replicated message forwarding. In the first experiment, we compared and analyzed the results of assigning the initial individual probabilities proportional to the node's buffer space. The results (see figure 2) show that the DSG with initial individual probabilities proportional to the buffer space (say A) performed better than DSG with nodes assigned equal probabilities (say B) for all the metrics namely message delivery ratio, message traffic ratio and message delay.

For the second experiment we compared our results i.e. the results of DSG-PC with that of DSG i.e. the initial individual probabilities were assigned proportional to the buffer space and periodic carriers were added. The review of results (see figure 3) indicates DSG-PC (exploiting both the scheduled as well as opportunistic contacts) outperforms DSG (exploiting only opportunistic contacts) on all the three metrics. The message delivery ratio for DSG-PC shows (see figure 3a) an improvement of 85-94% over DSG with replicated forwarding and 65-73% with non-replicated forwarding. Also the message traffic ratio (see figure 3b) for replicated forwarding reduced by 89-90% and that for non-replicated forwarding it reduced by 22-28% and the delay suffered by a message (see figure 3c) reduced by 29-40% for replicated forwarding and 38-50% for non replicated forwarding.

7.4 Impact of Periodicity of Scheduled Contact

The experiments for DSG-PC were repeated with periodic carriers having time periods 3k, 5k,10k, 15k, 20k and 30k sec (see figure 4). It was observed that the message delivery ratio (see figure 4a) increased when the periodicity of the carriers was varied from 3k sec to 10k sec. Beyond 10k sec i.e. for 20k and 30k sec the delay per message (see figure 4c) also increased considerably as was expected.

8 Conclusion and Future Work

The proposed algorithm improved message delivery ratio, message traffic ratio, and delay significantly by using buffer capacity to assign initial probabilities and exploitation of periodic contacts along with opportunistic contacts. While the cost of setting up this type of network will increase marginally due to cost of introducing a carrier, but since improvements in terms of outcome are substantial so cost will not be an issue. Further, as future work, an application of DSG may be widened by attaching communities to social groups. Attaching communities to the social groups may be useful to ensure that messages from one community do not travel through another.

References

1. Cabaniss, R., Bridges, J.M., Wilson, A., Madria, S.: Dsg-n2: A group-based social routing algorithm. In: 2011 IEEE Wireless Communications and Networking Conference (WCNC), pp. 504–509 (March 2011)
2. Cabaniss, R., Madria, S., Rush, G., Trotta, A., Vulli, S.S.: Dynamic social grouping based routing in a mobile ad-hoc network. In: Proceedings of the 2010 Eleventh International Conference on Mobile Data Management, MDM 2010, pp. 295–296. IEEE Computer Society, Washington, DC (2010)
3. Cerf, V., Burleigh, S., Hooke, A., Torgerson, L., Durst, R., Scott, K., Travis, E., Weiss, H.: Status of this memo interplanetary internet (ipn): Architectural definition
4. Costa, P., Mascolo, C., Musolesi, M., Picco, G.P.: Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks. IEEE Journal on Selected Areas in Communications 26(5), 748–760 (2008)
5. Daly, E.M., Haahr, M.: Social network analysis for routing in disconnected delay-tolerant manets. In: Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2007, pp. 32–40. ACM, New York (2007)
6. Fall, K.: A delay-tolerant network architecture for challenged internets. In: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, SIGCOMM 2003, pp. 27–34. ACM, New York (2003)
7. Hui, P., Crowcroft, J., Yoneki, E.: Bubble rap: social-based forwarding in delay tolerant networks. In: Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2008, pp. 241–250. ACM, New York (2008)
8. Jain, S., Fall, K., Patra, R.: Routing in a delay tolerant network. SIGCOMM Comput. Commun. Rev. 34(4), 145–158 (2004)
9. Jones, E.P.C., Li, L., Schmidtke, J.K., Ward, P.A.S.: Practical routing in delay-tolerant networks. IEEE Transactions on Mobile Computing 6(8), 943–959 (2007)
10. Keränen, A., Ott, J., Kärkkäinen, T.: The ONE Simulator for DTN Protocol Evaluation. In: Proceedings of the 2nd International Conference on Simulation Tools and Techniques, SIMUTools 2009, New York, NY, USA, ICST (2009)
11. Lindgren, A., Doria, A., Schelén, O.: Probabilistic routing in intermittently connected networks. SIGMOBILE Mob. Comput. Commun. Rev. 7(3), 19–20 (2003)
12. McNamara, L., Mascolo, C., Capra, L.: Media sharing based on colocation prediction in urban transport. In: Proceedings of the 14th ACM International Conference on Mobile Computing and Networking, MobiCom 2008, pp. 58–69. ACM, New York (2008)
13. Nelson, S.C., Bakht, M., Kravets, R., Harris III, A.F.: Encounter: based routing in dtns. SIGMOBILE Mob. Comput. Commun. Rev. 13(1), 56–59 (2009)
14. (Sandy) Pentland, A., Fletcher, R., Hasson, A.: Daknet: Rethinking connectivity in developing nations. Computer 37(1), 78–83 (2004)
15. Scott, J., Gass, R., Crowcroft, J., Hui, P., Diot, C., Chaintreau, A.: CRAWDAD trace (January 31, 2006), `http://crawdad.cs.dartmouth.edu/cambridge/haggle/imote/intel`
16. Shah, R.C., Roy, S., Jain, S., Brunette, W.: Data mules: modeling and analysis of a three-tier architecture for sparse sensor networks. Ad Hoc Networks 1(2-3), 215–233 (2003)

17. Vahdat, A., Becker, D., et al.: Epidemic routing for partially connected ad hoc networks. Technical Report CS-200006, Duke University (2000)
18. Vu, L., Do, Q., Nahrstedt, K.: Comfa: Exploiting regularity of people movement for message forwarding in community-based delay tolerant networks
19. Wang, Y., Wu, H.: Delay/fault-tolerant mobile sensor network (dft-msn): A new paradigm for pervasive information gathering. IEEE Transactions on Mobile Computing 6(9), 1021–1034 (2007)
20. Yuan, Q., Cardei, I., Wu, J.: Predict and relay: an efficient routing in disruption-tolerant networks. In: Proceedings of the Tenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2009, pp. 95–104. ACM, New York (2009)
21. Zhao, W., Ammar, M., Zegura, E.: A message ferrying approach for data delivery in sparse mobile ad hoc networks. In: Proceedings of the 5th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2004, pp. 187–198. ACM, New York (2004)

Integrated Approach for Multicast Source Authentication and Congestion Control

Karan Singh* and Rama Shankar Yadav

CSED, MNNIT, Allahabad, U.P., India
School of I.C.T., Gautam Buddha University, Greater Noida, India
karan@gbu.ac.in, rsy@mnnit.ac.in

Abstract. The coming age is information age in which data is being transmitted from network source to destination using unicast and multicast. Multicast services are very popular for transmission of huge information. Therefore, multicast network are growing day by day and it faces various problems such as reliability, security, congestion, connectivity scalability, fairness etc., due to exponential increment of network. Multicast Congestion is very serious problem to decrease the network utilization if network is not secure then condition may be worst and it is difficult to handle the situation. In this paper, we are providing secure multicast congestion control mechanism. In this mechanism global and local approach is proposed which provide the secure information in presence of congestion at minimum or any cost.

Keywords: Computer Network, Multicast Communication, Congestion Control, Source Authentication, Attack, Congestion Control, Security Goal, Layer System.

1 Introduction

Computer Network is essential part of our daily life. We perform various tasks such as email, newsfeed, stock information, IP TV, video conference etc. that use the unicast, broadcast and multicast transmission technology. In case of multicast huge amount of data is transfer from one computer to group of computer whereas it is efficient then unicast and broadcast but the design architecture raise various problem such congestion [1], security [12], fairness [2], reliability [17] etc. In case of congest network performance is decreased due to packet loss. So, each layer has different and independent authentication tree as well as signatures. Each receiver receives packets from joined layers then it verify the signature (reference signature is decrypted using public key known to receiver resulting to digest for the signature) now it compute the digest of received message using same hash algorithm as used at the sender side. If both digests matches, the received message is authentic. Besides of security mechanism for authenticity of source, the receiver performs the operations to maintain the desired performance and overload due to congestion in the system. The

* Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 16–30, 2013.

source S generates signature, computes authentication tree and generates packets of 3 independent layers [11, 20].

We can observe from figure 1 that the colors black, green and blue define for base layer (L_0), enhance layer 1 (L_1), enhance layer 2 (L_2) respectievely. The packets generated on different layers are of same color. Source sends packets with authentication information (hashes) from authentication tree in each layer independently. So, here is one authentication tree of each layer for one block (depend upon proposed approach). The receiver can join or leave the layer according to its own capacity or congestion situation. There are 3 receivers $R_{1,}$ R_2 and R_3 whereas receiver R_1, R_2 and R_3 have joined layers (L_0, L_1, L_2), layers (L_0, L_1) and layer (L_0). The next section discusses the constraints behind integrated security aware multicast congestion control approach for multicast communication.

1.1 Constraints while Integration

The source equipped with the security mechanism send group of packets rather than a single packet and the level of security depends upon the random behavior of attacker. In other hand, multicast congestion control approach uses multiple layer joining and adaptive deaf period concept for leaving layer. Security mechanism provides the authenticity while it increases the communication overhead which may be the reason behind congestion problem. In other hand, congestion control manages the overhead but it may create problems (such as packet during deaf concept) for security mechanism. Thus security mechanism and congestion control are orthogonal issues, so there can be many constraints during integration which create open ended question disused in figure 1.

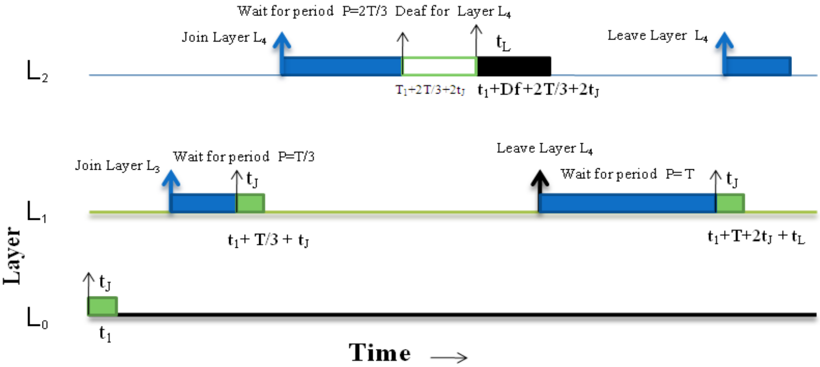


Fig. 1. Layer Deaf and Leaving Approach

In others case, figure 2 shows the deaf and leaving approach where base layer is L_0 (Black color, enhance L_1 (green color) and enhance L_2 (blue color) represented by line while box (🟩 , ⬛ , □ and 🟦) are represented for joining overhead, leaving overhead, deaf overhead and decision period respectively. Suppose R_1 join the layers L_0, $L_{1,}$ L_2 according to proposed method. At $t = t_1 + \frac{2T}{3} + 2t_J$, the receiver is overloaded and become deaf for the higher layer (L_2).

It can be observed that packets are lost when receiver is either overloaded or deaf for higher layer. Due to the lost packets the required secure information for such pakets are also lost. The chain could be break due to the congestion or loss in the secure information. At this stage the authenticity of the received stream is a major concern due to overloading or deaf decision.

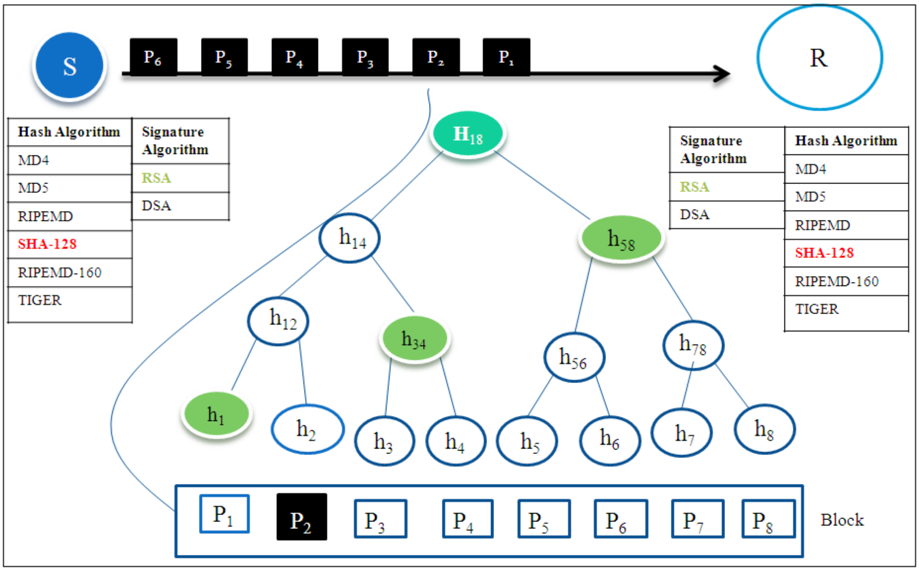


Fig. 2. Packet with Secure Information in a Layer

The communication overhead of packet in a layer depends upon number of hashes attached with packet, overhead of hash algorithm used, overhead of signature algorithm used. For example, we can see figure 2 where source is sending packets of base layer (L_0) to receivers. The hash values h1, h2, h3… h8 are computed by sender corresponding to packets P1, P2, P3.... P8 using hash algorithm and signature are generated by signature algorithm.

Here one source S is sending packets with secured information (number of hashes regarding hashes, for example secure information of P2 is h1, h34 h58) to receiver. It can be observed that more number of hashes, signature and increment in security level (to generate hashes) will create the communication overhead. Thus, on tuning with increased security level, hashes may lead to overloading and more packet loss. So, more attack probability leads to more congestion, more packet loss provide more security threats. It is a big challenge to manage the effect of increasing security overhead on overloading.

The tuning with increased security level, hashes may lead to overloading.

2 Related Work

2.1 Multicast Congestion Control

Computer networks uses the channels for transmit the data from source to receivers. If source rate increases the capacity of channel then congestion occur [7]. There are various multicast congestion control algorithms viz. RLM [4], TFMC [9, 10], FLID-DL [2], RLC [13], WEBREC [9], QIACCRM [5], EJLRDMC [11] etc. which control only the congestion and do not address the security threat. A few algorithms of congestions are describe as follows:

Receiver-driven Layered Multicast is the first well-known end to end congestion control for layered multicast. In RLM, receiver detects network congestion when it observes increasing packet losses. Receiver reduces the level of subscription if it experiences congestion. In the absence of loss, the receiver estimates the available bandwidth by doing the so-called join experiments when the join-timer expires. A join experiment means that a receiver increases the level of subscription and measures the loss rate over a certain period. If the join-experiment causes congestion, the receiver quickly drops the offending layer. Otherwise, another join-timer will be generated randomly and the receiver retains the current level of subscription and continues to do the join experiments for the next layer once the newly generated join-timer has expired.

Efficient Joining and Leaving for Receiver Driven Multicast Congestion Control (EJLRDMC) [11] have provided efficient layer joining and leaving through multiple layer joining and deaf leaving mechanism respectively.

Thus, we can see if source, router or receivers are working as a attacker then congest may increase more and network utilization will decrease so we need such type of mechanism which provide the authenticity of source and receivers. In next section we are providing secure multicast scheme to control the misbehavior of attack on system.

2.2 Multicast Source Authentication

Multicast source authentication provides [15, 16] the authenticity of sender to all receivers. This section is providing various types of secure multicast communication scheme which protect the network with security services such as authentication, Non-repudiation, Integrity etc. It divides the stream into blocks and embeds in the current block a hash of the following block. In this way, sign only the first block and then the properties of this single signature will propagate to the rest of the stream through the hash chaining .It is Off-line because entire stream is known in advance and this solution is not for fault tolerant.

EMSS [21] provides more or less probabilistic guarantees that it remains a hash-chain between the packet and a signature packet, given a certain rate of packet loss in the network. The robustness of the protocol to packet loss is proportional to the redundancy degree, k. In order to assure authentication of the stream, the sender continuously sends periodic signature packets .To verify authenticity of received packets, a receiver buffers received packets and waits for their corresponding signature packet. The signature packet carries the hashes that allow the verification of few packets. These latter packets carry, in turn, the hashes that allow verifying other packets, and so on until the authenticity of all received packets is verified.

In second approach, packet are sending the same things (key, hash value, hash chaining) with a block of packet. But in this approach main problem will come after packet loss. If at any packet or block, the approach fails, the packet loss should not exceed the threshold limit.

Hash chaining scheme can't tolerate packet loss and the receiver cannot verify authenticity if any future packet once any portion of data is lost in transit. He Jin [19] approach uses the hash tree for decreasing receiver's computation overhead and authenticity because one root hash has the all value of leaf hash. Hash chaining is used for decreasing communication overhead and signing. It has the very less computation overhead because no need to compute more than one time at receiver side to verify the authenticity. It has some more communication overhead.

Adaptive Multicast Source Authentication (AMSA) [20] provides the mechanism to authenticate the source in multicast environment efficiently. This approach is like the tree approach where authentication information has sent with digest value from root of tree to leaf to all receivers where root digest is signed by source only one time in one block.

If we are increased security level, hashes may lead to overloading and more packet loss. So, more attack probability leads [8, 14] to more congestion, more packet loss provides more security threats. The situation will be worst. In next section, we are providing our proposed mechanism to tackle such type of situation.

3 Proposed Work

The global approach leads to higher level of overhead and affect congestion adversely. Thus, it forces the receiver to go for either leaving, deaf or join operation more frequently and deteriorating the situation more and more overhead. In localized storage approach decision about amount of information to be preserved is based on the availability of information with its successor multicast group in multicast hierarchy. At the local level the highest layer subscriber store the reference authentication among the group. The subscribing node retains the authentic information with intimation predecessor node (multicast node). For example, according figure 3 the group manager G provides availability of layer to subscribing nodes R_1, R_2 and R_3 . Here R_3 subscribe the base layer L_0 whereas layers L_0, L_1 and layers L_0, L_1, L_2 are subscribed by R_2 and R_1 respectively.

Further, information stored at group managers lying at same level is also, local and its global one is store in its predecessor. Suppose receiver R_2 suffers from packet loss and switch to deaf or leave operation for layer L_2 (higher subscribed layer by receiver R_2). Later on either it resumes from deaf or join layer L_2 and requires authentic information, For authentic information it contact to group manager.

Table 1. Authentic Information of Layer*s* for one Block

Layer	Block ID	Bundle ID	Packet in Bundle	Hash Vales
L_0	B_0	BUN0	P_1	h_2, h_{34}, h_{58}
L_0	B_0	BUN1	P_3, P_4	h_1, h_4, h_{58}
L_0	B_0	BUN2	P_4, P_5, P_6, P_7	h_3, h_{12}, h_8
L_0	B_0	BUN3	P_8	h_7, h_{14}, h_{56}
L_1	B_0	BUN0	P_1	h_2, h_{34}, h_{58}
L_1	B_0	BUN1	P_3, P_4	h_1, h_4, h_{58}
L_1	B_0	BUN2	P_4, P_5, P_6, P_7	h_3, h_{12}, h_8
L_1	B_0	BUN3	P_8	h_7, h_{14}, h_{56}
L_2	B_0	BUN0	P_1	h_2, h_{34}, h_{58}
L_2	B_0	BUN1	P_3, P_4	h_1, h_4, h_{58}
L_2	B_0	BUN2	P_4, P_5, P_6, P_7	h_3, h_{12}, h_8
L_2	B_0	BUN3	P_8	h_7, h_{14}, h_{56}

The authentic information preserved on R_1 to R_3 are termed as local one whereas global information is preserved at group manager store the authentic information for block B_0.

Table 2. Terms Used

Terms	Explanation
G	Group manager
LG	Local Group
RR	Requesting Receiver
n_i	Number of layers available
n_r	Number of receiver in group
n_o	Number of level in hierarchical multicast architecture to access the authentication information between path RR to GM or S
M	Number of hashes in one block
SL (R_i)	Subscribe layer by receiver (R_i)
Max_SL (R_i)	Maximum subscribe layer by receiver (R_i)
Comp_MSL (G)	Computed maximum subscribe layers by receiver (R_i) in group G
ADD_R	Address of receiver
CO	Communication overhead (time take by RR to node) between RR to node (which have stored the authentic information) to access the reference authentic information
N_AI	No. of Authentic Information stored at node for one block

The group manager intimate R_2 that requires information is available with receiver R_1 of the multicast group of R_2. Thus, R_2 get authentic information within local group.In this case required information stored at group manager is for layer L_2

which is highest available as well as subscribed layer available at this group manager. The detailed required information to be stored for layer L_0, L_1 and L_2 for one block (first block) are given in table 1. The information managed at group manager is for a block at time. That is block B_0 information are replaced by block B_1 and so on. The proposed approach is summarized in form of algorithm 1.

The effectiveness of proposed localized based authentic information can be used in the example shown in figure 3. Here, topology has seven routers (RT1, RT2, RT3, RT4, RT5, RT6 and RT7) and there end receivers are connected to end router (G) which is RT7. The topology is considered as hierarchical architecture. The table 3 is showing storage of authentication information of one block in global and local approach when source is sending packets with authentic information through path S->RT1->RT2->RT7->R1 or R2 or R3. It can be observed from table 3 that local approach required less authentication information storage then global approach.

Algorithm 1. *Localized based authentic information preservation*

```
1. For i= 1 to $n_i$
2. For i= 1 to $n_r$
          Max_SL( $R_i$) = 0  // Initially no layer is subscribed
          Comp_MSL(G) = Max( SL( $R_1$), SL( $R_2$),.................. SL( $R_{nr}$) )
3. For ( I = 1 to $n_r$ )
          While ( SL ( $R_i$ ) $\leq$ $n_i$)
           DO (operation)
        Case: Operation =Join
                    Contact to G for authentic information
               If (required information is available with group)
                   i.   Provide it to RR
               Else
                   i.   Provide ADD_R in LG of RR where information
                        is available
                   ii.  Receiver gets information from local receiver
                   iii. SL ( $R_i$) = {New SL } U {Already SL}
                   iv.  Comp_MSL(G) = Max (SL( $R_1$),
                        SL( $R_2$).............. SL( $R_{nr}$)
                   v.   remove authentic information for layer
                        whose number is less than Max_SL (G)
                break;
        Case: Operation =Leave
                i.  SL ( $R_i$) = {Already SL } - {Leave Layer}
                ii. Comp_MSL (G) = max ( SL ( $R_1$), SL( $R_2$),..............
                    SL( $R_{nr}$))
               iii. add authentic information to the group manager
               break;
        Case: Operation = Deaf_ resumption
```

```
                  i.   Contact group manager for authentic information
               If (required information is available with group)
                  i.   Provide it to RR
             Else
                  i.   Provide ADD_R in LG of RR where information
                       is available
                  ii.  Receiver gets information from local receiver
   Break;
```

The significant meanings of symbols are given in table 2. Here, in global approach one block information of one layer preserve at source is 12 (3+3+3+3=12) hashes, so for 3 layer it store 36 (12*3=36) hashes at source.

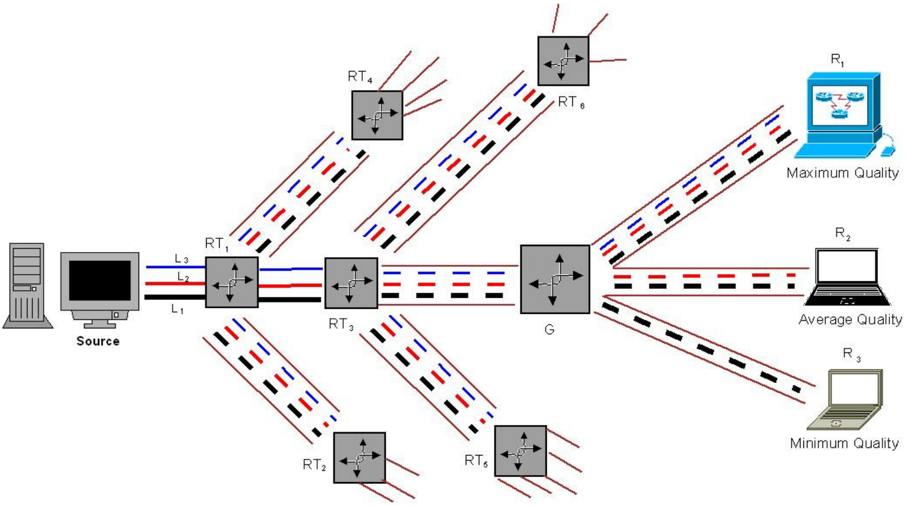


Fig. 3. Network Topology

On the other hand,each receiver store the one reference value of each layer to verify the packets i.e. 3 layer store the authentic information at all receiver which is 9 (3*3=9) hashes. Thus, total required store authentication information at all receivers are 45 (36+9=45) hashes. In other case of local approach the required authentication information for one block is stored at local group manager (only highest layer authentic information for one block i.e. 12 hashes) and maximum subscribe receiver (store all subscribe layer authentic information except highest layer i.e. 12*2=24) while receivers R_2, R_3 store the one referance authentic information for each layer (3*2=6) and R_1 store only one reference authentic information of highest layer.

Thus, total required stored authentic information at group and all receivers are 43 (12+24+1+6=43) hashes. However, storage of authentic information of localized based approach less than global based approach i.e. 2 hashes (45-43=2). For example number of layer (n_i) is 4 and number of receiver is 100 while maximum subscribe layer 4.

Table 3. Storage at node according to layer

Global Approach			Local approach		
Node	**Layer**	**N_AI**	**Node**	**Layer**	**N_AI**
S	L_0	12	S	L_0	0
S	L_1	12	S	L_1	0
S	L_2	12	S	L_2	0
RT_7	L_0	0	RT_7	L_0	0
RT_7	L_1	0	RT_7	L_1	0
RT_7	L_2	0	RT_7	L_2	12
R_1	L_0	1	R_1	L_0	12
R_1	L_1	1	R_1	L_1	12
R_1	L_2	1	R_1	L_2	1
R_2	L_0	1	R_2	L_0	1
R_2	L_1	1	R_2	L_1	1
R_2	L_2	1	R_2	L_2	1
R_3	L_0	1	R_3	L_0	1
R_3	L_1	1	R_3	L_1	1
R_3	L_2	1	R_3	L_2	1

In other case, for example shown in figure 3 where source is sending packets with authentic information through path S->RT1->RT2->RT7->R1 or R2 or R3. Suppose, communication overhead of nodes (request time to reach one node to other node) is equally distributed i.e 10 µs then communication overhead of request receiver (RR) to destination are illustrated by table 4. Here, in global approach for joining or deaf operation receivers send the request to source for access the reference authentic information of one layer, so these take 40 µs (10*4=40) communication overhead.

In other hand each receiver send the request to access the authentic information to a group manager if available or it provides the address of the receiver in the local group, so receiver take 10 µs (in best case) or 20 µs (in worst case) communication overhead. The communication overhead of requesting receiver (RR) is same bath approach i.e. 10 µs because receiver send to leaving request to group manager. It can be observed from table 4 that in deaf or join operation GBA communication overhead is 20 µs (40-20=20) more than LBA in worst case while LBA communication overhead is 30 µs (40-10=30) less than GBA in best case. Thus, receiver can access the reference authentic information while it performs join operation, deaf operation. Local based approach provide better performance than global based approach.

The communication overhead $\mathrm{CO_{iro}} = \sum_{r=1}^{r=Rn} n_r \sum_{o=1}^{o=RT_N} n_o * \mathrm{CO_{111}}$ where $\mathrm{CO_{111}}$ is communication overhead of one receiver to communicate first level node for only one layer authentication information and description of n_i, n_r, n_o are given in table 2.

At this cost (communication overhead) receiver access the authentic information in network overload situation and it verify the genuinity of source while performing the overload management operation.

Over the above provided secured information available irrespective of switching a receiver into deaf or leaving a layer on the occurrence of congestion. Up to now we have considered that intensity of attack is same all the time. However, in case intensity of attacker varies with time more prompt hash technique is required to apply.

Table 4. Communication overhead between RR to node

Global Approach				**Local Approach**				
Operation	**RR Node**	**Layer**	**CO (µs) (Worst)**	**Operation**	**RR Node**	**Layer**	**CO (µs) (Best)**	**CO (µs) (Worst)**
Join	R_1	L_0	40	Join	R_1	L_0	10	20
	R_1	L_1	40		R_1	L_1	10	20
	R_1	L_2	40		R_1	L_2	10	20
	R_2	L_0	40		R_2	L_0	10	20
	R_2	L_1	40		R_2	L_1	10	20
	R_3	L_0	40		R_3	L_0	10	20
Leave	R_1	L_0	10	Leave	R_1	L_0	10	20
	R_1	L_1	10		R_1	L_1	10	20
	R_1	L_2	10		R_1	L_2	10	20
	R_2	L_0	10		R_2	L_0	10	20
	R_2	L_1	10		R_2	L_1	10	20
	R_3	L_0	10		R_3	L_0	10	20
Deaf	R_1	L_2	40	Deaf	R_1	L_0	10	20
	R_2	L_1	40		R_2	L_1	10	20
	R_3	L_0	40		R_3	L_2	10	20

4 Results and Discussion

In this section simulation has been carried out to evaluate the performance of the proposed global and local level approach. The key parameters for performance measurement are stored authentic information, computation time, verification time, authentic packet ratio and throughput.

The effect of variation of block size, number of receivers, number of layers, deaf duration etc. are over these key parameters. The legends used in this section are listed in the table 5. The next subsection briefs about experimental setup used.

Table 5. Storage at node according to layer

Legends	Expanded form of legends	Description of legends
GBA	Global Based Approach	Receiver access the reference authenticates information from global level.
LBA	Local Based Approach	Receiver accesses the reference authenticate information from global level.

4.1 Experimental Setup Used

The simulation experiment has been carried out on Intel Core 2 dual processor 2.0 GHz, 3.0 GB RAM, 80 GB HDD machine supports with network simulation version 3.0 under Linux operating system. In this simulation topology the key component are sender (where message has been originated) and end router where multiple receivers are connected multicast. The roll of the intermediate router is to perform more routing decision and provide the authentic information to successor node. End router maintain multicast group and provide the authentic information a global as well as local level where as receivers stores regarding authentic information, computes the hashes and verify the genuinity.

On the others hand source compute hashes, make a bundle from packet and send it end router, from it is delivered to the multicast receivers. We have implemented the example figure 2 as simplest topology which is simplified form of multicast system in figure 3. It gives routers, source (sender) and multicast receivers. The network is heterogeneous in term bandwidth uniformly distributed in range (10-100) MBPS. The buffer used at each receiver is 100KB. The other simulation parameters are listed in table 6.

Table 6. Simulation Parameters

Parameter	Value used (fixed) (range)
Packet Size (Byte)	(256)(64, 128, 256,512,1024)
Hash Size (Byte)	(20)(16, 20,24,32)
Signature Size (Byte)	(128)(-)
Block Size (No. Packets)	(8)(2, 4, 8, 16,32)
Rate (Packet/Sec)	(10)(-)
Queue Size (No. Packet)	(100)(-)
Threshold (THARS)	5
Bundle size	(8)(1,2, 4, 8, 16,1)
Network bandwidth(MB)	(10-100)(-)
Link delay (ms)	(10-50)(-)

These values are same used in 11][12][13][20][21][22][23][25][28][30][31][36] [37] [38][40][42][48][44]. The next subsection deals with simulation results and it analysis.

4.2 Results and Analysis

The effect of variation of block size, number of receivers, number of layers, deaf duration etc. are over stored authentic information, computation time, verification time, authentic packet ratio and throughput. First we analysis the effect of variation in packet size followed by number of receivers.

Effect of Variation in Number of Level in Multicast Architecture

The effect of variation in number of level in multicast architecture on the joining overhead can be seen from figure 4. It has been observed that with increment in number of level in architecture then joining overhead increases and applicable for each one. This is because more the number of levels in architecture required more time to access this authentic information. The increment of local based approach is less than global one because in local based approach received the authentic information form group manager or neighbor receiver.

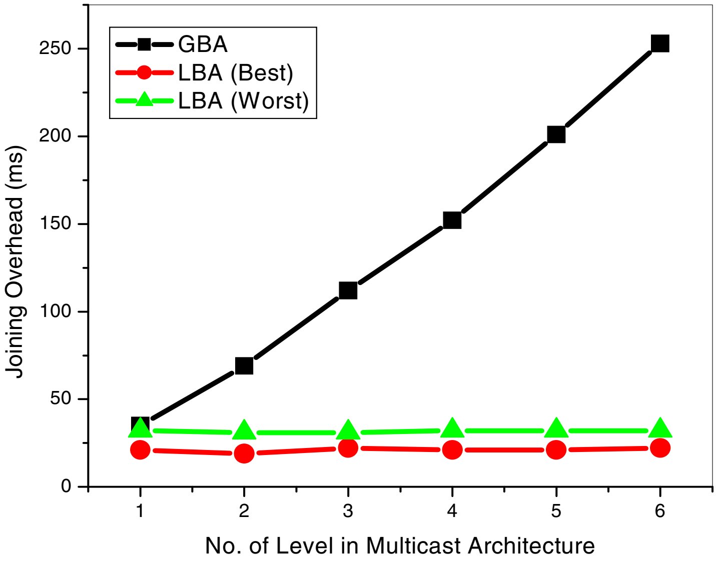


Fig. 4. Joining Overhead w.r.to No. of Level in Architecture

Effect of Variation in Number of Layers

The effect of variation in number of layer on the joining overhead can be seen from figure 5. It has been observed from figure 3 that with increment in number of layers increases the joining overhead which is applicable GBA, LBA (worst) and LBA (best)

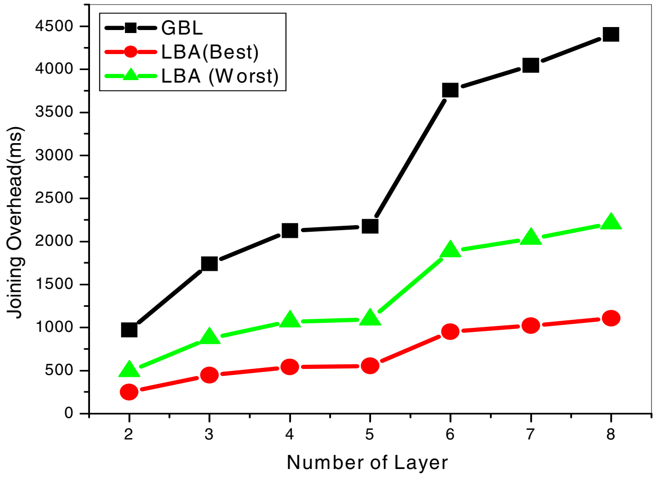


Fig. 5. Joining Overhead w.r.to No. of layer

approach. The reason behind increment in joining overhead is that receiver access more authentic information which take more time to access these authentic information.

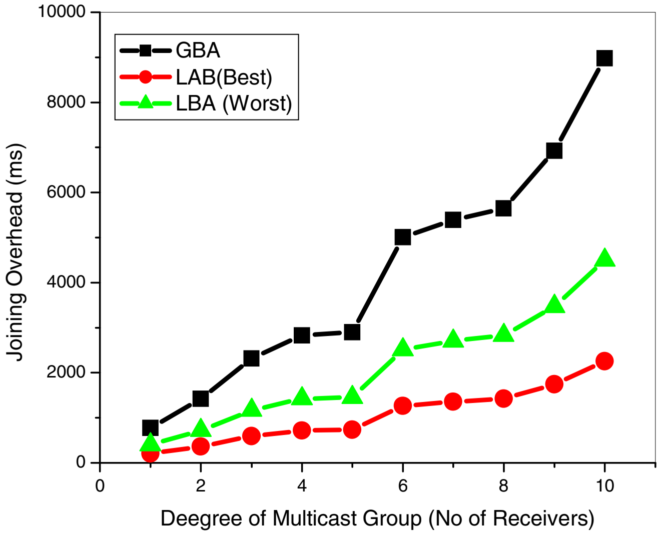


Fig. 6. Joining Overhead w.r.to degree of Multicast Group

In other word, it can be seen that increment in joining overhead of local based (LBA) approach is less than global (GBA) based approach. This is due to local based approach received the authentic information from local and it will take less time to access these information.

Effect of Variation in Number of Receivers

The effects of variation in number of receiver affect the joining overhead which can be seen from figure 6. It can be observed that joining overhead increases with increment in number of receivers in a group and this is applicable for all approach. However, lesser increment is observed for local approach as compared to global one.

5 Conclusion

In this paper, we proposed integrated security aware congestion control approach to improve the security of multicast system in presence of security threats. The proposed approaches provide the authentic information in layered multicast architecture for source authentication in presence of network overload. For this, we have proposed global based and local based approach. The aim of proposed work is to increase throughput and reduce the overhead to access the authentic information. The proposed global based approach (GBA) stores the authentic information at source end. When network is overloaded then receiver performs the deaf/leaving operation then authentic information of next successor packets is also lost. Due to this loss of authentic information, receiver is unable to verify the genuinity of source. So, the receiver receives authentic information from source at the cost of increased overhead. In local based approach (LBA) the group manager stores the authentic information stored at predecessor node and neighbor receiver. In case of overload, receiver will access the authentic information from predecessor node (best case) or neighbor receiver (worst case). The authentic information from layered multicast architecture is received when applying the overload management mechanism, the parameters such as level of architecture, number of layer and number of receiver which effect the joining overhead. The simulation results show that the joining overhead is less in LBA than GBA. The effectiveness of the proposed algorithm has been discussed through examples and extensive simulation results. The proposed security aware multicast congestion control approach increases the security and reduces the overhead in presence of security threats and network overload.

References

1. Yin, D.S., Liu, Y.H., et al.: A new TCPfriendly congestion control protocol for layered multicast. In: Proc. IASTED Conference on Internet and Multimedia Systems and Applications, Innsbruck, Austria (February 2006)
2. Byers, J., Frumin, M., et al.: FLID-DL: congestion control for layered multicast. In: Proc. NGC 2000, Palo Alto, USA, pp. 71–81 (November 2000)
3. Kulatunga, Fairhurst: TFMCC Protocol Behaviour in Satellite Multicast with Variable Return Path Delays. IEEE (2006)
4. McCanne, S., Jacobson, V., Vetterli, M.: Receiver-driven layered multicast. In: Proceedings of ACM SIGCOMM, New York, USA, pp. 117–130 (August 1996)
5. Johansen, S., Kim, A.N., Perkis, A.: Quality Incentive Assisted Congestion Control for Receiver-Driven Multicast. IEEE Communications Society ICC 2007 (2007)

6. Kammoun, W., Youssef, H.: An adaptive Mechanism for End-to-End Multirate Multicast Congestion Control. In: Proceeding of The Third International Conference on Digital Telecommunications, pp. 88–93 (2008)
7. Li, B., Liu, J.: Multirate video Multicast over the Internet: An Overview. IEEE Network (January/February 2003)
8. Bruhadeshwar, B., Kulkarni, S.S.: Balancing Revocation and Storage Trade-offs in Secure Group Communication. IEEE Trans. on Dependable And Secure Computing 8(1), 58–73 (2011)
9. Rizzo, L.: A TCP-friendly single-rate multicast congestion control scheme. In: Proc. ACM SIGCOMM, Stockholm, Sweden, pp. 17–28 (August 2000)
10. Floyd, S., Handley, M., Padhye, J., Widmer, J.: Equation based congestion control for unicast applications. In: Proc. ACM SIGCOMM, Stockholm, Sweden, pp. 43–56 (August 2000)
11. Singh, K., Yadav, R.S.: Efficient Joining and Leaving for Receiver Driven Multicast Congestion Contro. International Journal of Computer Applications 1(26), 110–116 (2010)
12. Singh, K., Yadav, R.S.: Overview of secure multicast Congestion Control. In: International Conference on Soft Computing and Intelligent Systems (ICSCIS 2007), Jabalpur (December 2007)
13. McCanne, S., Jacobson, V., Vetterli, M.: Receiver-driven Layered Multicast. In: Proceedings of ACM SIGCOMM (August 1996)
14. Athens/Glyfada, Greece, Replay Attack of Dynamic Rights within an Authorised Domain. In: Proc. of IEEE, Third International Conference on Emerging Security Information, Systems and Technologies (2009)
15. RFC 4046, Multicast Security (MSEC) Group Key Management Architecture (April 2005)
16. RFC-3740, The Multicast Group Security Architecture (March 2004)
17. Mokhtarian, K., Hefeeda, M.: Authentication of Scalable Video Streams With Low Communication Overhead. IEEE Trans. on Multimedia 12(7), 730–742 (2010)
18. Gorinsky, S., Jain, S., Vin, H., Yongguang: Design of Multicast Protocols Robust Against Inflated Subscription. IEEE/ACM Transactions on Networking 14(2) (April 2006)
19. He, J.-X., Xu, G.-C., Fu, X.-D., Zhou, Z.-G.: A Hybrid and Efficient Scheme of Multicast Source Authentication. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. IEEE (2007)
20. Singh, K., Yadav, R.S., Sharma, A.K.: Adaptive Multicast Source Authentication. In: IEEE Proceeding of International Advance Computing Conference, IACC 2009, March 6-7 (2009)
21. Perrig, et al.: Efficient Authentication and Signing of Multicast Streams over Lossy Channels. In: IEEE Symp. Security and Privacy 2000 (2000)
22. Singh, K., Yadav, R.S.: Multicast Congestion Control in Adversary Environment. In: IPCSIT, vol. 31. IACSIT Press, Singapore (2012)

Network Selection for Remote Healthcare Systems through Mapping between Clinical and Network Parameter

Rajeev Agrawal and Amit Sehgal

G.L. Bajaj Institute of Technology and Management, Plot No. 2, Knowledge Park – III,
Greater Noida, UP, India-201306
{rajkecd,amitsehgal26}@gmail.com

Abstract. The paper presents fuzzy based approach to select best suited network for remote healthcare services. A direct mapping between clinical parameters and corresponding QoS network parameters is done. The paper proposes an application independent integrated system where stage of the disease is identified based on the fuzzified clinical parametric values which are critical for that disease. Based on critical nature of the disease i.e. stage of disease, the requirements of network QoS are defined in linguistic terms. This eliminates strict sense condition on any specific network and selects most suited network out of all available networks. The objective is to avoid denial of service in case of unavailability of a network with high QoS values and also to conserve resources in case the patient is in normal condition medically. A simulation study has been presented to verify the selection of the network based on stage of disease.

Keywords: Index Terms: heterogeneous networks, network selection, remote healthcare, fuzzy logic, QoS.

1 Introduction

Availability of healthcare services to the population in every corner has undoubtedly become one of the priorities for Health Organizations. This requirement is more challenging for the Low Income Countries (LIC) and Lower Middle income countries (LMIC) [1]. Population in these countries is predominantly rural and distributed across distant geographically diverse locations in countries like India. Despite several advances and developments in the medical field, its access is restricted to some privileged section of the society residing in urban areas of falling under high income group. The hierarchical structure of healthcare facilities further restricts the access to specialty health services to the rural population.

The fast evolving information and communication technology supported by increasing use of mobile devices has resulted in the development of e-health and m-health services. Several such remote healthcare systems have been developed and tested [2 – 4]. However, these systems suffer from limitation of their dependence on a particular type of network. Their use is therefore restricted to the areas falling under

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 31–41, 2013.

the coverage zone of that network. Further, demand for high level of Quality of Service by these mission critical services still restricts its use in the remote areas. The concept of integration of technologies and user level customization introduced by 4G systems will provide a solution to these problems. The objective is to provide remote healthcare service irrespective of the type of communication link available. Specifying network requirements and selecting optimum network remains an open issue for such systems.

In this paper, we present a fuzzy based approach to generate network requirements in terms of QoS parameters based on the physiological data generated by data acquisition devices i.e. body sensors. There are sensor based systems available which acquire clinical parameters from human body and transmit them through a compatible communication system [5]. However, in the era of heterogeneous/4G systems, there are reconfigurable and customizable devices with ability to communication through more than one type of networks or communication links [6, 7]. Fig. 1 presents a remote healthcare system based on such devices and body sensors. Selecting optimum network out of several available network is still an open issue and under research. We intend to use such devices for the mission critical application of remote healthcare services. Therefore, a direct mapping between clinical parameters and network QoS parameters for network selection has been presented in this paper. To further keep the approach generalized and applicable to several diverse medical situations, fuzzification of clinical parameters has been done before it can be used for generating network requirements.

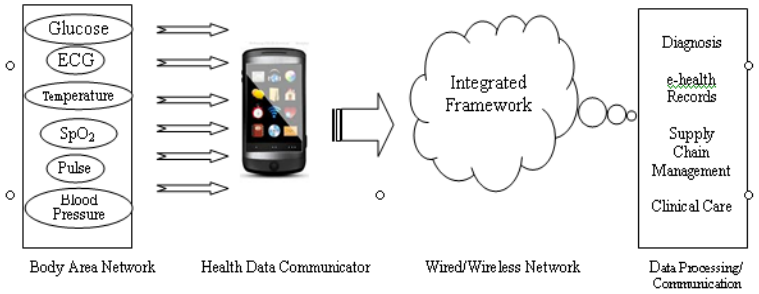


Fig. 1. Remote healthcare system based on mobile devices communicating through heterogeneous networks. Body sensors monitoring various clinical parameters provide necessary information to frame selection criterion at the communicator device.

Rest of the paper is organized as follows. Section II presents fuzzy functions for some of the commonly used clinical parameters. A mapping scheme between fuzzy clinical and network QoS parameters is introduced in section III. A clinical case study has been taken for this purpose. Section IV concludes the paper with summarizing contribution of the work and its future scope.

2 Fuzzy Functions for Clinical Parameters

Since our focus is on healthcare services the paper proposes a network selection scheme considering QoS parameters required for the transmission of physiological information being monitored. The device needs to process the data received from the body sensors and suggest the condition of the patient and decide the seriousness of the application. Based on this decision, QoS requirements are generated to select the optimum network to transmit this data. We propose a fuzzy logic system to process the patient's physiological parameters. In spite of this, human body is so complex that it would be impossible to cover every single measurable body parameter. Some of the commonly used vital signs of human body have been considered in this paper. Table 1 presents a list of such vital signs along with their normal range values. The range of values representing different medical conditions used in this paper is based on inputs received from various medical practitioners whose identities have not been disclosed here. For each of these parameters, fuzzy functions F(parameter) are defined in linguistic terms.

Table 1. Body parameters being monitored alongwith their prescribed normal range of values

Parameter	Normal range
Body Temperature	36.5 – 37.5 ^{o}C
Blood Pressure (systolic)	120 mmHg
Blood Pressure (diastolic)	80 mmHg
Blood Sugar	100 mg/dL
Blood Oxygen Level	85 – 100 mmHg
Oxygen Saturation	97 – 100 %
Heart or Pulse Rate	72 (daily average)

Body Temperature. Most of the medical ailments in human body result in variation of the body temperature. It is, therefore a most commonly examined physiological parameter. In our proposed work, we have defined this vital sign using fuzzy logic having five membership functions. Temperature range for hypothermia and hyperthermia has been considered to select the lower and upper limit.

F(temperature) = {LowCritical, LowMarginl, Normal, HighMarginal, High Critical}

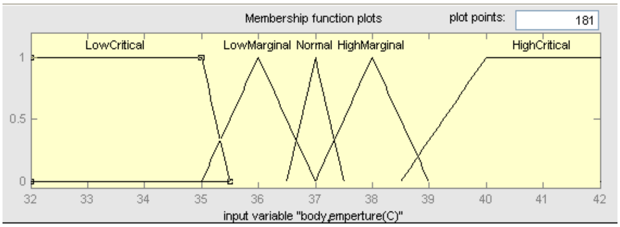


Fig. 2. Membership functions for body temperature

Heart Rate or Pulse Rate. A daily average value of 72 beats per minute (bpm) is considered to be normal for an average human body. Neglecting any irregular activity, the heart beats at a rate 30% faster during day time as compared to night while sleeping. This makes overall range for the entire day to 55 – 100 bpm. It is to be noted that fuzzy equivalent used for this parameter in this paper is for the average values.

F(pulse) = {LowCritical, LowMarginl, Normal, HighMarginal, High Critical}

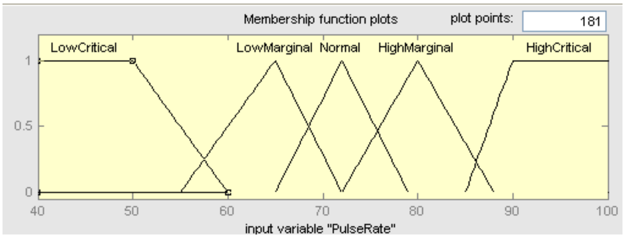


Fig. 3. Membership functions for heart rate or pulse rate

Many more such representations can be obtained based on different conditions of the patient such as during physical activity, during rest and so on. A set of If – Then statements can be used to select the appropriate condition among the set of available conditions.

Blood Pressure (Systolic and Diastolic): It is a parameter which, compared to the above two parameters, shows irregular pattern if observed for entire day. The mean blood pressure for an average healthy adult human being is 120 mmHg for systolic and 80 mmHg for diastolic but varies 20mmHg for systolic and 10 mmHg for diastolic above and below during different activities throughout the day. The fuzzy equivalent of these two parameters involves five membership functions each.

F(BP-sys) = {LowCritical, LowMarginl, Normal, HighMarginal, High Critical}

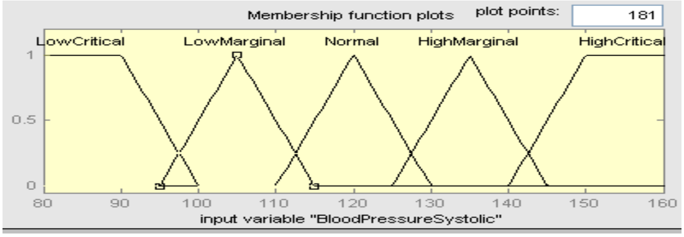


Fig. 4. Membership functions for blood pressure – systolic

F(BP-dia) = {LowCritical, LowMarginl, Normal, HighMarginal, High Critical}

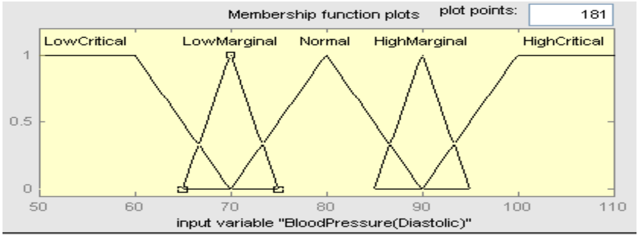


Fig. 5. Membership functions for blood pressure -diastolic

Blood Sugar: Blood sugar or blood glucose level in human blood is measured as mg/deciliter or mill moles per liter. 1 mmol/L is equivalent to 18 mg/dL. Its normal value is 100 mg/dL when fasting and can go up to 140 within one hour of having food. For diabetic patients, it can be as high as 400 or above and can drop down to 70 or less. High or low blood sugar level can often lead to various complications like coma or death. A five level fuzzy function has been used for this parameter also.

F(sugar) = {LowCritical, LowMarginl, Normal,HighMarginal, High Critical}

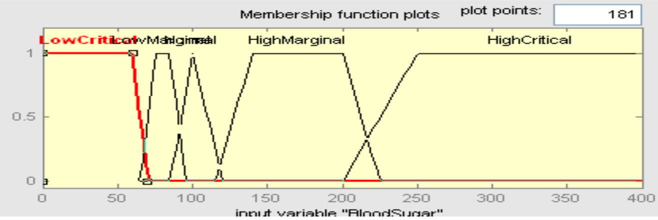


Fig. 6. Membership functions for blood sugar level

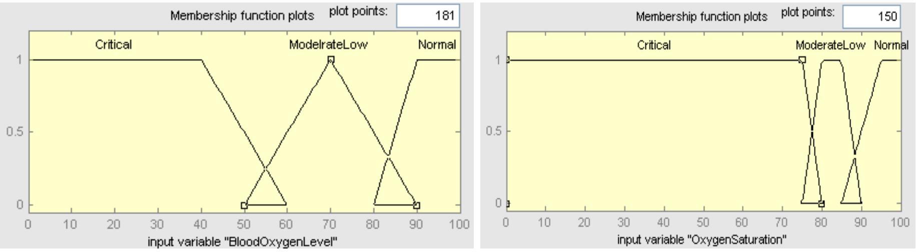


Fig. 7. Membership functions for blood oxygen level and oxygen saturation

Blood Oxygen Saturation. Ability of the lungs to supply oxygen to the blood is measured through the blood oxygen level. It is used to evaluate the oxygenation and saturation of haemoglobin in the blood. There are three parameters involved – (a) partial oxygen pressure (mmHg) in arterial blood (PaO_2 in Pulmonary Capillary or PAO_2 in Alveolus, differing by 10 mmHg), (b) direct measurement of the percentage of blood oxygen saturation level (SaO_2) and (c) indirect measurement of the percentage of blood oxygen saturation level (SpO_2). First term refers to dissolved oxygen and the normal value is 90 – 100 mmHg. A level below 80 mmHg is the beginning of moderate hypoxia and its drop to 40 results in critical hypoxia. The second and third terms are for oxygen bound to hemoglobin and the required level of saturation is 97 – 98 %. Decrease in the level of all these parameters represents medical ailment.

Two fuzzy variables namely blood oxygen level and oxygen saturation have been defined each having three membership functions as shown in fig. 7.

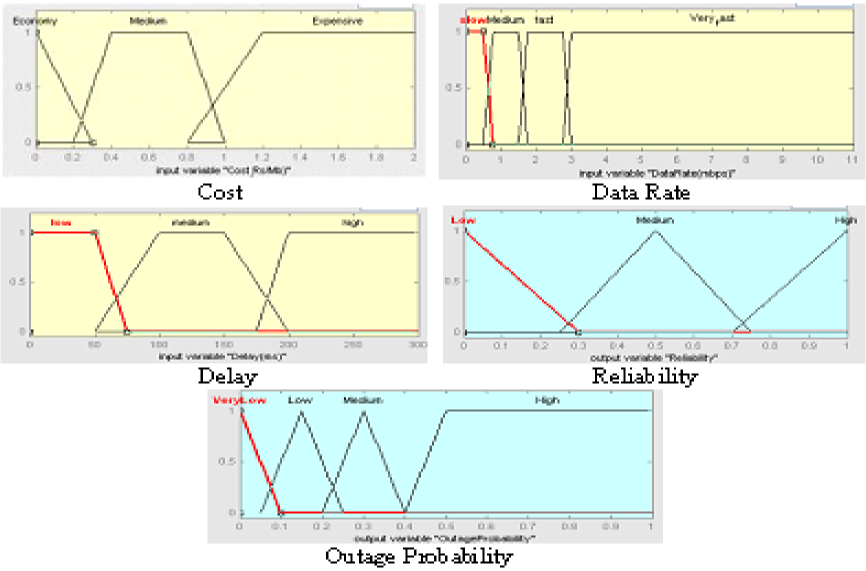
F(blood-oxy) = {Critical, ModerateLow, Normal}
F(lood-sat) = {Critical, ModerateLow, Normal}

3 Mapping Clinical Parameters with Network Parameters

The clinical data acquired through body sensors is used for several healthcare applications e.g. diagnosis and post hospital monitoring. There are two fold objectives of this remote monitoring system.

1. To keep a regular/time based record of the patient's condition without getting admitted to a hospital or healthcare centre.
2. To generate alert signal when one or more clinical parameters crosses the predefined upper or lower threshold limit.

The objective of this paper is to propose a monitoring system which does not depend on any specific communication technology but works with the most suitable of all the available links. We, therefore, make some elementary clinical inference about state of



Cost

Data Rate

Delay

Reliability

Outage Probability

Fig. 8. Fuzzy functions for QoS parameters in linguistic terms

the patient being monitored to generate QoS requirements for efficient transmission of the sensor's output data. Depending on the disease for which the patient is being monitored, a set of fuzzy rules are written which give a set of QoS parameters as output variables. Five such network QoS parameters have been used in this paper viz. Cost of network usage, Data Rate, End-to-End Delay, Reliability (in terms of Packet Loss and Bit Error Rate-BER) and Outage Probability. Since the values of these parameters for any available networks may depend on a large number of factors and thus vary from place to place, it is not justifiable to define the QoS requirements in the form of crisp values. We, therefore generate QoS requirements in linguistic terms which are then represented by fuzzy variables as shown in fig. 8.

Corresponding fuzzy functions G(parameter) are defined as:

- G(cost) = {Economy, Medium, Expensive}
- G(data-rate) = {Slow, Medium, Fast, VeryFast}
- G(delay) = {Low, Medium, High}
- G(reliability) = {Low, Medium, High}
- G(outage-prob) = {VeryLow, Low, Medium, High}

3.1 Clinical Case Study - Diabeters

Diabetes, a common and globally widespread disease, has been taken as a clinical case to study the proposed network selection scheme. Six clinical parameters introduced in section II are generally monitored for diabetic patients. A fuzzy rule base is defined taking clinical parameters as input variables and network parameters as output variables. Based on deviation in the observed values of these parameters from their normal values, we define four different stages of the patient viz. Normal (Nm), Marginally High (MH), Very High (VH) and Critical (Cr). Each of these stages generates a different set of QoS requirements in terms of network parameters defined in fuzzy terms earlier in this section. Table 2 gives the lookup table between four disease stages and corresponding QoS requirements. A total of 70 fuzzy rules were created. 15 of them represent Normal Stage. Marginally High and Very High stages are represented by 20 rules each. Critical stage results from outcome of 15 rules. The proposed scheme was tested on the pre-recorded data of 50 diabetic patients.

Fig. 9 present the direct fuzzy-to-fuzzy mapping system between clinical and network QoS parameters. Mamdani rule base model has been used to draw fuzzy

Table 2. QoS requirements based on clinical stage of the patient

QoS Parameters	**Fuzzy Network QoS Values for Different Disease Stages**			
	Normal	**Marginally High**	**Very High**	**Critical**
Cost	Economy	Medium	--	--
Data rate	Medium	Medium	Fast	VeryFast
Delay	High	Medium	Low	Low
Reliability	Medium	Medium	High	High
Outage Probability	High	Medium	Low	VeryLow

inferences. Functions F(parameter) defined for Clinical parameters have been used as input fuzzy functions. The output is available in the form of functions G(parameters) defined for network parameters.

Fig. 10 presents graphical view of QoS requirements derived by using proposed scheme on sample data. Samples have been taken so as to include all the four stages of disease defined in this case study. For simple representation, QoS levels have been shown by using numbers. Table 3 gives the linguistic terms corresponding to each level for all the five network parameters used as output fuzzy variables.

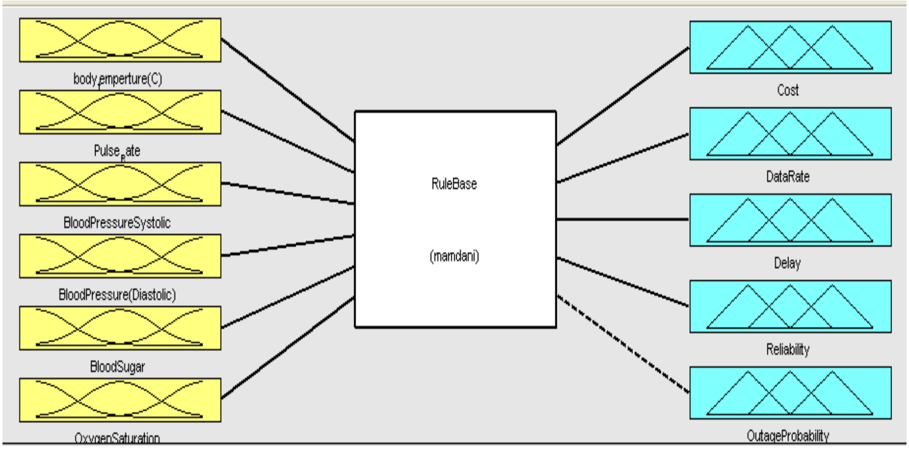


Fig. 9. Fuzzy based mapping system between clinical and network QoS parameters

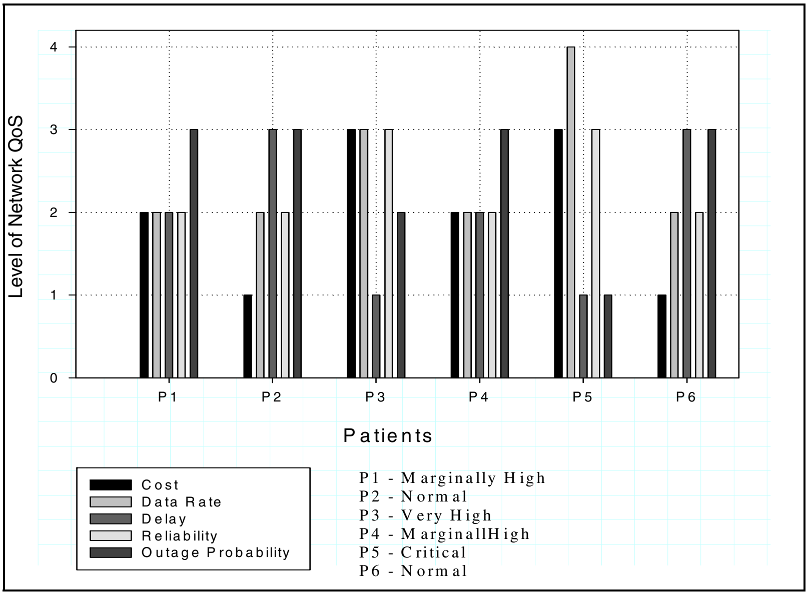


Fig. 10. Mapped QoS level requirements for patients from sample data

4 Testing the Network Selection

To verify the proposed scheme, a sample heterogeneous test environment with four different networks (shown in fig. 11) has been simulated. The QoS values for these test networks were adjusted according to the networks available in the author's research place. Four patient's nodes correspond to the health data communicators shown in fig. 1, each having four network interfaces compatible with four test networks.

Table 3. QoS levels in linguistic terms corresponding to numeric levels used in fig. 10

QoS Parameters	**Linguistic Terms for QoS Levels in Fig. 11**			
	Level 1	**Level 2**	**Level 3**	**Level 4**
Cost	Economy	Medium	Expensive	--
Data rate	Slow	Medium	Fast	Very Fast
Delay	Low	Medium	High	--
Reliability	Low	Medium	High	--
Outage Probability	Very Low	Low	Medium	High

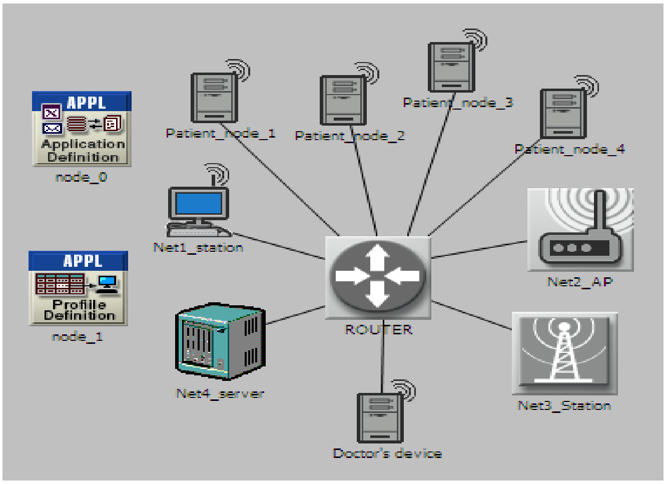


Fig. 11. Simulation setup with four test networks, four patient devices as source, and one medical practitioner's device as destination

Table 4. four patients with different disease stages and their corresponding selected networks as a result of simulation

Patient No.	Disease Stage	Selected Network
1	Marginally High	**N2**
2	Normal	N1
3	Marginally Hgh	N2
4	Very High	N3

Clinical parameter values for four different diabetic cases falling under disease stages – Normal, Marginally High and Very High were used to generate the network requirements in terms of QoS parameter (Data for Critical stages was not available and hence not simulated). A local medical practitioner was consulted for this purpose whose identity has been kept hidden on request. Network selection scheme introduced in [8] was used for this purpose. It was verified after simulation that patients falling under different stages of the disease used different networks to transmit the clinical data recorded by the body sensors. Fig. 12 and table 4 give the simulated results.

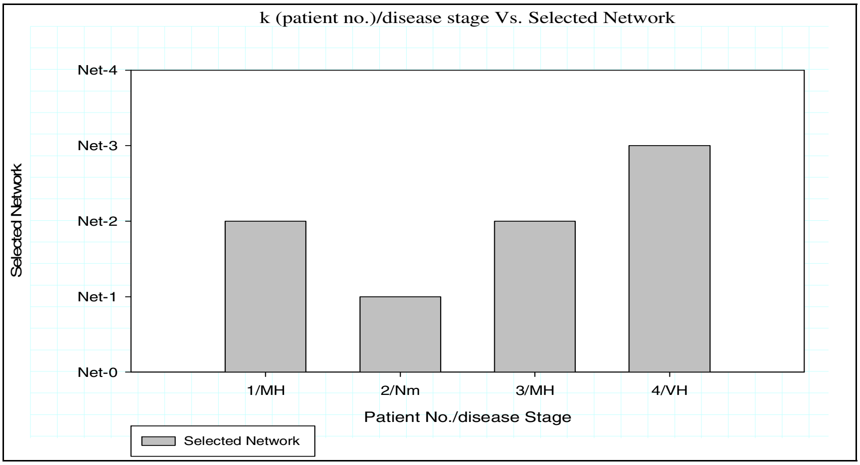


Fig. 12. Selected Network for patients at different disease stages

5 Conclusion

The Proposed work presents a fuzzy based approach to map the stage of patient's disease with the selection of an optimum network for remote monitoring services. The selection of right network will ensure the transfer of vital clinical parameters to the control centre irrespective of the location of the patient and availability of high performance network. The values of clinical parameters have been fuzzified and mapped into a set of linguistic network QoS levels as desired QoS requirements. These requirements can be used as input to any suitable network selection algorithm. The simulation study presents a roadmap towards a complete integrated remote healthcare system.

References

1. http://www.who.int/healthinfo/global_burden_disease/.../en/index.html
2. Lin, S.-S., Hung, M.-H., Tsai, C.-L., Chou, L.-P.: Development of an Ease-of-Use Remote Healthcare System Architecture Using RFID and Networking Technologies. Journal of Medical Systems (March 2012)

3. Nita, L., Cretu, M., Hariton, A.: System for remote patient monitoring and data collection with applicability on E-health applications. In: 7th International Symposium on Advanced Topics in Electrical Engineering, Bucharest, pp. 1–4 (May 2011)
4. Agrawal, R., Sneha, S., Sehgal, A.: QoS Based Adhoc Probabilistic Routing Strategy for e-health Services. International Journal of Scientific & Engineering Research 3(5), 1–5 (2012)
5. Jurik, A.D., Weaver, A.C.: Body Sensors: Wireless Access to Physiological Data. IEEE Software, 71–73 (January/February 2009)
6. Sehgal, A., Agrawal, R.: QoS Based Network Selection Scheme for 4G Systems. IEEE Transactions on Consumer Electronics 56(2), 560–565 (2010)
7. Martin, J., Amin, R., Eltawil, A., Hussien, A.: Using Reconfigurable Devices to Maximize Spectral Efficiency in Future Heterogeneous Wireless Systems. In: Proceedings of 20th International Conference on Computer Communications and Networks (ICCCN), pp. 1–8 (July-August 2011)

Performance Analysis of SMAC Protocol in Wireless Sensor Networks Using Network Simulator (Ns-2)

Gayatri Sakya[*] and Vidushi Sharma

JSS Academy of Technical Education, Noida
Gautam Buddha University, Greater Noida
gayatri.sakya@rediffmail.com, Vidushi@gbu.ac.in

Abstract. Energy effeciency of medium access control has been an active research area in wireless sensor networks since past few years. SMAC stands for Sensor-MAC protocol, which is designed on the basis of periodic listen-sleep mechanism of nodes for avoiding energy wastage because of idle listening. SMAC reduces energy consumptions because of collision, overhearing, control packet overhead and idle listening. This paper discusses the basic attribute of MAC protocols, their classification and the importance of SMAC protocol in wireless sensor networks. SMAC is developed primarily for Mote platform, and thereafter also implemented in Network Simulator-2.So without real hardware one can analyze the performance of SMAC under various application specific scenarios with NS-2.In this paper, the performance of SMAC protocol is analyzed under high and low traffic rates with different duty cycles in single hop scenario without the routing effect. The residual energy is also measured in each scenario. Since wireless sensor networks are application specific, so the behavior of SMAC is studied when the data transport performance and hence the throughput and jitter also plays an important role along with the energy effeciency. Finally it has been shown that under higher traffic loads, if the value of duty cycle is increased to optimum value, the residual energy of the node is improved with a better throughput.

Keywords: Wireless Sensor Networks, SMAC, residual energy, throughput, NS-2.

1 Introduction

Advancements in wireless information system, embedded systems and the VLSI technology has enabled the development of wireless sensor networks which consist of thousands of tiny sensor nodes deployed in Adhoc or structured preplanned manner. These [1] are low power devices equipped with one or more sensors, a processor, limited memory, a power supply, a radio, and an actuator. Since the sensor nodes have limited memory and are typically deployed in unattended locations, a transceiver is implemented for wireless communication to transfer the data to the base station.

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 42–51, 2013.

[2]Battery is the main source of power in a sensor node. Wireless sensor networks are used in range of applications such as monitoring, tracking and surveillance of borders, in industry for factory instrumentation, in metro cities to monitor traffic and road conditions, in engineering to monitor buildings structures, in environment to monitor forest, oceans etc. To serve these different applications of wireless sensor networks, the protocol stack has not been standardized yet, and the research is continued on each layer to design energy efficient protocols suitable for specific applications. The transceiver of sensor nodes consumes maximum amount of energy [2]. One solution to reduce energy consumption is to develop the energy efficient communication protocols. The[2] MAC protocol has been the research area since past few years because by designing a good MAC protocol, the energy effeciency of the nodes may be increased which is the prime concern in case of wireless sensor networks [1].

1.1 The MAC Protocol Design Issues

The MAC protocols for wireless Adhoc networks such as 802.11 DCF is not suitable for wireless sensor networks because it does not consider the nodes with limited battery[4].The good design of the MAC protocol should prevent energy wastage due to packet collisions, overhearing, excessive retransmissions, control overheads, and idle listening. It should also adapt to topology and network changes efficiently. Various MAC protocols with different objectives are proposed for wireless sensor networks. The main attribute of MAC protocol on which authors concentrated is the energy effeciency [6]. Other important attributes are scalability and adaptability to changes like network size, node density and topology. Secondary attributes are throughput, latency and per node fairness. The MAC layer defines two main criterions for designing efficient MAC protocol. First is to detect the reasons for energy waste and second is the communication pattern which the network follows since these patterns decides the behavior of traffic which the MAC protocol will handle[6]. A wide range of energy efficient MAC protocols for example [7][8][9][10] are proposed which are categorized according to channel accessing approaches ,into contention-based, TDMA-based, hybrid, and cross layer MAC protocols . In recent research [15][16][17] it is pointed out that wireless sensor networks when used in industrial control or monitoring applications then apart from energy and latency the throughput and packet delivery ratio should also be taken into consideration.

1.2 Sensor-MAC Protocol Design

Sensor-MAC (S-MAC), is a contention based Medium Access Control protocol for Wireless Sensor Networks proposed by SCADDS project group at USC/ISI[10] which is different from traditional IEEE 802.11DCF for Adhoc networks. Implementation of the protocol is done on Rene Motes, developed at UCB as test beds. Mote uses TinyOs as an efficient event driven operating system. For SMAC, energy efficiency and self-configuration ability are the primary goals, while others attributes, like latency and fairness, are secondary. While designing the protocol, it is assumed that the major sources of energy waste are collision, overhearing, control packet overhead and idle listening (keep on listening to receive possible traffic which is not sent). In MAC protocols such as IEEE 802.11 nodes must listen to the channel to receive possible traffic which consumes 50–100% of the energy required for receiving. But [3] S-MAC tries to reduce the waste of energy from all the above sources. SMAC consists of three major components: periodic listen and sleep, collision and overhearing avoidance, and message passing. . During sleep, the node turns off its radio, and sets a timer to awake it later. Periodic listen and sleep is used to avoid idle listening. Contention mechanism [5] is same as in IEEE 802.11 to avoid collision. Overhearing is avoided by letting interfering nodes go to sleep when they hear RTS/CTS packets, So that they may not hear the long data packet and the ACK.NAV value is used for this purpose. Message passing is used to reduce control overhead in contrast with 802.11.In message passing; long messages are broken into small fragments and transmitted as a burst. Extra delay is caused because of node periodic sleeping and is further improved by the technique of adaptive listening [5]. In the second section of this paper , we have discussed the implementation of SMAC in NS-2 and in third section the simulation experiments has been carried out in NS-2.35. In fourth section the analysis of results has been done and finally section fifth concluded the paper.

2 SMAC Model and Its Implementation in NS 2.35

S-MAC is based on the Mote platform that runs TinyOs operating system. For hardware implementation, which needs real hardware as its running platform, at times it is not easy to carry out the performance analysis for the experimentation purposes and hence S-MAC has been also implemented in NS2[11][12][13], the network simulator. SMAC has been tested on the real platform but little work has been done on the NS2 for further improving the SMAC for the mission critical applications in wireless sensor networks. Besides this, [9][10][14]researchers have studied SMAC performance for energy and latency. But for mission critical applications one needs to analyze the performance of S-MAC protocol in terms of residual energy of nodes, throughput , packet delivery fraction and the impact of different duty cycles to analyze the protocol fully. Frame structure of SMAC is given in Figure 1.The listen period is further divided into two parts. SYNC periods designed for SYNC packets, which are broadcast packets and solve synchronization problems between neighboring nodes.

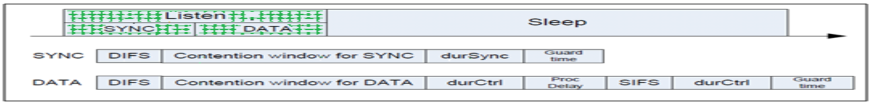


Fig. 1. Frame intervals (listen + sleep) [3]

Table 1. The default Energy Model for SMAC

Idle Power	1.0 watts
TxPower	1.0 watts
RxPower	1.0 watts
SleepPower	0.001watts
TransitionPower	0.2 watts
TransitionTime	0.005seconds

3 Simulation of SMAC Protocol in NS-2

Environment. Simulation experiment is performed by writing the TCL script in NS-2.35, installed in Cygwin environment. The results are obtained by analyzing the trace file obtained in new trace file format. The updated energy model has been used in this paper and different scenarios are considered to carry out simulation.

Topology. Five S-MAC nodes form a single hop star topology, with four sources and one central sink which is given in figure 2.The UDP agents are used at source nodes and they are attached to the CBR traffic.

Input Parameters. The packet inter-arrival time is varied from .01seconds (considered highest traffic load) to 50 seconds (lower traffic load). Duty cycle is varied from 1% to 50%. Syncflag is set to 1, S-MAC runs with periodic sleep. If it is set to 0, SMAC runs without periodic sleep. Duty cycle controls the length of sleep. Here we have set the SelfConfigFlag to 1. Packet size is set to 100Bytes. The initial Energy of nodes is set to 1000J. The simulation runs for 5000 seconds. Maximum numbers of packets are 1000. All four nodes start sending the packets after 40 seconds, which is required for synchronization of nodes.

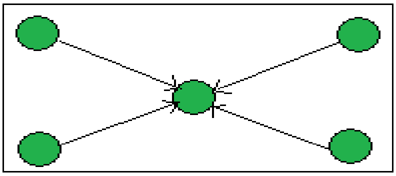


Fig. 2. Central sink node (Receiving) with four source (sending) nodes

Performance Parameters. Performance parameters which are considered in the paper are the Residual Energy and Throughput under different duty cycles and message inter arrival time.

- **Residual Energy:** In this analysis the residual energy is given by the amount of remaining energy of the node after simulation time. So the new trace file format using updated energy model gives us the residual energy after each event at each node.

- **Throughput:** We compute the throughput using the payloads received at MAC layer.

Throughput (Kbps) = Packets received at the sink node/ (T1-T2) ... (1)
T1= time when first packet sent by source nodes
T2= time when last packet received by sink node.

4 Simulation Results

The residual energy and the throughput are measured under different traffic loads with varying duty cycle.

- **High traffic scenarios**

Scenario 1 & Scenario2: In scenario1, the duty cycle was varied from 1% to 30% and the residual energy of sink node is observed for MIAT=.01s which is considered to be the highest traffic load. The variation is shown in Fig.3. From Table 2 and Table 3, it is seen that on duty cycle 1% and 10%, under high traffic load, no data is sent. Under highest traffic load (0.01s), if duty cycle is increased to 30%, the residual energy of the sink node becomes 0.069698 Joules and the throughput is .07kbps which is more than the 20% duty cycle (0.01 Kbps). So in mission critical applications where traffic load instantly increases, the SMAC duty cycle can configure to 30% to achieve higher throughput. This is because there is negligible difference in residual energy. Under high traffic loads if the duty cycle is increased to 50%, there is increase in throughput .44Kpbs, but it will be at the cost of residual energy. It is observed that under high traffic load, the residual energy becomes negligible at 1000s at 20% and 30% duty cycle and the throughput is also very low.

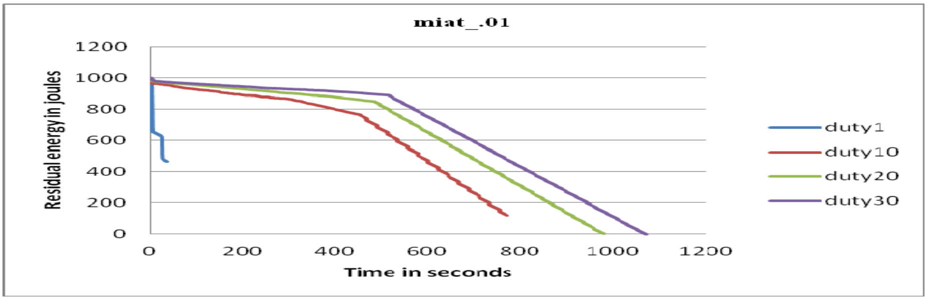


Fig. 3. Residual Energy vs Time for Message inte-arrival time .01s

In scenario2,as seen from Fig 4, varying the duty cycle from 1% to 30% the residual energy of sink node is observed for miat=1seconds which is considered to be the higher traffic rate. From Table.2, It is seen that, for miat=1s (higher traffic rate), residual energy decreases with increase in duty cycle. For 30% duty cycle, the residual energy is maximum. From Table 3, it is seen that the throughput is also increased to 1.61Kbps.So it is observed that for traffic rate 100bytes/second, the

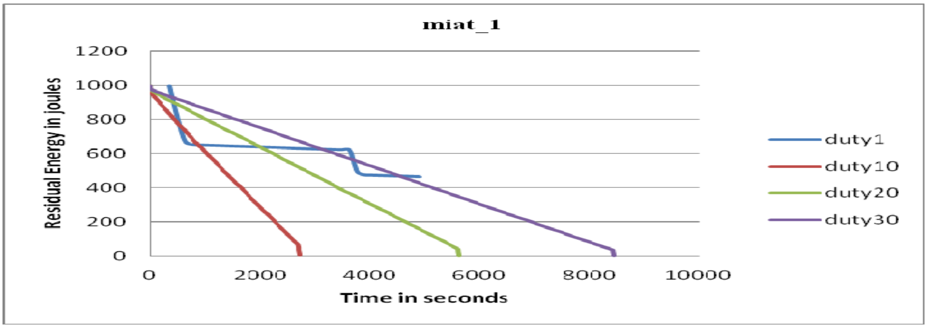


Fig. 4. Residual Energy vs Time for Message inte-arrival time 1s

SMAC can perform better when duty cycle is increased to 30% in single hop scenario. The throughput achieved is 1.61 kbps. This is the scenario where the residual energy is maximum at end of simulation.

- **Moderate traffic scenario**

Scenario 3: In Fig. 5 is shown that for miat(message inter-arrival time)=5seconds, which is considered to be the medium traffic load, varying the duty cycle from 1% to 30% ,residual energy of sink node is varied. From Table 2 and Table 4, it is seen that 20% duty cycle is best because it saves residual energy and at the same time the throughput is also same at both 20% and 30% duty cycle. The throughput achieved is .64 kbps.

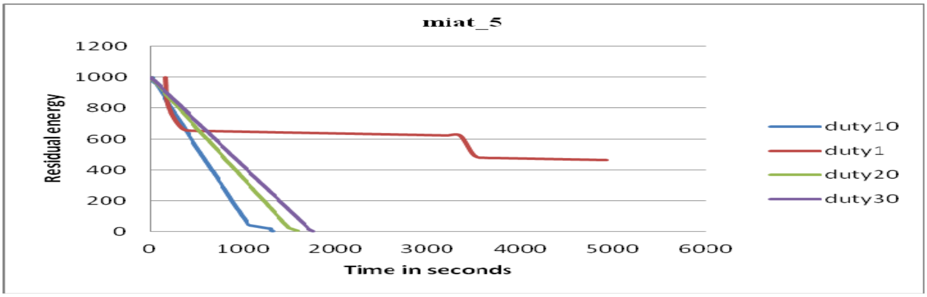


Fig. 5. Residual Energy vs Time for Message inte-arrival time 5s

- **Low traffic rate scenarios**

Scenario 4 & 5: These are the scenarios with lower traffic loads. In Fig.6 and Fig.7, varying the duty cycle from 1% to 30% the residual energy of sink node is observed for MIAT=25seconds which is considered to be the lower traffic rate. Varying the duty cycle from 1% to 30% the residual energy of sink node is observed for MIAT=50seconds which is considered to be the lower traffic rate.

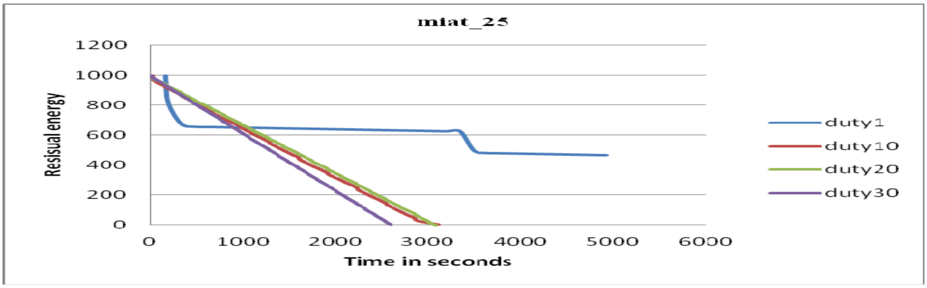


Fig. 6. Residual Energy vs Time for Message inte-arrival time 25s

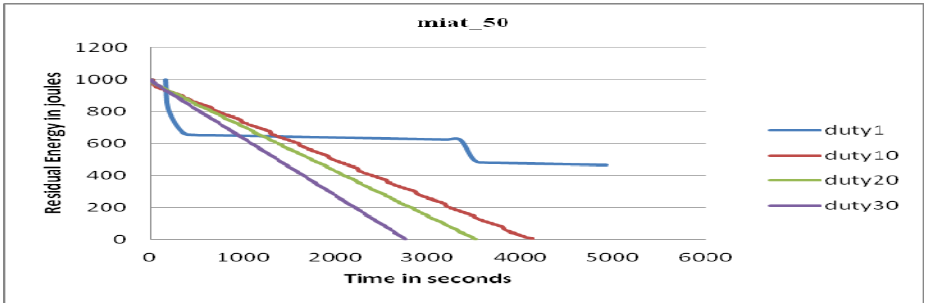


Fig. 7. Residual Energy vs Time for Message inte-arrival time 50s

As per the table 2, the graph in Fig 8 has been plotted for residual energy under different traffic loads for different duty cycles. For miat=25s and miat=50s (low traffic rate), for better results of residual energy we should keep the duty cycle low. Form Table 5,it is seen The throughput achieved is .13kbps when miat is 25s and .06kbps when miat is 50s.Since the throughput is not improved with increasing the duty cycle, hence we should keep the duty cycle low to achieve better residual energy.

Table 2. The residual energy with different duty cycles and different traffic loads

MIAT	Re.Ener	duty	Re.Ener	duty	Re.Ener	duty	Re.Ener	duty
0.01	463.6788	1	119.1136	10	2.601626	20	0.069698	30
1	463.6788	1	5.707545	10	0.694504	20	1.05225	30
5	463.6788	1	1.394872	10	7.193864	20	1.382184	30
25	463.6788	1	0.966691	10	0.148844	20	1.285791	30
50	463.6788	1	7.746957	10	1.224436	20	1.057871	30

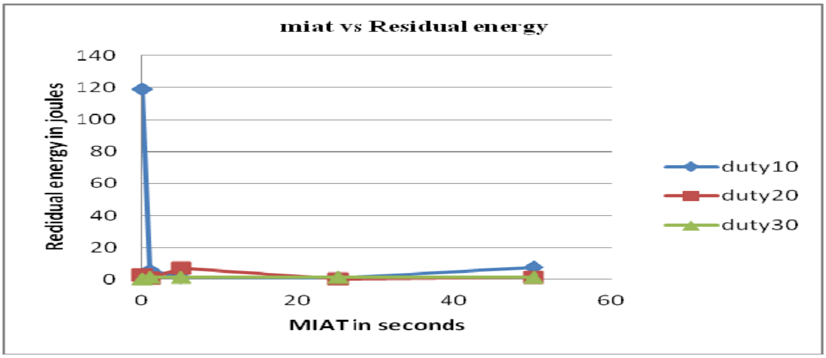


Fig. 8. Residual energy vs Message inter-arrival time

Measurement of Throughput:

Scenario1 & Scenario 2

Table 3. Throughput (MIAT=.01s & MIAT=1s)

	MIAT=.01s					MIAT=1s		
duty %	T(Kbps)	t1(start)	t2(stop)		duty %	T(Kbps)	t1(start)	t2(stop)
1%	0	40	0		1%	0	40	1037
10%	0	40	0		10%	0.42	40	1149.42
20%	0.01	40	210.95		20%	1.06	40	1062.73
30%	0.07	40	157.52		30%	1.61	40	1054.9
50%	0.44	40	111.93		50%	2.71	40	1037

Scenario3

Table 4. Throughput under different duty cycles (MIAT=5s)

	MIAT=5s		
duty %	T(Kbps)	t1(start)	t2(stop)
1%	0	0	0
10%	0.53	40	1056.28
20%	0.64	40	1505.14
30%	0.64	40	1719.69

Scenario 4 & Scenario 5

Table 5. Throughput (MIAT=25s & MIAT=50s)

	MIAT=25s					MIAT=50s		
duty %	T(Kbps)	t1(start)	t2(stop)		duty %	T(Kbps)	t1(s)	t2(s)
1%	0	40	0		1%	0	40	0
10%	0.13	40	2891.94		10%	0.06	40	4046.93
20%	0.13	40	3040.94		20%	0.06	40	3490.49
30%	0.13	40	2566.88		30%	0.06	40	2741.97

The graph has been plotted for throughput vs. duty cycle at different message inters arrival time in Fig.9. It is observed that for 50% duty cycle and message inter-arrival time=1s, the maximum throughput is achieved, but when the residual energy is also considered then the 30% duty cycle is the best choice. At 30% duty cycle, the throughput is maximum and the residual energy is almost equivalent to 20% duty cycle as seen from Fig.8.Varing duty cycle from 1% to 20% ,high throughput is not achieved which is at times the requirement for some applications in wireless sensor networks.

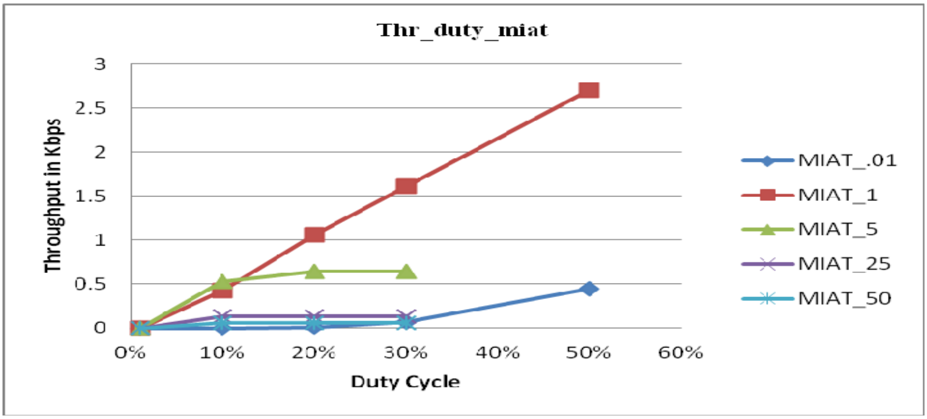


Fig. 9. The Throughput vs duty cycle at different message inter-arrival time(MIATs)

5 Conclusion

We have analyzed the S-MAC protocol in terms of throughput and residual energy. It is observed that it's not necessary that increasing the duty cycle percentage will always decrease the residual energy. At higher traffic rates, we can achieve energy effeciency (residual energy) and throughput with the proper choice of duty cycle. So [13]for mission critical application where apart from energy effeciency, the data transport performance is also important, the S-MAC protocol can be used with the proper choice of duty cycle to achieve better throughput and energy effeciency. In single hop scenario, improvement in SMAC protocol is needed when message inter-arrival time is much less (.001s). [14][15][16][17] In multihop scenario, the same analysis can further be performed and some enhancements in SMAC can be done accordingly, to improve its performance for applications where the throughput, energy effeciency and the latency all are important at higher traffic rates.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A Survey on Sensor Networks. IEEE Communications Magazine 40(8), 102–114 (2002)
2. Yick, J., Mukherjee, B., Ghosal, D.: Wireless Sensor Network Survey. Elsevier Computer Networks 52(12), 2292–2330 (2008)
3. Ye, W., Heidemann, J., Estrin, D.: An energy-efficient MAC protocol for wireless sensor networks. In: Proc. IEEE INFOCOM, New York, NY, pp. 1567–1576 (June 2002)
4. LAN MAN Standards Committee of the IEEE Computer Society. Wireless LAN medium access control (MAC) and physical layer (PHY) specification, IEEE Std. 802.11-1997 edn. IEEE, New York (1997)
5. Ye, W., Heidemann, J., Estrin, D.: Medium Access Control with Coordinated Adaptive Sleeping for Wireless Sensor Networks. IEEE/ACM Transactions on Networking 12(3), 493–506 (2004)
6. Demirkol, I., Ersoy, C., Alagöz, F.: MAC protocols for. Wireless Sensor Networks: a Survey. IEEE Communications Magazine 44(4), 115–121 (2006)
7. Enz, C.C., El-Hoiydi, A., Decotignie, J.D., Peiris, V.: WiseNET: An Ultralow-Power Wireless Sensor Network Solution. IEEE Computer 37(8) (August 2004)
8. Jamieson, K., Balakrishnan, H., Tay, C.: Sift: A MAC Protocol for Event-Driven Wireless Sensor Networks. MIT Lab. Comp. Sci., Tech. rep. 894 (May 2003)
9. Lu, G., Krishnamachari, B., Raghavendra, C.S.: An Adaptive Energy-Efficient and Low-Latency MAC for Data Gathering in Wireless Sensor Networks. In: Proc. 18th Int’l. Parallel and Distrib. Processing Symp., p. 224 (April 2004)
10. Misra, S., Mohanta, D.: Adaptive listen for energy-efficient medium access control in wireless sensor networks. Journal Multimedia Tools and Applications 47(1) (March 2010)
11. SCADDS: Scalable Coordination Architectures for Deeply Distributed Systems web page, `http://www.isi.edu/scadds/projects/smac/`
12. The Network Simulator - ns-2 homepage, `http://www.isi.edu/nsnam/ns/`
13. The VINT project. The NS Manual. UC Berkeley, LBL, USC/ISI, and Xerox PARC, `http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf`
14. Ameen, M., Islam, S.M.R., Kwak, K.: Energy Saving Mechanisms for MAC Protocols in Wireless Sensor Networks. International Journal of Distributed Sensor Networks, 16 pages (2010) Article ID 163413
15. Suryachai, P., Roedig, U., Scott, A.: A Survey of MAC Protocols for Mission-Critical Applications in Wireless Sensor Networks. IEEE Communication Surveys & Tutorials 14(2) (Second Quarter 2012)
16. Misra, S., Woungang, I., Mishra, S.C.: Guide to Wireless Sensor Networks. Springer-Verilog London Limited (2009)
17. Karl, H., Willig, A.: Protocols and Architectures for Wireless Sensor Networks. John Wiley & Sons (2005)
18. Greis, M.: Tutorial for the network simulator ns, `http://www.isi.edu/nsnam/ns/tutorial/index.html`
19. Smac-users – Discussions by users of S-MAC web page, `http://mailman.isi.edu/mailman/listinfo/smac-users`
20. Energy Model Update in ns-2 web page, `http://www.isi.edu/ilense/software/smac/ns2_energy.html`

Routing Protocols in Mobile Ad-Hoc Network: A Review

Bahuguna Renu, Mandoria Hardwari lal, and Tayal Pranavi

Department of Information Technology,
G.B. Pant University of Agriculture & Technology, Pantnagar
{renubahuguna,drmandoria,diligent.virgos}@gmail.com

Abstract. Mobile ad-hoc network comprises of wireless nodes that communicate each other by exchanging the information. The path chosen for transferring the information from one node to another node is called routing and the protocols used is called routing protocols. The requirement of routing protocol is to send and receive information among the nodes with best suited path with the minimum delay. Correct and efficient route establishment between a pair of nodes is the primary goal of routing protocol . Many routing protocols for manet have been proposed earlier. Performance analysis of routing protocol is a significant challenge in the research area. This paper, gives a review work done on existing protocols characteristics of MANET and comparison between them.

Keywords: Routing protocols, MANET, Proactive, Reactive, Hybrid.

1 Introduction

Mobile Adhoc Network (MANET) is a collection of independent mobile nodes that can communicate to each other. MANETs being researched by several organizations and institutes. MANETs employ the traditional TCP/IP structure to provide end-to-end communication between nodes. However, due to their mobility and the limited resource in wireless networks, each layer in the TCP/IP model require redefinition or modifications to function efficiently in MANETs. One interesting research area in MANET is routing.

Most applications in the MANET are based upon unicast communication. Thus, the most basic operation in the IP layer of the MANET is to successfully transmit data packets from one source to one destination. The forwarding procedure is very simple in itself: with the routing table, the relay node just uses the destination address in the data packet to look it up in the routing table. If the longest matching destination address is found in the table, the packet is sent to the corresponding next hop. The problem that arises is how the routing table is built in the nodes in the MANET.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 52–60, 2013.

2 Existing Protocols in MANET

Routing Protocols:
Routing protocols basicaly divided into severals parts:

2.1 Table driven (proactive) routing protocols
2.2 Source initiated (demand driven/reactive) routing protocols
2.3 hybrid routing protocols

2.1 Table Driven Routing Protocol

In this type of routing protocols maintain consistent and up to date routing information of each node in the network. These protocols store there routing information on each node and when there is any changes in network topology updation has to be made throughout the network.various protocols are shown into the fig-1.and the basic characterstics of table driven routing protocol are:

Table 1. Characterstics of table driven routing protocol [16]

Protocol	RS	No. of Tables	Frequency of updates	HM	Critical nodes	Characteristic feature
CGSR	H	2	Periodic	No	Yes, Cluster head	Clusterheads exchange routing information
DSDV	F	2	Periodic and as required	Yes	No	Loop free
FSR	F	3 and a list	Periodic and local	No	No	Controlled frequency of updates
HSR	H	2(link-state table& location management)	Periodic, within each subnet	No	Yes, Cluster head	Low CO and Hierarchical structure
OLSR	F	3(Routing, neighbour & topology table)	Periodic	Yes	No	Reduces CO using MPR
STAR	H	1 and a 5 lists	Conditional	No	No	Employes LORA and/or ORA. Minimize CO
WRP	F	4	Periodic	yes	No	Loop freedom using predecessor info

R =routing structure; HM=hello message; H=hierarchical; F=flat; CO=control overhead; LORA=least overhead routing approach; ORA=optimum routing approach; LM=location manager.

Table 2. Complexity comparison of proactive routing protocols [16]

Protocol	CT	MO	CO	Advantages/Disadvantages
CGSR	O(D)	O(2N)	O(N)	Reduced CO/cluster formation and maintenance
DSDV	O(D.I)	O(N)	O(N)	Loop free/high overhead
FSR	O(D.I)	O(N^2)	O(N)	Reduces CO/high memory overhead, reduced accuracy
HSR	O(D)	O(N^2.L)+(S) +O(N/S)+ (N/n)	O(n.L)/I +O(1)/J	Low CO/location management
OLSR	O(D.I)	O(N^2)	O(N^2)	Reduced CO and contention/2-hop neighbor knowledge required
STAR	O(D)	O(N^2)	O(N)	Low CO/high MO and processing overhead
WRP	O(h)	O(N^2)	O(N)	Loop free/memory overhead

CT=convergence time; MO=memory overhead; CO=control overhead; (1)=a fixed number of update tables is transmitted; V =number of neighbouring nodes;N =number of nodes in the network; n=average number of logical nodes in the cluster; I =average update interval; D=diameter of the network; S=number of virtual IP subnets; h=height of the routing tree; X =total number of LMs (each cluster has an LM); J =nodes to home agent registration interval; L=number of hierarchical level.

2.2 Source Initiated Demand Driven Routing Protocol

In on-demand routing protocols routes are generated as and when we required. When a source wants to send any information to a destination,it invokes the route discovery mechanisms to find the path to the destinations. The route remains valid till the destination is reachable or until the route is no longer needed. various protocols are shown into the fig-1.

Table 3. Basic characteristics of reactive routing protocols [16]

Protocol	**RS**	**Multiple routes**	**BC**	**Route metric method**	**RMI**	**Route reconfiguration strategy**
AODV	F	No	Yes, hello messages	Freshest & SP	RT	Erase route then SN or local route repair
ABR	F	No	Yes	Strongest Associativity &SP	RT	LBQ
ARAN	F	Yes	No	SP	RT	Use alternate route or back track until a route is found
DSR	F	Yes	No	SP, or next available in RC	RC	Erase route the SN
FORP	F	No	No	RET & stability	RT	A Flow HANDOFF used to use alternate route
LAR	F	Yes	No	SP	RC	Erase route then SN
SSA	F	No	Yes	Strongest signal strength & stability	RT	Erase route then SN
TORA	F	Yes	No	SP, or next available	RT	Link reversal & Route repair

RS=routing structure; H=hierarchical; F=flat; RT=route table; RC=route cache; RET=route expiration time; SP=shortest path; SN=source notification; BC= Beacons; RMI= Route maintained in; LBQ=localised broadcast query.

Table 4. Complexity comparison of reactive routing protocols [16]

Protocol	**TC [RD]**	**TC [RM]**	**CC [RD]**	**CC [RM]**	**advantage**	**Disadvantage**
AODV	O(2D)	O(2D)	O(2N)	O(2N)	Adaptable to highly dynamic topologies	Scalability problems,large delays,hello messages
ABR	O(D+P)	O(B+P)	O(N+R)	O(A+R)	Route stability	Scalability problems
ARAN	O(D+P)	O(D+P)	O(N+R)	O(A+R)	Low overhead, small control packet size	Flooding based route discovery process

Table 4. (*Continued*)

DSR	O(2D)	O(2D)	O(2N)	O(2N)	Multiple routes, Promiscuou s overhearing	Scalability problems due to source routing and flooding, large delays
FORP	O(D+P)	O(B+P)	O(N+R)	O(N+R)	Employees a route failure minimisatio n technique	Flooding based route disovery process
LAR	O(2S)	O(2S)	O(2M)	O(2M)	Localised route discovery	Based on source routing, flooding is used if no location information is available
SSA	O(D+P)	O(B+P)	O(N+R)	O(A+R)	Route stability	Scalability problems, large delays during route failure and reconstruction
TORA	O(2D)	O(2D)	O(2N)	O(2A)	Multiple routes	Temporary routing loops

TC=time complexity; CC=communication complexity; RD=route discovery; RM=route maintenance; CO=control overhead; D=diameter of the network;
N =number of nodes in the network; A=number of affected nodes; B=diameter of the affected area; G=maximum degree of the router; S =diameter of the nodes in the localised region; M =number of nodes in the localised region; X =number of clusters (each cluster has one cluster-head); R=number of nodes forming the route reply path, RREP, BANT or FLow_SETUP; P =diameter of the directed path of the RREP, BANT or FLow_SETUP; jEj=number of edges in the network.

2.3 Hybrid Routing Protocol

Hybrid routing protocols are a new generation of protocol, which are both proactive and reactive in nature. These protocols are designed to increase scalability by allowing nodes with close proximity to work together to form some sort of a backbone to reduce the route discovery overheads. various protocols are shown into the fig-1.

Table 5. Basic characteristics of hybrid routing protocols [16]

Protocol	RS	Multiple routes	BC	Route metric method	Route maintained in	Route reconfiguration strategy
ZHLS	H	Yes, if more than one virtual link exists	No	SP or next available virtual link	Intrazone and interzone tables	Location request
ZRP	F	No	Yes	SP	Intrazone and interzone tables	Route repair at point of failure and SN

RS=routing structure; H=hierarchical; F=flat; SP=shortest path; SN=source notification; Bc=beacons.

Table 6. Complexity comparison of hybrid routing protocols [16]

Protocol	TC [RD]	TC [RM]	CC [RD]	CC [RM]	advantage	Disadvantage
ZHLS	Intra:O(I)/ Inter:O(D)	O(I)/O(D)	O(N/M)/O(N+V)	O(N/M)/ O(N+V)	Reduction of SPF, low CO	Static zone map required
ZRP	Intra:O(I)/ Inter: O(2D)	O(I)/O(2D)	O(Z_N)/ O(N+V)	O(Z_N)/ O(N+V)	Reduce retransmission	Overlapping zones

TC=time complexity; CC=communication complexity; RD=route discovery; RM=route maintenance; I =periodic update interval;N =number of nodes in the network; M =number of zones or cluster in the network; ZN =number of nodes in a zone, cluster or tree; ZD =diameter of a zone, cluster or tree; Y =number of nodes in the path to the home region; V =number of nodes on the route reply path; SPF=single point of failure; CO=control overhead.

3 Comparison of Protocols

Comparison between different routing protocols are shown here.

Table 7. Parametric Comparison [14]

Parameters	Proactive Protocols	Reactive Protocols	Hybrid Protocols
Availability of routing information	Available when required	Always available stored in tables	Combination of both
Latency	High due to flooding	Low due to routing tables	Inside zone low outside similar to Reactive protocols
Mobility	Support Route maintenance	Periodical updates	Combination of both
Periodic Updates	Not needed as route available On demand	Yes. Whenever the topology of the network changes	Yes needed inside the zone
Scalability level	Not suitable for large networks	Low	Designed for large networks
Storage capacity	Low generally Depends upon the number of routes	High ,due to the routing tables	Depends on the size of Zone, inside the zone Sometimes high as Proactive protocol
Routing Overhead	Low	High	Medium
Routing Philosophy	Flat	Flat/Hierarchical	Hierarchical
Routing Scheme	On demand	Table driven	Combination of both

Table 8. Pros and Cons Comparison [14]

Protocol	Advantages	Disadvantages
Proactive	Proactive Information is always available. Latency is reduced in the network	Overhead is high, Routing information is flooded in the whole network
Reactive	Path available when needed overhead is low and free from loops.	Latency is increased in the network
Hybrid	Suitable for large networks and up to date information available Complexity increases	Complexity increases

4 Conclusion

The paper begins with a brief introduction of routing protocols in mobile ad-hoc Networks. we reviewed and studied the features of different protocols used in mobile ad-hoc network then we are discussed all the above review persented in this paper. Finally, the stated review work of the routing protocols were discussed in this paper.

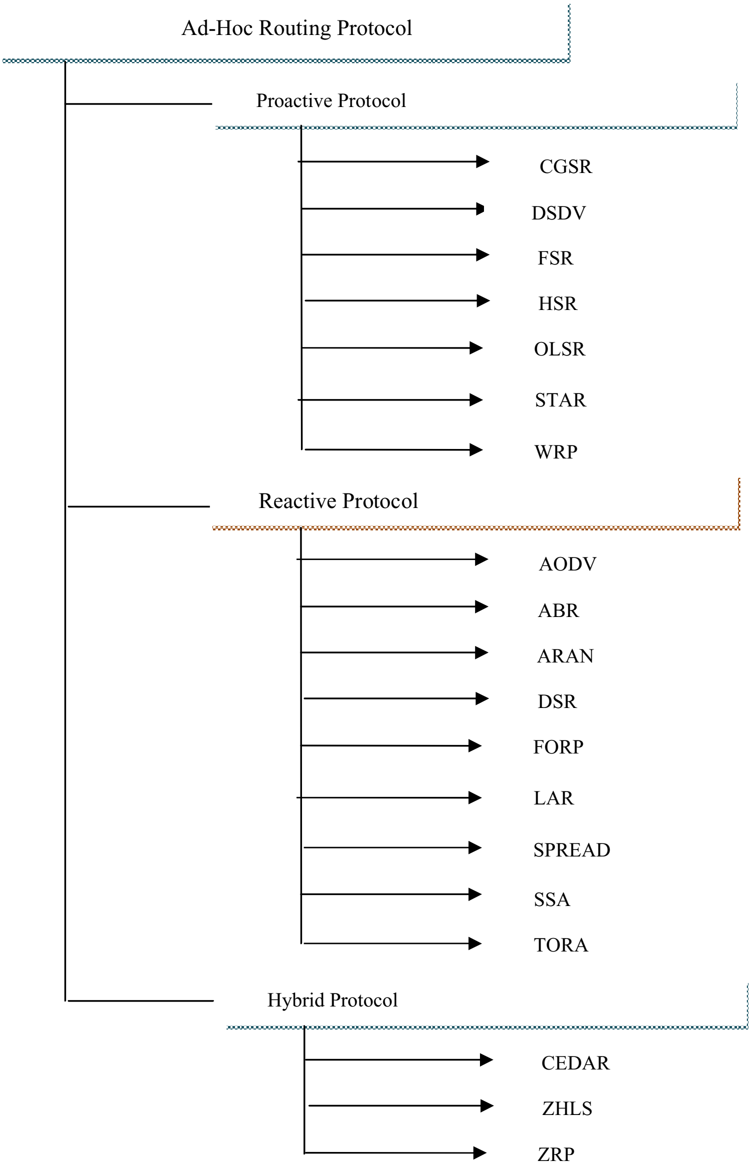


Fig. 1. Ad-hoc Routing Protocol(classification)

References

1. Royer, E.M., Toh, C.K.: A Review of Current Routing Protocols for Ad-Hoc Mobile Wireless Networks. IEEE Personal Communications (1999)
2. lee, S.-J., Gerla, M., Toh, C.K.: A Simulation Study of Table-Driven and On Demand Routing Protocols for Mobile Ad Hoc Networks. IEEE Network (1999)
3. Broch, J., Maltz, D.A., Johnson, D.B., et al.: A Performance Comparison of Multi-Hop Wireless Ad Hod Network Routing Protocols. In: MOBICOM 1998 (1998)
4. Park, V., Corson, S.: Temporally-Ordered Routing Algorithm (TORA) Version 1 Functional Specification (2001), `http://draft-ietf-manet-tora-spec-04.txt`
5. Bae, S.H., le, S.-J., Su, W., Gerla, M.: The Design, Implementation, and Performance Evaluation of the On-Demand Multicast Routing Protocol in Multihop Wireless Networks. IEEE Network (2000)
6. Perkins Charles, E., Royer Elizabeth, M., Das Samir, R., Marina Mahesh, K.: Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks. IEEE Personal Communications (2001)
7. Song, J.-H., Wong, V.W.S., Leung, V.C.M.: Efficient On-Demand Routing for Mobile Ad Hoc Wireless Access Networks. IEEE Journal on Selected Areas in Communications 22(7) (2004)
8. Pirzada, A.A., McDonald, C., Datta, A.: Performance Comparison of Trust-Based Reactive Routing Protocols. IEEE Transactions on Mobile Computing 5(6) (2006)
9. Bai, R., Singhal, M.: DOA: DSR over AODV Routing for Mobile Ad Hoc Networks. IEEE Transactions on Mobile Computing 5(10) (2006)
10. Abusalah, L., Khokhar, A., Guizani, M.: A Survey of Secure Mobile Ad Hoc Routing Protocols. IEEE Communications Surveys & Tutorials 10(4) (Fourth Quarter 2008)
11. Xu, H., Wu, X., Sadjadpour, H.R., Garcia-Luna-Aceves, J.J.: A Unified Analysis of Routing Protocols in MANETs. IEEE Transactions on Communications 58(3), 911–922 (2010)
12. Sanjay kumar, P., Prasant kumar, P., Puthal, B.: Review of routing protocols in sensor and adhoc networks. International Journal of Reviews in Computing © 2009-2010 IJRIC (2009-2010)
13. Geetha, J., Gopinath, G.: Ad Hoc Mobile Wireless Networks Routing Protocols – A Review. Journal of Computer Science 3(8), 574–582 (2007)
14. Robinpreet, K., Mritunjay Kumar, R.: A Novel Review on Routing Protocols in MANETs. Undergraduate Academic Research Journal (UARJ) 1(1) (2012) ISSN : 2278 – 1129
15. Hsieh, H.-Y., Sivakumar, R.: Routing: On Using the Ad Hoc Network Model in Cellular Packet Data Networks. In: Proc. ACM MOBIHOC 2002 (2002)
16. Mehran, A., Tadeusz, W., Eryk, D.: A review of routing protocols for mobile ad hoc networks (2004)
17. Sunil, T., Ashwani, K.: A Survey of Routing Protocols in Mobile Ad-Hoc Networks. International Journal of Innovation, Management and Technology 1(3) (August 2010) ISSN: 2010-0248
18. Fahim, M., Nauman, M.: MANET Routing Protocols vs Mobility Models: APerformance Evaluation. IEEE, ICUFN (2011)
19. Charu, W., Kumar, S.S.: Mobile Ad-Hoc Network Routing Protocols:A Comparative Study. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) 3(2) (2012)

Carrier Aggregation for Enhancement of Bandwidth in 4G Systems

Jolly Parikh and Anuradha Basu

Bharati Vidyapeeth's College of Engineering,
Delhi, India
{jolly.parikh,anuradha.basu}@bharatividyapeeth.edu

Abstract. Since ITU-R had officially completed the formal definition of Third Generation (3G) systems in 1997, focus has been shifted to the Fourth Generation (4G) wireless cellular systems. The paper provides an overview of the different aspects of the proposed carrier aggregation technique, which would enable LTE-A systems to fully utilize the wider bandwidths up to 100MHz and as well maintain the backward compatibility with LTE systems. The contiguous and non contiguous carrier aggregation techniques have been discussed and their deployment scenarios have been illustrated. The technique of carrier aggregation will not only provide a wide bandwidth of 100 MHz but shall also help in achieving higher peak data rates and better coverage for medium data rates.

Keywords: carrier aggregation (CA), contiguous and non contiguous component carriers (CCs), deployment scenarios, LTE-Advanced systems.

1 Introduction

Fast and more efficient mobile internet access demands, pressurize the mobile service providers to think about adopting the advanced version of IMT, i.e. IMT-advanced. The IMT-advanced requirements of the peak data rates of 1Gbps and 500 Mbps in downlink and uplink respectively, can be achieved by using the wider bandwidths of up to 100 MHz [1]. Such wider portions of continuous spectrum is rarely available in practice for cellular based mobile communication use i.e. below 3 GHz. 3GPP proposed carrier aggregation (CA) technology, in release 10, as a potential solution for increasing the LTE bandwidth [2]. In CA, multiple component carriers (CC) of smaller bandwidths, belonging to same or different spectrum bands are aggregated by the operators to scale their spectrum bandwidths so as to enable high data rates in downlink as well as uplink transmission.

These CCs follow LTE release 8 numerology and core physical layer design, there by guaranteeing the LTE-Advanced systems (release 10 and beyond) to be backward compatible with the LTE systems (release 8 and 9). Release 10 users can access multiple spectrum bands belonging to contiguous or noncontiguous frequency bands simultaneously, to send and receive data [3]. Legacy users can access the system using one of the aggregated CCs. With CA, spectrum efficiency can be increased due

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 61–74, 2013.

to the feature of aggregation of non-contiguous carriers, proper utilization of different carriers results in various deployments scenarios of both homogenous and heterogeneous networks. The LTE compatible CC, enable operators to migrate from LTE to LTE-Advanced systems and at the same time continue service to LTE users [4]. Apart from providing higher peak data rates, CA also provides better coverage for medium data rates. Here the use of lower orders of modulation and lower code rates, reduces the required link budget, transmission power and interference [5]. Figure 1 [6] shows the improvement in the bandwidth of LTE-Advanced with the increase in the number of CCs.

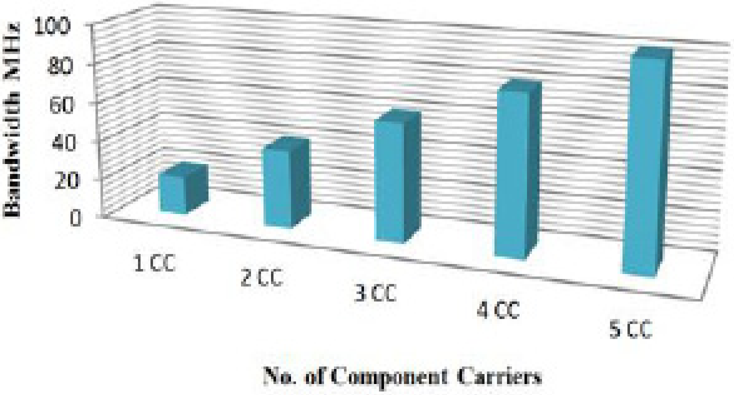


Fig. 1. Relationship between component carriers and Bandwidth MHz [6]

In spite of all these advantages, allocation of multiple CCs to power limited LTE-A user equipments (UEs), experiencing unfavorable channel condition, is not advisable. This is because when a UE has reached its maximum transmission power, increasing bandwidth does not result in increase in data rates for a UE, transmitting simultaneously over multiple CCs. The transmission power reduces due to effects of increased PAPR and peak to average power ratio and intermodulation [7].

2 Carrier Aggregation Configurations in LTE Advanced Systems

In order to ensure backward compatibility to legacy release 8 users, LTE-A systems aggregate multiple release 8 CCs for providing wider transmission bandwidth [8]. Multiple LTE carriers, with bandwidth of upto 20 MHz each, can be transmitted in parallel to/from an LTE-A supporting terminal. LTE devices which do not support LTE-A feature will use one of these 20 MHz aggregated CC.

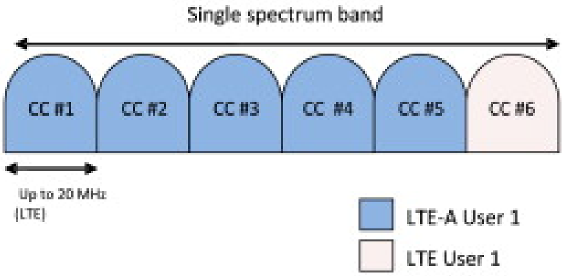


Fig. 2(a). Carrier aggregation in contiguous bandwidth

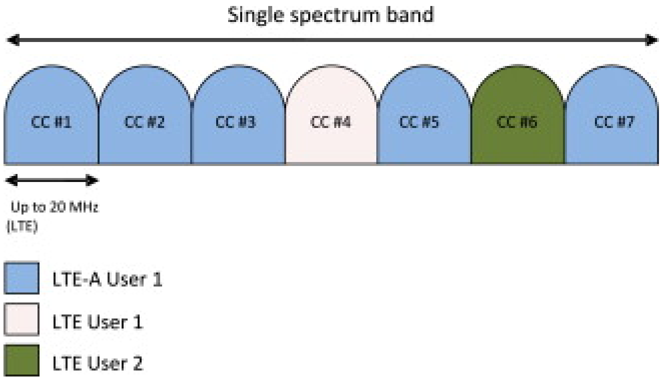


Fig. 2(b). Carrier aggregation in noncontiguous bandwidth, single band

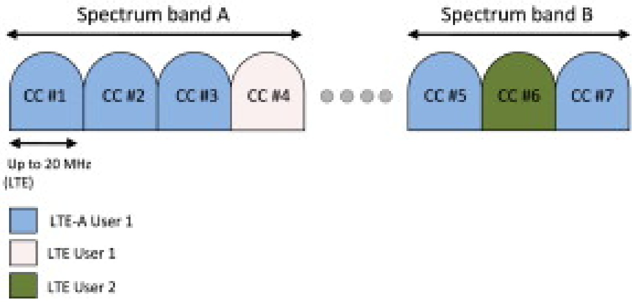


Fig. 2(c). Carrier aggregation in non-contiguous bandwidth, multiple bands

The aggregation of carriers can be done in different ways. Figure 2 (a,b,c) [5] shows the different CA types. In the Intraband aggregation with frequency contiguous CC mode, the available multiple spectrum bands of upto 20 MHz each and adjacent to each other, can be used to form a 40 MHz band single spectrum. Intra band aggregation with noncontiguous CC makes use of CCs in same frequency band e.g. 800 MHz but not necessarily adjacent to each other. Whereas in Interband aggregation with non-contiguous CCs, carriers belonging to multiple bands (located in different frequency bands) are aggregated to serve a single unit of LTE-A UE [9] e.g. 20 MHz (800 MHz) + 20 MHz (2.1 GHz) or 20 MHz (1.8 GHz) + 20 MHz (2.6 GHz) CCs aggregation combinations can be used to form a 40 MHz transmission bandwidth for downlink with interband noncontiguous CA. Thus this technique provides a practical approach for the operators to fully utilize either the current spectrum resources of frequency bands, in the frequency range below 4 GHz [10] which have already been allocated to legacy systems like GSM and UMTS, or the unused and scattered frequency bands in the range of frequencies >4GHz. Table I [5] lists the primary proposed LTE-A deployment scenarios.

Table 1. Primary LTE-Advanced Deployment Scenarios.

Scenario no.	**Description**	**Transmission BWs of LTE-A carriers**	**No. of LTE-A CCs**	**Bands for LTE-A carriers**	**Duplex modes**
1	Single-band contiguous spec. alloc. @ 3.5 GHz band for FDD	UL: 40 MHz DL: 80 MHz	UL: Contiguous 2x20 MHz CCs DL: Contiguous 4x20 MH CCs	3.5 GHz band	FDD
2	Single-band contiguous spec. alloc. @ Band 40 for TDD	100 MHz	Contiguous 5x20 MHz CCs	Band 40 (3.5 GHz band)	TDD
3	Multi-band non-contiguous spec. alloc. @ Bands 1, 3 and 7 for FDD	UL: 40 MHz DL: 40 MHz	UL/DL: Non-contiguous 10 MHz CC@Band 1+10 MHz CC@Band 3+20 MHz CC@Band 7	Band 3 (1.8 GHz), Band 1 (2.1 GHz), Band 7 (2.6 GHz)	FDD
4	Multi-band non-contiguous spec. alloc. @ Bands 39, 34 and 40 for TDD	90 MHz	Non-contiguous 2x20 + 10 + 2x20 MHz CCs	Band 39 (1.8 GHz), Band 34 (2.1 GHz), Band 40 (2.3 GHz)	TDD

A noncontiguous FDD deployment scenario has been shown in Figure 3. Here a 40 MHz system is formed using 10 MHz (1.8 GHz) + 10MHz (2.1 GHz) + 20 MHz (2.6 GHz). Spurious emissions into adjacent bands are taken care of by guard bands. Figure 3 illustrates the contiguous CA scenario wherein 2 x 20 MHz CCs form a 40 MHz system. Here, the available spectrum is more efficiently utilized due to narrower or no guard band between adjacent carriers of same eNB. To ensure the spacing of multiples of 300 KHz between the centre frequencies of the adjacent carriers, unused sub carriers are used [11]. Prioritized deployment scenarios for LTE-A were proposed in [12]. Aggregation in 3.5 MHz band is also planned.

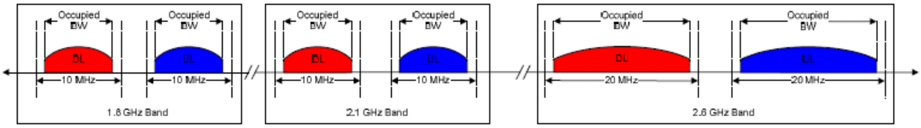


Fig. 3. Non contiguous FDD deployment over multiple bands

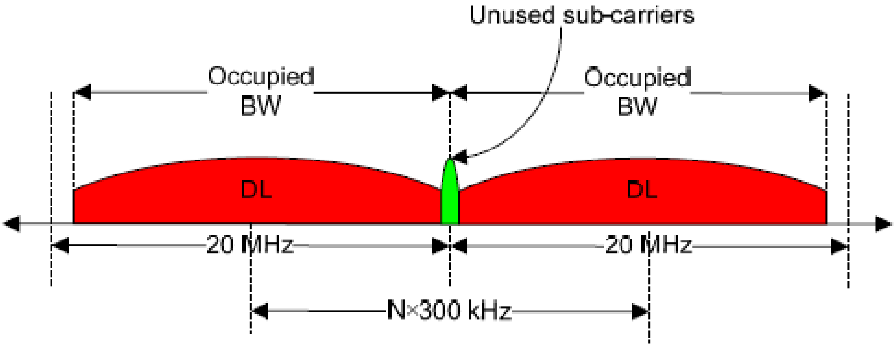


Fig. 4. Contiguous FDD deployment in single band

For a LTE-A UE unit, the contiguous CA is easier to be implemented as it can be realized with single FFT and a single radio frequency (RF) unit. Resource allocation and management algorithms can be easily implemented for this scheme. But the complexity of the LTE-A UE increases in case of noncontiguous CA, as the radio network planning phase and the design of the RRM algorithms should consider that different CCs will exhibit different path loss and Doppler shifts [13].

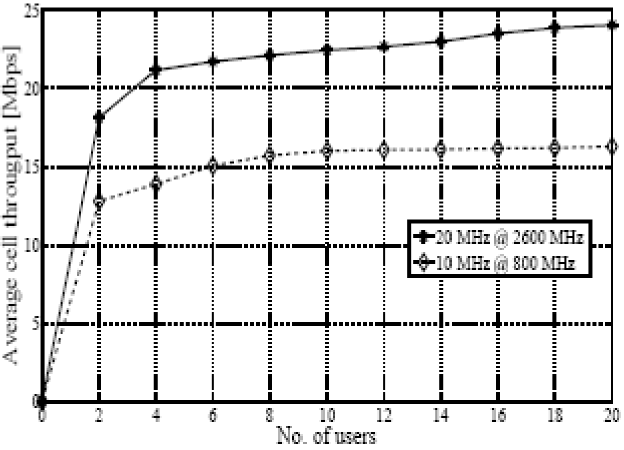


Fig. 5. Cell Throughput comparison of 20 MHz and 10 MHz

In [14], Fourat Haider and et. al have studied the effect of path losses on the cell throughput in different single carrier frequency bands. As shown in figure 5, in 2.6GHz band there is only 50% of increase in cell throughput, even when the bandwidth is doubled, as compared to operation in the 800 MHz band. Whereas, figure 6 shows that due to frequency diversity, if CA is implemented at 2.6 GHz, the throughput is 50% higher than single carrier transmission at 800 MHz and 2.6 GHz.

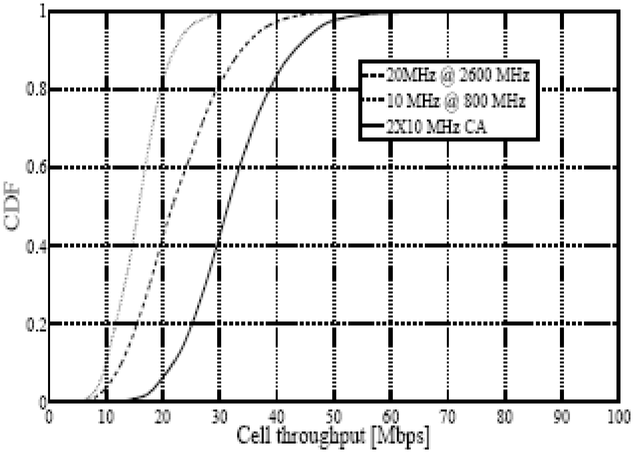


Fig. 6. CDF of Cell Throughput

Yuan G. and et. al. have discussed the effects of Doppler shifts on BER performance under different modulation schemes [10]. On account of large Doppler shifts, systems with high speed mobiles will have self interference or intersystem interference. To overcome this problem, in contiguous CA schemes, LTE technical specification [15] suggests about allocating 10% of total bandwidth specifically for inserting guard bands between adjacent component carriers. Figure 7 shows the BER performance with and without Doppler frequency shift.

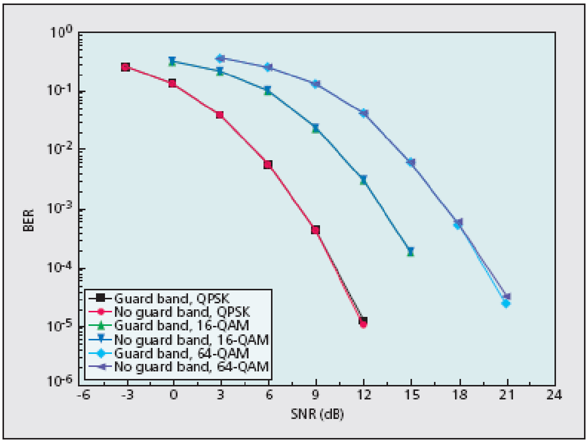


Fig. 7. BER Performance With and Without Doppler Frequency Shift

3 Downlink and Uplink Carrier Configurations in LTE-A Systems

In case of symmetric configurations, uplink and downlink carriers are always paired. In asymmetric CA configurations there are multiple downlink CCs for a UE and only one uplink CC. This causes ambiguity in the selection of downlink CC. The LTE-A eNB has difficulty in identifying the CC to which the UE will anchor in DL at the time of random access response to the UE. As per release 10, the LTE radio interface can be configured with any number of carriers (upto 5 carriers), of any bandwidth, inclusive configuration of downlink and uplink, but the number of uplink carriers cannot exceed the number of downlink carriers [16].

4 Deployment Scenarios of LTE-A Carrier Aggregation

The goal of CA is to improve the data rates for users within overlapped areas of cells. For this LTE release 10 has agreed upon several deployment scenarios [17] for design of LTE-A CA systems. Figure 8 [17] shows some of these deployment scenarios exemplifying how, in real network, CA could be deployed in flexible manner. In practice, we can consider large number of CCs and also deployments with mixed scenarios, but here only 2 CCs – CC1 and CC2 – operating at F1 and F2 frequencies respectively have been considered.

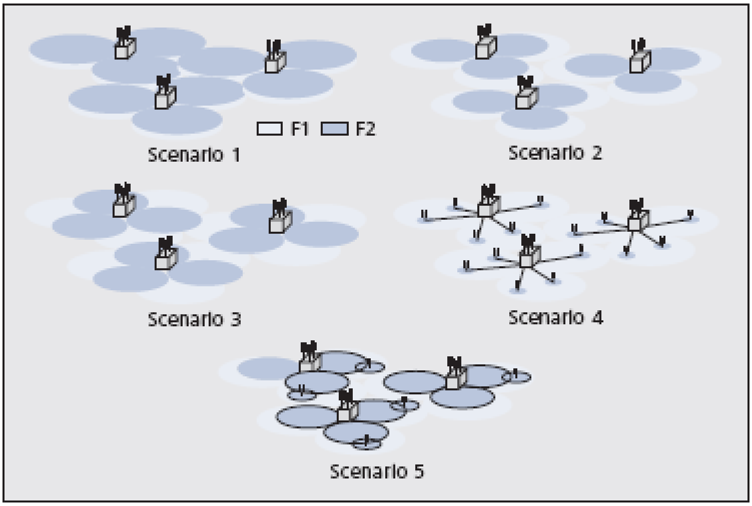


Fig. 8. Carrier aggregation deployment scenarios (F2>F1) [17]: a) scenario 1; b) scenario 2; c) scenario 3; d) scenario 4.

Various factors like the type of area involved in the existing deployed network i.e. urban, suburban or rural; usage of common antennas for all the CCs used; presence or absence of hotspots in the coverage area, determine the most efficient scenario of deployment.

Deployment Scenario 1. This is the most envisaged scenario. Here, eNB antennas with F1 and F2 carrier frequencies are collocated. F1 and F2 belong to same band. The antennas have same beam directions/patterns for both the CCs. Figure 8a shows that the antennas provide nearly same coverage on both carriers as the path loss will be similar within band. Both carriers can support mobility.

Deployment Scenario 2. Here, the collocated eNB antennas with F1 and F2 carrier frequencies operate on CCs belonging to different bands. As shown in figure 8b the coverage for a CC of higher frequency may be smaller as compared to that of the CC of low frequency because there are larger path losses in the higher frequency band. The carrier in the lower frequency band supports mobility whereas high frequency band carrier enables higher data rates and throughput. To solve the problem of intercell interference it is essential to have different coverage for the eNB antennas. For this, they are operated at different transmit power levels and CCs of same band may be deployed at the eNBs. Higher user throughputs are possible, in either cases of CA, at places where overlapping of the coverage of CCs occur.

Deployment Scenario 3. In this scenario eNB antennas operating at F1 and F2 are collocated. The CCs belong to different bands. The antennas are having different beam directions/patterns to support the different sectorization schemes (e.g. 120° sectoring or 60° sectoring). Deployment shown in figure 8c provides improved data rates and throughput at the sector boundaries of cells operating at F1. This is achieved by intentionally shifting the direction of the antenna beams, of cells operating at F2 frequency, towards the boundaries of cells of F1 frequency. CA can be implemented where the coverage area overlaps for the CCs belonging to same eNB.

Deployment scenario 4. In this case the eNB with F1 frequency CC provides macro coverage, while the remote radio heads (RRHs) of F2 frequency CC are placed at traffic hotspots to enable throughput by another CC. The coverage of eNB operating at F1 frequency determines the mobility. The CCs are of different bands. The RRH cells are connected to the eNB via optical fibers so as to enable aggregation of CCs between the macrocell and RRH cell. The operators having deployments as shown in figure 8d can improve the system throughput even with the help of low cost RRH equipments.

Deployment scenario 5. The deployment shown in figure 8e is similar to that of scenario 2 but with additional deployment of frequency selective repeaters and RRH. This scenario has the limitation that the frequency selective repeaters boost certain CCs only as a result of which there is variation in the propagation delay across boosted and non boosted CCs. This in turn requires separate transmission timing control for each CC during UL transmission. This type of scenarios are not considered for LTE release 10 for the UL transmission but are considered in a later release wherein interband CA can also be used for the UL transmission, so as to support traffic growth, spectrum allocation, and feasibility in the device implementation. As for DL transmission, release 10 does consider scenarios with RRH and repeaters [18].

5 Spectrum and Network Sharing among Service Providers

With the advancement in technology, different clients of a service provider are able to use UEs that can support various Radio Access Technologies (RATs) like the LTE, WIMAX, HSPA. This feature enables the service providers to provide a coverage to all of their users by developing different RATs [5]. It solely depends on the operator to decide as to which RAT (s) should the UE be attached to so as to achieve optimum spectrum utilization and the required QoS. Figure 9 shows the operation of Multi RAT scenarios.

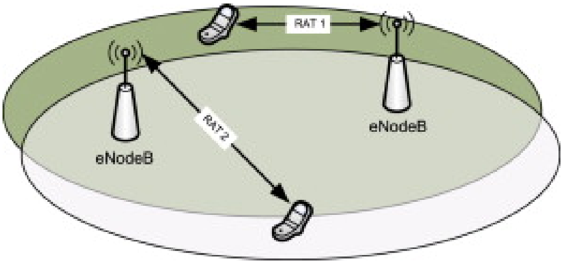


Fig. 9. Multi-RAT scenario

Every RAT will need different spectrum resources. It is the responsibility of the eNB/base stations to manage the spectrum (resources) of the RAT in use. Service providers adopt spectrum sharing concept (network sharing) which is supported by 3GPP [20,21]. Figure 10 [19] depicts the different spectrum aggregation scenarios for FDD. Herein, the service providers can get access to resources of different networks thereby reducing their initial investments. The 2 scenarios of spectrum sharing are shown in figure 11. The spectrum band is shared, between operators using either in noncontiguous carrier aggregation (case 1) or contiguous Carrier Aggregation (case 2) depending upon the way the spectrums are used by the operators. The general scenarios for multioperator network sharing are identified by 3GPP in [20]. FDD and TDD spectrum sharing on dynamic basis have been discussed in [22].

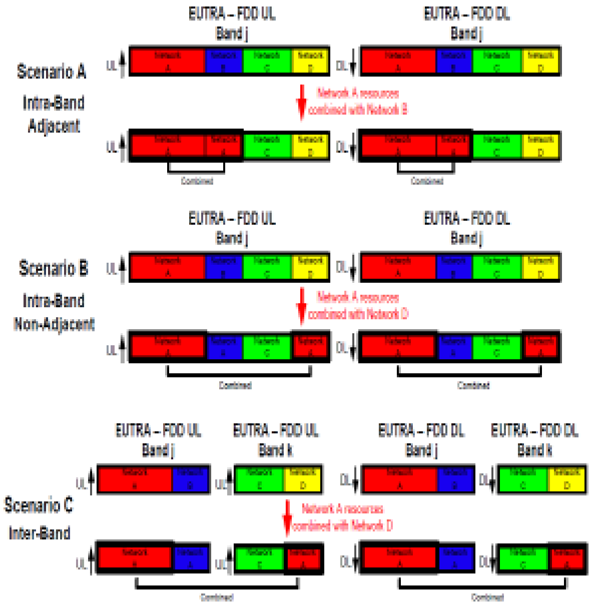
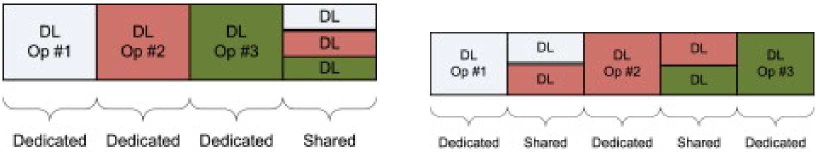


Fig. 10. Spectrum Aggregation Scenarios for FDD [19]



(a) Scenario 1

(b) Scenario 2

Fig. 11. Spectrum sharing scenarios

6 Band Aggregation

The spectrum being already overcrowded, the regulatory bodies are finding it difficult to allocate a contiguous band of 100 MHz to a single operator. The tables 2 and 3 show the bands assigned for E-UTRA (LTE) [23, 24]. It can be seen that they are not broad enough to provide a 100 MHz bandwidth which is essential to meet the basic requirement of IMT-A as defined by ITU (see figure 12). Hence it will be required by the operators to combine the various bands, as shown in table 4, as per the availability of the spectrum resources [4]. Future releases will have to support interband carrier aggregation for TDD, both UL-DL configurations, on different bands in order to ensure coexistence with the already deployed TDD systems.

Table 2. Operating Bands for LTE FDD

LTE Operating Band	Uplink (UL) Operating Band (MHz)	Downlink (DL) Operating Band (MHz)	Main Regions of Use
1	1920 - 1980	2110 – 2170	Asia, Europe
2	1850 - 1910	1930 – 1990	Americas, Asia
3	1710 - 1785	1805 – 1880	Americas, Asia, Europe
4	1710 - 1755	2110 – 2155	Americas
5	824 - 849	869 – 894	Americas
6	830 - 840	875 – 885	Japan
7	2500 – 2570	2620 – 2690	Asia, Europe
8	880 – 915	925 – 960	Asia, Europe
9	1749.9 – 1784.9	1844.9 – 1879.9	Japan
10	1710 – 1770	2110 – 2170	Americas
11	1427.9 – 1452.9	1475.9 – 1500.9	Japan
12	698 – 716	728 – 746	USA
13	777 – 787	746 – 756	USA
14	788 – 798	758 – 768	USA
15	Reserved	Reserved	Reserved
16	Reserved	Reserved	Reserved
17	704 – 716	734 – 746	USA
18	815 – 830	860 – 875	Japan
19	830 – 845	875 – 890	Japan
20	832 – 862	791 – 821	Europe
21	1447.9 – 1462.9	1495.9 – 1510.9	Japan
22	3410 – 3500	3510 – 3600	

Table 3. Operating Bands for LTE TDD

LTE Operating Band	Band Allocation (MHz)	Main Regions of Use
33	1900 – 1920	Asia (not Japan), Europe
34	2010 – 2025	Asia, Europe
35	1850 – 1910	Americas
36	1930 – 1990	Americas
37	1910 – 1930	
38	2570 – 2620	Europe
39	1880 – 1920	China
40	2300 – 2400	Asia, Europe
41	2496 – 2690	USA

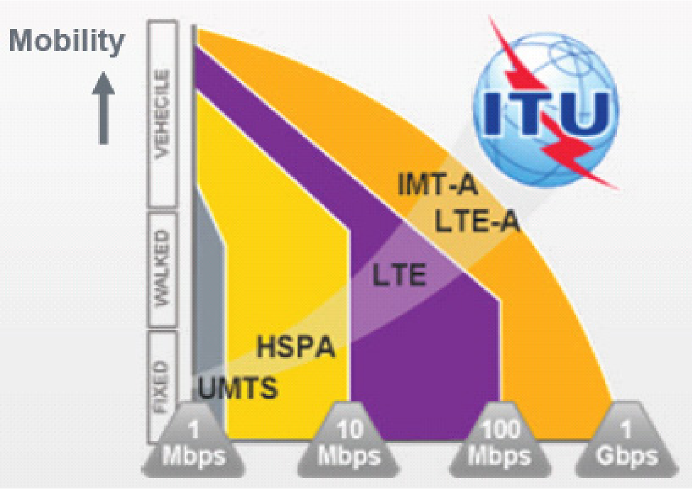


Fig. 12. Data Rates Requirements of IMT-A defined by ITU

Table 4. CA Band or Band Aggregation

CA Band or Band Aggregation	Operator	Duplex Mode
Band 4 + Band 17	AT&T	FDD
Band 2 + Band 17	AT&T	FDD
Band 4 + Band 5	AT&T	FDD
Band 5 + Band 17	AT&T	FDD
Band 41	Clearwire	TDD
Band 38	CMCC	TDD
Band 20 + Band 7	Orange	FDD
Band 3 + Band 7	Telia Sonera	FDD
Band 4 + Band 12	US Cellular	FDD
Band 5 + Band 12	US Cellular	FDD
Band 4 + Band 13	Verizon	FDD

7 Conclusion

This article provides an overview of the carrier aggregation technique used in LTE-Advanced systems for increasing the bandwidth and thereby improving the data rates of LTE-Advanced users. The deployment of CA based system enables efficient spectrum utilization and also increases the cell and user throughput significantly. The fully backward compatibility feature of CA for LTE-Advanced enables the coexistence of legacy Rel. 8 terminals and LTE-Advance terminals. To meet the IMT-Advanced peak data rate requirements, the initial focus of 3GPP was on intraband CA. Release 10 supports interband CA in Downlink for a limited number of bandwidth combination whereas Release 11 will provide full support for non-contiguous carrier aggregation.

References

1. 3GPP TR 36.913 v8.0.0: Requirements for further advancements for E-UTRA (LTE-Advanced) (June 2008)
2. Recommendation ITU-R.M.1645: Framework and overall objectives of the future development of IMT 2000 ans systems beyond IMT 2000 (June 2003)
3. Dahlman, E., Parkvall, S., Skold, J.: 4G LTE/LTE-Advanced for Mobile Broadband, pp. 132–134. Elsevier, UK (2011)
4. Shen, Z., Papasakellariou, A., Montojo, J., Gerstenberger, D., Xu, F.: Overview of 3GPP LTE-Advanced Carrier Aggregation for 4G wireless Communications. IEEE Communications Magazine (2012)
5. Akyildiz, I., Gutierrez-Estevez, D., Chavarria, R.: The evolution to 4G cellular systems: LTE-Advanced. Physical Communication 3, 217–244 (2010)
6. Yonis, A., Abdullah, M., Ghanim, M.: Design implementation of Intra band Contiguous Component Carriers on LTE-A. International Journal of Computer Applications 41(14) (March 2012)
7. 3GPP R4-091910: LTE-A MC RF requirements for contiguous carriers (May 2009)
8. Wang, H., Rosa, C., Pedersen, K.I.: Performance Analysis of Downlink Inter-band Carrier Aggregation in LTE-Advanced. In: IEEE Vehicular Technology Conference, September 1-5 (Fall 2011)
9. 3GPP, TR 36.815 V9.1.0: Further advancements for E-UTRA LTE-Advanced feasibility studies in RAN WG4, Rel. 10 (June 2010)
10. Yuan, G., Zhang, X., Wang, W., Yang, Y.: Carrier aggregation for LTE-Advanced mobile communication systems. IEEE Communications Magazine 48(2), 88–93 (2010)
11. Ratasuk, R., Tolli, D., Ghosh, A.: Carrier Aggregation in LTE-Advanced. In: IEEE 71st Vehicular Technology Conference (VTC 1010- Spring) (2010)
12. Krouk, E., Semenov, S.: Modulation and coding techniques in wireless communications, pp. 600–603. Wiley, UK
13. Ingemann, K., Frederiksen, F., Rosa, C., Nguyen, H., Garcia, L., Wang, Y.: Carrier aggregation for LTE-Advanced: Functionality and Performance Aspects. IEEE Communication Magazine (June 2011)
14. Haider, F., Hepsaydir, E., Binucci, N.: Performance Analysis of LTE-Advanced Networks in Different Spectrum Bands. In: Wireless Advanced (WiAD) Conference, July 20-22 (2011)
15. 3GPP TS 36.104 v. 9.1.0: Base Station (BS) RadioTransmission and Reception, Tech. Spec. Group Radio Access Network, Rel. 9 (September 2009)
16. Anritsu: LTE-Advanced: Carrier Aggregation, white paper (September 2011)
17. 3GPP TS 36.300 v10.3.0: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall description, stage 2 (Release 10), TSG RAN
18. Iwamura, M., Etemad, K., Fong, M., Nory, R., Love, R.: Carrier Aggregation Framework in 3GPP LTE-Advanced. IEEE Communication Magazine (August 2010)
19. 4G Americas: 4G Mobile broadband evolution: 3GPP Release 10 and beyond HSPA, SAE/LTE and LTE-Advanced, pp. 51-52 (February 2011)
20. 3GPP, TR 22.951 Service aspects and requirements for network sharing, Tech. Rep., (December 2009), http://ftp.3gpp.org/specs/html-info/22951.htm

21. 3GPP, TS 23.251 Network sharing; architecture and functional description, Tech. Rep. (March 2010), http://ftp.3gpp.org/specs/html-info/23251.htm
22. CELTIC/CP5-026 WINNER+, D1.2 initial report on system aspects of flexible spectrum use, Tech. Rep. (January 2009)
23. Georgoulis, S.: How to Test Carrier Aggregation in LTE-Advanced Networks, White Paper (2012), http://www.eetimes.com/design/test-and-measurement/4376129/How-to-test-carrier-aggregation-in-LTE-Advanced-networks
24. Rhode & Schwarz: LTE-Advanced Technology Introduction Application Note (March 2010)

A Stable Energy-Efficient Location Based Clustering Scheme for Ad Hoc Networks

Anubhuti Roda Mohindra and Charu Kumar

Jaypee Institute of Information Technology,
Sector-128, Jaypee Wish Town, Noida, U.P., India
{anubhuti.mohindra,charu.kumar}@jiit.co.in

Abstract. *Clustering* is an approach applied in mobile ad hoc networks to organize the hosts in groups called *Clusters*. The paper focuses on implementing a Clustering technique that conserves *Energy*, provides *Stability* and keeps the *Communication Traffic* and *Overheads low* using *Location based approaches* by setting *Thresholds* for different parameters in dynamic ad hoc environment of uncontrolled movement of nodes, low bandwidth of wireless channel and energy limited nodes. The mechanism uses pools of Primary and Secondary cluster heads (CHs) to provide uninterrupted communication due to node mobility. The protocol saves network resources by reducing the information exchange amongst the nodes and limiting it to clusters and CHs. The paper includes a case that models different scenarios based on the proposed protocol. Also, characteristic analysis of our clustering scheme with some of the existing algorithms is done which shows the strength of this scheme over others.

Keywords: Clustering, Mobile Ad Hoc network (MANET), Stability, Location Services, GPS, Energy Conservation, ELS.

1 Introduction

A Mobile Ad-hoc Network (MANET) is a temporary self-configuring wireless network of autonomous mobile nodes that communicate without any preinstalled infrastructure or central administration over the wireless links. MANETs provide an efficient method for a dispersed set of nodes to establish communication without the need of an infrastructure [3].

Every Node acts as a router as well as a host and forwards the packets. Routers are free to move in an uncontrolled manner, thereby leading to constant topology change. Initial setup configuration, routing support and maintenance in mobile networks are some of the most significant challenges that make research in ad hoc networks very interesting. First, uncontrolled node mobility results in a highly dynamic network with rapid topological changes causing frequent route failures. A good network mechanism for this environment is required to dynamically adapt to the changing topology. Second, the underlying wireless channel working as a shared medium, provides much less and variable bandwidth than wired networks making available bandwidth per

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 75–85, 2013.

node even lower. Therefore, the right routing protocol should be bandwidth-efficient by expending a minimal overhead for computing routes so that much of the remaining bandwidth is available for the actual data communication. Third, nodes run on batteries which have limited energy supply. In order for nodes to stay and communicate for longer periods, it is desirable that the routing approach be energy-efficient as well. This provides another reason why overheads must be kept low. Thus a mechanism must be there in the network that meets the conflicting goals of dynamic adaptation and low overhead to deliver good overall performance [13].

Currently there is a growing interest in designing protocols for mobile ad hoc networks and one of the most researched topics in this respect is *Clustering*. It is an approach applied in mobile adhoc networks to organize the hosts in groups called *Clusters* in hierarchical or flat fashion. The interconnected Clusters communicate inside themselves as well as amongst themselves [6]. Its architecture makes it an extremely efficient approach which helps in improving the network performance; provide scalability in dynamic environment and reducing energy consumption of power limited nodes. The purpose of any clustering algorithm is to produce and maintain a connected cluster where every node can play a different role in different situation [8]. A *Cluster Head* (CH) is chosen for every Cluster using any of the various available schemes for CH Election, which acts as a Base Station for the cluster communication. Many Clustering schemes have been proposed for MANETs where they aim to meet certain needs of the system [7].

There are numerous *Advantages* of Clustering for MANETs [6]: Better *Protocol Performance* for MAC Layer; Improved *Power Consumption*; *Routing Tables' size minimized;* Reduced *Transmission Overhead*; *Aggregated* topology information; *Bandwidth and Energy conservation* are just a few of them. On the other hand, there are some areas of concern that draw the attention of researchers [6]: *High Maintenance cost* because of large message exchanges when topology changes; Some nodes are more *Power consuming* like Cluster Heads and Gateways leading to shutdown of the nodes; *Stability* of Clusters and *Location Management.*

In this paper we focus on the design of an algorithm that implements Clustering while Conserving Energy of the nodes, providing Stability to reduce reclustering and keeping the Communication Traffic and Overheads low using Location Services. There have been many schemes that have implemented each of these features separately to achieve effective routing and adaptability for the vast application set of mobile ad hoc networks.

This paper focuses on a scheme that provides Energy Conservation and Stability in Clustering by managing the Power levels of information signals between the CHs and other members [1] [3]. Further we add Location Management Services to provide Efficient and Fast Routing along with Quick Recovery in case of Link Failures [2]. The algorithm uses a Hierarchical approach with Location services for attaining a Stable Clustered Network. To achieve Energy Conservation for every node the CH and the cluster are chosen depending on the proximity of the CH to the node, so that less power is used while sending and receiving packets. The protocol manages both Intercluster and Intracluster Location information using tables and also includes Path Activation and Deactivation process for accuracy. We assume that all the nodes are

GPS enabled. Also, Mobility parameters are included to achieve Stability in the network which is extremely critical in vehicular ad hoc applications [3].

In Section 2 we present some of the existing related clustering approaches for stability, location management and energy conservation in terms of their strengths and weaknesses. Section 3 describes the proposed clustering approach. In Section 4 we perform a case evaluation under different scenarios and verify the working of our protocol. Section 5 gives a characteristic analysis of our protocol and some previously proposed protocols. Finally, Section 6 concludes the paper.

2 Survey of Related Work

There have been previous works on Energy Conservation in Clustering like 'New Clustering Schemes for Energy Conservation in Two-Tiered mobile Ad Hoc Networks' by Ryu, in which hierarchical clustering schemes that optimize the configuration of the cluster in a way where every node has maximum energy conservation and minimum drop in communication rate are proposed [1]. The schemes designed take into account the *small size and battery supplied power characteristics of nodes* working in dynamic real time event–driven scenarios and reconfigurations. Two schemes have been designed in the paper- *Single Phase and Double Phase* Clustering. Both the algorithms are based on Paging from the Master Nodes and Acknowledgement from the Slave Nodes. Every Master node sends paging signal to all the Slave nodes at the same Power level. Every Slave Node will send an Acknowledgement to one Master Node from which it received highest power level because that Master must be nearest to that Slave, thereby saving the Transmitted Power. It is seen that the Power consumption of every node decreases with both less number and high number of nodes. Although these schemes provide a mechanism for energy conservation and lowering call drop rate. But they are designed assuming static topology which is quite unfeasible considering dynamic nature of mobile adhoc networks. Also, they consider Error-free conditions and no channel contention, which is again very difficult to achieve.

In another paper based on Location Services to improve network performance [2] the protocol manages Location information of a Source, a Destination or any in-between host. The proposed scheme works on single level because multilevel schemes require high Bandwidth because of high Communication overhead. Each CH acts as a Location Server for providing location information which results in Less Delay in searching for a Location. The protocol provides Intracluster management by maintaining two tables – Local Connection Table (LC) and Intracluster Routing Table (Intra R). For Intercluster management, Location State Table (LS) have been used in addition to LC and Intra R tables. But there are certain limitations like the direction and moving rate are considered constant. Also, the protocol works well when the numbers of clusters are less and Cluster Size is Large.

One of the stable clustering algorithms, is proposed in [3] for pseudo linear mobile ad hoc networks. In such networks the movement of the nodes is highly linear in one direction without much change in their motion parameters. It is designed on one-hop approach and based on *Doppler's Shift*. The working is in two phases – Initial stage Cluster Formation and Progressive stage Cluster Maintenance. Stability of the network is dependent on the estimated duration of the communication links which depends on DV which further depends on the ratio of the frequency of received signal to the frequency of known communication signal. It was shown that in Cluster maintenance stage, with increase in the speed of a node, the number of members leaving a cluster is more in case number of NULL nodes is less. Authors also found, with increase in the number of NULL nodes, number of clusters formed decrease but they become more stable. The schemes proposed manage CH Contention very well and Cluster size can vary depending on the application. But the schemes don't take Ripple effects of clustering are into account and won't work in case there are any changes in the direction of the nodes. Considering all these approaches, it is seen that though some of the issues are addressed but none of them meets all the requirements of an efficient clustering. In the next section we propose a protocol which implements all of the following: *Stability, Energy Conservation, Location Services and Low overhead.*

3 Proposed Clustering Scheme: ELS

In our Clustering approach, in *the initial set up stage* every node connects to a CH only and not amongst themselves. The scheme utilizes the grouping approach and reduces the information updates by limiting it to a cluster and amongst CHs. Every Cluster is formed by two sets of nodes - Cluster Heads and Cluster Nodes that have unique id numbers. For every cluster our scheme makes use of *Primary* and *Secondary Cluster heads*. Pool of a predefined number of nodes are created for backup of both Primary and Secondary CH using election schemes like Lowest Id, Highest Degree, Weights etc. [8].

3.1 System

Every Cluster node sets up connection with a primary CH that is nearest to it i.e. sending maximum power so as to provide maximum *Energy Conservation*. The CHs work on a set of rules designed for efficient working of the network:

(i) During communication, the power of signal transmitted by the CH varies depending on the power left with the node. A minimum threshold value is set 'T_p' and if the power of the primary CH gets below that value, a new CH is made the primary CH from the pool of elected CHs. The size of any cluster depends on the distance till which minimum power 'P_{min}' of every CH is maintained. This helps in preventing breakdown of the CHs.

(ii) Also, the Bandwidth capacity 'B_c' allotted for communication to every CH is fixed. When a CH's majority communication channels are busy and the capacity left with it gets below the minimum set capacity value, we delegate the future communication tasks to the Secondary CH with again capacity ‘B_c’ and it also starts working as a CH in the cluster. It provides uninterrupted network through Load balancing.

3.2 Location Management

We assume that every node in the network is *GPS enabled* and therefore every CH is able to get and store Location Information of all the nodes in their Cluster and other Cluster Heads. In our approach every CH (Primary as well as Secondary) stores two tables: *Routing Location Table* and *Network Location Table*. The first table - Routing Location Table stores Intracluster information that includes id numbers of the cluster nodes, location coordinates available through GPS and distance to every cluster node (calculated using location coordinates). The other table – Network Location Table stores Intercluster information which includes other CHs’ id numbers, shortest route to every other CH in terms of number of hops, node ids in every cluster and neighboring CHs that are one hop away.

Table 1. Cluster Head C-4 Intracluster Table

Node Ids	Node Location	Distance from CH
I-1	12,20,45	7.7’88deg N
I-2	30,60,20	10.3’120degN W
I-3	22,50,12	5.5’90degS

Table 2. Cluster Head C-4 Intercluster Table

CH Ids	Node Ids in Cluster	Shortest Path to CH	Neighboring CHs
C-1	I-1,I-3,I-8	C-3, C-9	C-3, C-12
C-2	I-9,I-5,I-4,I-6	C-2, C-5, C-6, C-9	C-2
C-3	I-2,I-7,I-10	C-8, C-10, C-15	C-8, C-9,C-14

3.3 Cluster Communication

The algorithm proposes the communication in the Clustered network through different types of messages.

(i) Any node that either wants to join a particular cluster or change its cluster membership will send a JOIN packet to the CHs of that cluster. In return, if the current CH (primary or secondary) has the required communication capacity, it will acknowledge the node by sending a JOIN-ACCEPT packet to that node. Also, the CH adds the node information in its tables.

(ii) If a node wants to leave a cluster or is moving to another cluster, it will see a decrease in the strength of the power signal from its current CH that will be below the threshold value set. Node will then send a BYE message to the CH, which in turn removes the node from it member list in the table and responds with a BYE ACCEPT packet.

(iii) Whenever a Cluster node changes its location, UPDATE-NODE packet is sent to the CH with the new information. The CH updates its tables and sends an acknowledgement through UPDATE-NODE ACCEPT packet.

(iv) Whenever there is a change in the CH like a change in the location of CH or its members, change in the membership of nodes or a new CH assignment, UPDATE-CH packet is broadcasted by the CH to all other CHs in the network with aggregated information of the cluster over a fixed time period. Also, if a new CH is designated it gets the copies of all the tables with the previous CH.

(v) Any node that wants to send a message to any other node in the same cluster or a node outside will send a REQUEST message to the CH to get the information of the destination node. If the destination node is within the cluster, CH provides the location info and the next hop in a REQUEST REPLY message. The node then uses Dijkstra's algorithm for calculating shortest path to its destination. The communication of information between cluster nodes of the same cluster is then done through a NODE-MSG packet.

(vi) On the other hand if the node is outside the cluster, CH asks the source node to send it the information that needs to be sent to the destination in REQUEST REPLY message.

Node then sends a NODE-MSG packet to it CH, which sends all the information to the head of that cluster in which destination node is present in a PASS-INFO packet. CH uses CH-MSG to pass the information to the required node. In the packet, source will be the original node that actually sent the message.

Any message can be sent via any alternate intermediate CH in case the intermediate CH on shortest path fails. This is possible because all CHs have the clustered network map.

3.4 Stability

To prevent duplication of packets and loops, we attach a *Sequence Number* with every packet. If Acknowledgement of any form is not received, nodes resend the message after lapse of a time period 'T_{rep}'. To maintain a stable cluster, our algorithm sets up *a Minimum Connection Time, a Maximum and a Minimum Mobility Factor.*

For every node that wants to join a cluster, its estimated Connection time to the CH and its mobility is calculated by using the algorithm in [10]. If the predicted Connection time of that node satisfies the Minimum set Threshold, only then it is allowed to be a part of the Cluster. Also, if any node or a CH has mobility factor is higher than the Maximum set value, it means it will not allow the network to stay stable. Therefore, the node is not made a part of that cluster. On the other hand if two CHs which are extremely close to each other, if their mobility factor is lower than the Minimum value, it means they will stay in each other's vicinity for a large amount of time and thus only one of them is required to service all the nodes. CH with higher power level at that point of time is allowed to stay a CH and the other becomes a node under it.

4 Case Evaluation

In this section, to better understand the proposed scheme, we consider a MANET environment and implement our protocol under various scenarios.

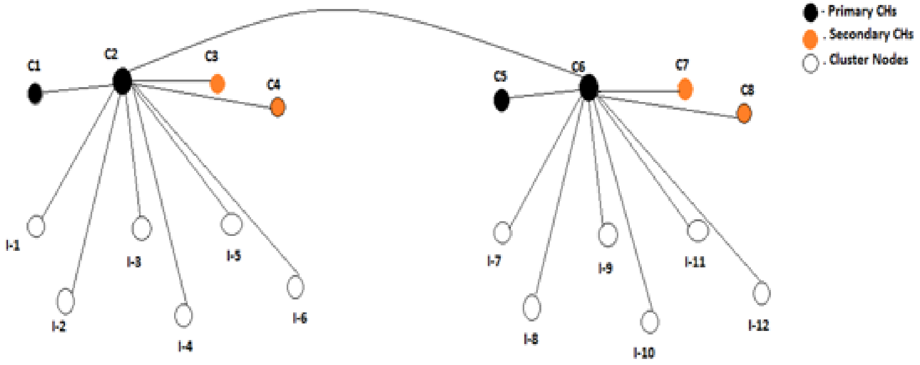


Fig. 1.

As we can see in *Fig. 1*, our network is divided into two clusters. In every cluster there are two primary CHs - (C1,C2,C5,C6) and two secondary CHs - (C3,C4,C7,C8) with currently C2 and C6 being the currently designated CHs. I-1 to I-6 are the members of Cluster 1 and I-7 to I-12 form Cluster 2.

- According to the proposed algorithm, the CHs C2 and C6 have a fixed power level 'PWR'. While communicating let us assume for CH- C2 *: $PWR < T_p$*

  As a result, the other primary CH- C1 is designated the task of communication and all nodes of the cluster (I-1 to I-6) are now connected to C1. The Intercluster and Intracluster tables from C2 are then passed on to C1 for future communication.
- The currently designated CHs- C1 and C6 are allocated a fixed Bandwidth Capacity B_c such that : $B_c1 + B_c2 + B_c3 + B_c4 + B_c5 + B_c6 + B_cRemaining = B_c$ *(for C1)*

  Let us suppose that $B_cRemaining$ < *Minimum Capacity*, therefore we make secondary CH 'C3' a designated CH of Cluster 1 *(Fig. 2)* and it is also assigned a Bandwidth Capacity – B_c to handle the future connections. Again, the new CH gets copies of the tables.

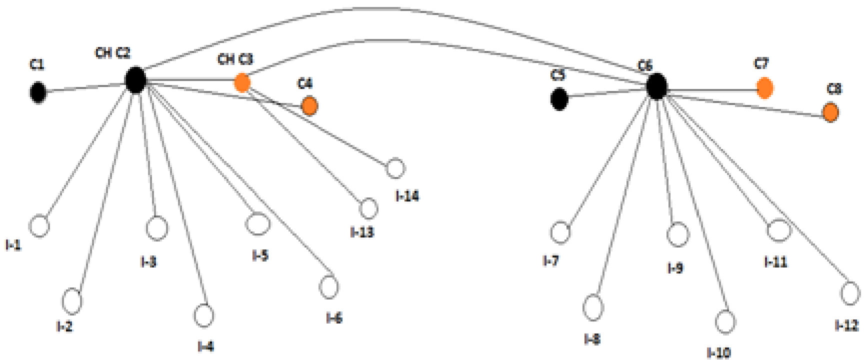


Fig. 2.

- In Fig. 2, Nodes I-13 and I-14 want to join cluster1 and will send JOIN packet to the secondary CH 'C3'. In return, C3 will establish connections with them as it has the required communication capacity and then it will acknowledge the request by sending a JOIN-ACCEPT packet to I-13 and I-14. Also, the CH adds the node information in its tables.
- Next let's consider a scenario where I-8 moves to a position where it finds that CHs of Cluster 1 are closer than that of Cluster 2. According to the energy conservation principle of our algorithm, it will now join Cluster 1.
- The power received by I-8 from CH 'C6' is below the minimum threshold value. So, a BYE message is sent to C6 which in turn acknowledges the request with a BYE ACCEPT. C6 also removed I-8 from its table.

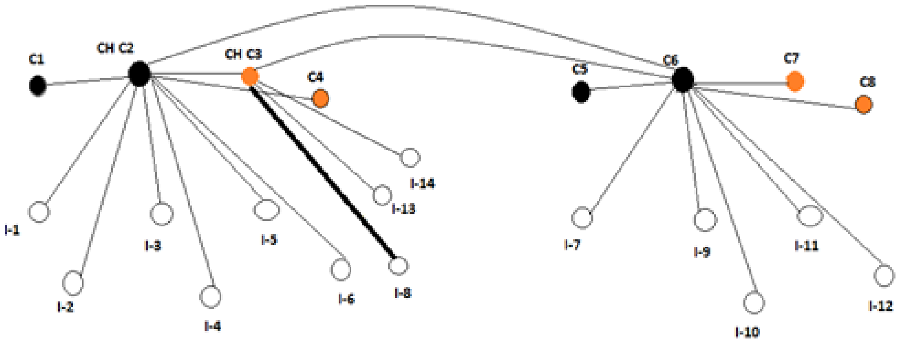


Fig. 3.

- I-8 will then send a JOIN packet to C3 and C2. Since C2 has no more capacity, C3 sends a JOIN-ACCEPT packet and I-8 becomes a part of Cluster 1 (Fig. 3)
- Since I-8 has left Cluster 2, C6 sends the UPDATE-CH packet after a fixed time interval containing this information about node I-8 to C2 and C3, so that they can update their tables. On the other hand, since I-8 after leaving Cluster 2 has joined Cluster 1, C3 will also send an UPDATE-CH to C6.
- In another scenario, let us suppose I-4 in Cluster-1 wants to send a message to I-10 in Cluster-2. I-4 will send a REQUEST message to C-2. Since destination is not in the same cluster, C-2 will send a REQUEST REPLY message to I-4 asking for the message it wants to send to I-10.

  I-4 sends all the information in NODE-MSG packet to C-2 which further passes on this information to C-6 through PASS-INFO packet. C-6 will look up its Intracluster table, get the location of the destination and use the shortest path to deliver the message in CH-MSG packet to I-10.

5 Characteristic Analysis

In this section we provide a characteristic analysis of the proposed algorithm along with some other representative algorithms that address the problems of stability, energy conservation and location management individually. The table provides comparison of performance and other characteristic features of each protocol.

It is seen that the Transmission overhead in the network is Highest in case of [1] because of the use of paging signals for energy conservation, and Lowest in case of [2] when the cluster size is not too large because of the availability of location of all nodes and therefore less location updates. Our proposed algorithm also doesn’t require high overhead because of the Location Tables with every CH. Also, Single & Double Phase Clustering Protocol [1] provides faster processing and computing power along with Less Call Drops. But, it works on static topology during the clustering process, which is quite unfeasible. KCLC & KCMBC Protocol [2] takes into account Duplicate copies and Delivery Loops by adding sequencing and in case of Inter-cluster, link failure the packet can still reach the destination as the new route information can be provided by intermediate CHs and so it acts as a Self-Recovery protocol. Again, this protocol assumes an environment in which direction and moving rate are considered constant. Therefore, the results that have been derived don’t model the real time scenarios that closely. In DDVC & DDLC Protocol [3] Cluster size can vary depending on the application and they manage CH Contention very well, but won’t work in cases where there are any changes in the direction of the nodes.

The proposed algorithm ELS takes into consideration the dynamic nature of the network and the direction and moving rate are not taken constant. In addition, the protocol uses a very efficient mechanism for CH management which ensures Energy and Bandwidth conservation along with fault free working of network. The tables managed by CHs help in achieving Less Delay in searching for a Location, thereby reducing overhead and allowing self recovery with the help of alternate redundant nodes. It is suggested that Single & Double Phase Clustering Protocol [1] takes care

Table 3. Characteristic Analysis

S.No	Characteristics	Single & Double Phase Clustering Protocol [1]	KCLC & KCMBC Protocol [2]	DDVC & DDLC Protocol [3]	ELS Protocol [4]
1	**Basis of Protocol**	Energy based	Location based	Stability based	Energy, Location & Stability based
2	**QOS Metric**	Power	Mobility	Mobility	Power & Mobility
3	**Route Discovery**	Proactive	Proactive	Reactive	Reactive
4	**Mobility Support**	Low	Medium	Medium	High
5	**Architecture**	Hierarchical multi hop	Flat k hop	Hierarchical single hop	Hierarchical multi hop
6	**Route Redundancy**	No	Yes	No	Yes
7	**Reclustering**	Required	Required	Not Required	Not Required
8	**Transmission Overhead**	High	Less	Medium	Medium
9	**Bandwidth Conservation**	Achieved to quite an extent	Limited	Limited	Achieved to quite an extent
10	**Stability**	No	No	Yes	Yes
11	**Energy Conservation**	Yes	No	No	Yes
12	**Maintenance**	No Mechanism	Effective Mechanism	Effective Mechanism	Effective Mechanism
14	**Call Drop Rate**	Very Low	Medium	Medium	Low
15	**Fault Tolerance**	Good	Excellent	Less	Good
16	**Environment**	Static Topology	Constant Direction & Moving Rate	Fixed Direction	Dynamic Topology
17	**Self Recovery**	To some extent	Yes	Yes	Yes

of the Energy Conservation aspect of Clustering, KCLC & KCMBC Protocol [2] works on Location Management, DDVC & DDLC Protocol [3] tries to achieve stable clustering and ELS protocol incorporates all three aspects of clustering by implementing Location services, Energy efficient CHs and high Stability. Thereby achieving higher QOS than the other protocols.

6 Conclusion

In this paper we have presented an improvised clustering protocol that gives better network performance by combining various QOS metrics together. In MANET scenarios the algorithm works on energy conservation and high stability of clusters with location management that provide efficient and fast routing along with quick recovery in case of Link Failures. Our scheme makes use of two sets of Cluster heads - Primary and Secondary with unique ids. Thresholds have been set for Power level, Capacity and Link duration to achieve Energy Conservation. All nodes are GPS enabled and every CH stores Intercluster and Intracluster table to manage the Location services. Updates are required to be sent only to the CHs, thereby reducing the pressure on Network Bandwidth. This algorithm is well suited for dynamic environment and scenarios where direction and moving rate are not constant. The proposed scheme shows better performance than the previous algorithms because it combines reduced overhead, self recovery, Less Delay in searching along with energy efficient nodes. In the future, we plan to further analyze the mobility and link duration characteristics of our proposed algorithm in multihop Ad-Hoc Environment.

References

1. Ryu, J.-H., Song, S., Cho, D.-H.: New Clustering Scheme for Energy Conservation in Two-Tiered Mobile Ad Hoc Networks. IEEE Transaction on Vehicular Technology 51(6), 1661–1668 (2002)
2. Leng, S., Zhang, L., Fu, H., Yang, J.: A Novel Location-Service Protocol Based on k-Hop Clustering for Mobile Ad Hoc Networks. IEEE Transaction on Vehicular Technology 56(2), 810–817 (2007)
3. Sakhaee, E., Jamalipour, A.: Stable Clustering and Communications in Pseudolinear Highly Mobile Ad Hoc Networks. IEEE Transaction on Vehicular Technology 56(8), 3769–3777 (2008)
4. Rodoplu, Meng, T.H.: Minimum energy mobile wireless networks. IEEE J. Select. Areas Commun. 17, 1333–1344 (1999)
5. Anupama, M., Sathyanarayana, B.: Survey of Cluster Based Routing Protocols in Mobile Ad Hoc Networks. International Journal of Computer Theory and Engineering 3(6), 806–815 (2011)
6. Yu, J.Y., Chong, P.H.J.: A survey of clustering schemes for mobile ad hoc networks. Commun. Surveys tuts. 7(1), 32–48 (2005)
7. Umamaheshwari, Radhamani, G.: Clustering Schemes for Mobile Adhoc networks: A Review. In: International Conference on Computer Communication and Informatics, pp. 1–6 (2012)
8. Camp, T., Boleng, J., Wilcox, L.: Location information services in mobile ad hoc networks. In: Proc. IEEE ICC, pp. 3318–3324 (2001)
9. Su, W., Lee, S.J., Gerla, M.: Mobility prediction in wireless networks. In: Proc. IEEE MILCOM, Los Angeles, CA, pp. 491–495 (2000)
10. Stojmenovic, I.: Position-based routing in ad hoc network. IEEE Commun. Mag. 40(7), 128–134 (2002)
11. Basu, P., Khan, N., Little, T.D.C.: A mobility based metric for clustering in mobile ad hoc networks. In: Proc. ICDCSW, Mesa, pp. 413–418 (2001)
12. Inspirenignite routing for MANETS, http://www.inspirenignite.com/routing-in-mobile-ad-hoc-networks/

Comparative Analysis of Contention Based Medium Access Control Protocols for Wireless Sensor Networks

Chandan Kumar Sonkar[1], Om Prakash Sangwan[2], and Arun Mani Tripathi[3]

[1] Department of Computer Science and Engineering,
Dr. K.N. Modi Institute of Engineering and Technology, Modinagar, Ghaziabad, India
[2] School of ICT Gautam Buddha University Greater Noida, India
[3] Computer Science and Engineering, Integral University Lucknow, India
{c.sonkar1986,amttheking}@gmail.com,
opsangwan@gbu.ac.in

Abstract. Wireless sensor "motes" is small or tiny embedded systems equipped with radios for wireless communication in the networks, which depend on batteries as a power source. The development of Medium Access Control (MAC) protocols which search is for to minimize the energy consumption in the wireless sensor network. Recent contention based MAC protocols reduce energy usage by placing the radio in a low power sleep state when not sending or receiving the message. In this paper the main emphasis is on the analysis of the Contention Based MAC Protocols and energy consumption in the networks. Based on the work done by various researchers conclude that S-MAC is the backbone of all the MAC protocols for wireless sensor networks. Our proposed work investigated the energy usage and compared the performance of IEEE 802.11 MAC protocol with S-MAC protocol on different modes like without periodic sleep and with periodic sleep on different performance metrics like Remaining Energy vs Time, Energy consumption vs global packet id and Average End to End Delay vs Time. The performance of S-MAC protocol improves on the basis of duty-cycle parameter, which determines the length of sleep period in a frame and this parameter is a variable. So changing the duty cycle will change the performance of S-MAC protocol. Finally, we have used the different routing protocol with S-MAC to evaluate the energy consumption. The experimented worked done on Network Simulator Ns2- 2.34.

Keywords: Medium Access Protocol, Wireless Sensor Network, Idle listening, Sleep State, S-MAC, Energy Consumption, Duty cycle.

1 Introduction

Wireless communications start from the late 1800s, when M.G. Marconi did the pioneer work establishing the first successful radio communication systems have been developing and evolving with a furious pace. In the early stages, wireless communication systems were dominated by military usages and supported accordingly to military needs and requirements. Over the past few years, the world

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 86–103, 2013.

has become increasingly mobile. As a result traditional ways of networking the world have proven inadequate to meet the challenges posed by our new collective lifestyle. Recent advances in processing, storage, and communication technologies have advanced the capabilities of small-scale and cost-effective sensor systems, which are composed of a single chip with embedded memory, processor, and transceiver.WSN has a great ability of obtaining data and it can work under any situation, at any time, in any place, which makes it useful in many important fields. So, the military department, industrial circle and academic circle of many countries all over the world are paying great attention to it. It also becomes a hot issue in research at home and abroad today, and it is regarded as one of the ten influencing technology in the 21st century [33].

1.1 Wireless Sensor Networks (WSNS)

Wireless sensor networks are consisting of thousands of extremely small and cheap devices that can sense the environment and communicate the data as required. Wireless sensor networks have emerged as one of the first real applications of ubiquitous computing. Sensor networks play a key role in bridging the gap between the physical and the computational world by providing reliable, scalable, fault tolerant and accurate monitoring of physical phenomena. A Wireless sensor network is defined as being composed of a large number of nodes, which are deployed densely in close proximity to the phenomenon to be monitored. As shown in fig. 1 many sensor nodes are scattered in a sensor field and each of these nodes collects data and its purpose is to route this information back to a sink [13]. The network must possess self-organizing capabilities since the positions of individual nodes are not predetermined. Cooperation among nodes is the dominant feature of this type of network, where groups of nodes cooperate to disseminate the information gathered in their vicinity to the user [28].

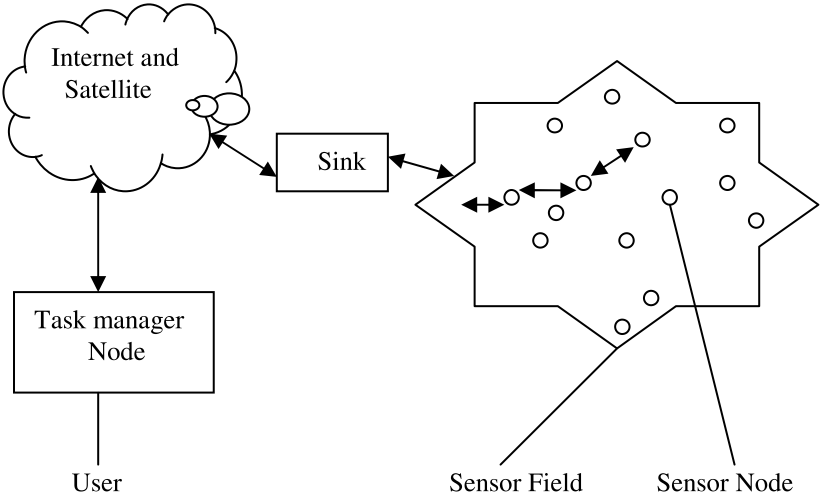


Fig. 1. Sensor nodes scattered in a sensor field

1.2 Features of Wireless Sensor Networks

The design of Wireless Sensor Networks is determined by the sensor nodes characteristics and its application. The important features of WSNs are discussed below.

Energy Limitations. In wireless sensor networks energy is the main issue, whiles the design of the wireless sensor networks. The chip devices which are used in the sensor nodes depend on the battery which provides the energy. Batteries are the most commonly used sources of energy. Thus, we need a mechanism polices for the efficient utilization of the energy resources for long time.

Resistance to Node Failure. The node failure is also the major cause of energy wastage in wireless sensor networks. WSN is dynamic systems and resistance to node failures. There may be changes in the network topology, may be caused by node failure due to various factors such as depleted batteries, environmental factors (fire, flood), an intruder's attack etc.

Scalability. Wireless Sensor Networks may contain hundreds or even thousands of sensor nodes. The WSN should be scalable, meaning that the performance of these networks should be minimally affected by a change in network size. In most of the cases, recharging or replacing batteries is not possible, and adding new sensor nodes is the only way to prolong the lifetime of the network.

Deployment. The deployment of can be in various ways it depend on the requirement, application and environmental condition. It can be deployed randomly over the monitoring field or sensor field. After deployment, the sensor nodes in most applications remain static. Depending on the deployment strategy, suitable communication protocols should be developed based on the existing network topology in order to support the WSN functionality.

Quality of Service (QoS). Quality of service is the most important parameter of the network which gives the information about the reliability of the networks. It means satisfying the application goals by meeting the quality of service requirements which is one of the basics requirements.

1.3 Mac Protocols

There are many existing MAC protocols for wireless sensor networks. In this section a wide range of the MAC protocols is listed with their comparison [10].

- Sensor-MAC (S-MAC)
- WiseMAC
- Traffic-Adaptive MAC Protocol (TRAMA)
- Data gathering-MAC (D-MAC)

- Timeout-MAC (T-MAC)
- Dynamic Sensor-MAC (DSMAC)

One fundamental task of the MAC protocol is to avoid collisions from interfering nodes. The MAC sub-layer uses MAC protocol to ensure that signals sent from different stations across the same channel don't collide. There are many MAC protocols that have been developed for wireless Voice and data communication networks.

Existing MAC protocols can be divided into two broad categories-

- Scheduled based protocols e.g. TDMA, FDMA, CDMA etc.
- Contention based protocols e.g. IEEE 802.11, CSMA etc.**Mac Protocol Design Consideration**

The medium access control protocols for the wireless sensor network have to achieve two objectives.

- The first objective is the creation of the sensor network infrastructure. A large number of sensor nodes are deployed and the MAC scheme must establish the communication link between the sensor nodes.
- The second objective is to share the communication medium fairly and efficiently. To design the efficient MAC protocol for wireless sensor networks, the following characteristics are to be considered [31].

Energy Efficiency. The energy is the most important factor in the wireless sensor nodes. The sensor nodes are battery powered and it is often very difficult to change or recharge batteries for these sensor nodes.

Latency. Latency requirement basically depend on the application in the sensor network.

Throughput. Throughput requirement are also varies with different application in the wireless sensor network.

Duty Cycling. Duty cycling is also one of the important mechanisms is used for energy efficient MAC protocol in sensor network.

1.5 Sources of Energy Consumption at the MAC Layer

From the point view of energy dissipation, four major sources of energy waste are caused by MAC layer problems [39]. Retransmission of process is due to the collision or congestion. In WSNs, all nodes are capable of transmitting data through the same broadcast channel. As a tiny communication device, each sensor node may have only one receiving antenna; therefore, if two or more transmissions from multiple sources arrive at the same time, a collision will happen, and none of transmitted packets can be received correctly.

Idle channel sensing In order to eliminate or reduce collisions, nodes must sense the channel continuously to obtain scheduling information or wait before sending data until the channel is detected idle. In either case, extra sensing energy is needed.

Overhearing is sharing a common wireless medium; the data transmitted by one node can reach all the other nodes within their transmission range. A node then may receive packets not destined for it. This is referred to as overhearing and it also wastes energy.

Overhead due to control messages, a lot of MAC protocols operate by exchanging control messages for signaling, scheduling, and collision avoidance, which will consume extra energy. Therefore, in order to design an energy-efficient MAC protocol, collisions must be avoided as much as possible. Many approaches have been proposed, but it is difficult to achieve all energy-conserving objectives at the same time.

2 Related Works

According to the work done by [13] in this they discussed the present communication architecture for sensor networks and proceed to survey the current research pertaining to all layers of the protocol stack that is physical, data link, network, transport and application layers. They defined sensor network as being composed of a large number of nodes, which are deployed densely in close proximity to the phenomenon to be monitored. Each of these nodes collects data and its purpose is to route this information back to a sink. They propose that sensor network must possess self-organizing capabilities since the positions of individual nodes are not predetermined [17]. The author [3] examines that how CSMA based medium access can be adapted for sensor networks. However, these approaches are not directly applicable due to the following characteristics of sensor networks-

- Network operates as a collective structure
- Traffic tends to be periodic and highly correlated
- Equal cost per unit time for listening, receiving and transmitting

The authors outline a CSMA-based MAC and transmission control scheme to achieve fairness while being energy efficient. The adaptive rate control proposed uses loss as collision signal to adjust transmission rate in a manner similar to the congestion control in TCP [3].

In 2002 another author [39] gives the novel technique about S-MAC, a medium-access control (MAC) protocol designed for wireless sensor networks. S-MAC uses three novel techniques to reduce energy consumption and support self-configuration. To reduce energy consumption in listening to an idle channel, nodes periodically sleep. Neighboring nodes form virtual clusters to auto-synchronize on sleep schedules.

Inspired by PAMAS, S-MAC also sets the radio to sleep during transmissions of other nodes. Unlike PAMAS, it only uses in-channel signaling. Finally, S-MAC applies message passing to reduce contention latency for sensor-network applications that require store-and-forward processing as data move through the network. Finally the authors point out that the experiment results show that, on a source node, an 802.11-like MAC consumes 2–6 times more energy than S-MAC [39]. The authors [40] include significant extensions in the protocol design, implementation, and experiments of S-MAC work. This paper presents S-MAC, a medium access control protocol specifically designed for wireless sensor networks. Energy efficiency is the primary goal in the protocol design. Low-duty-cycle operation of each node is achieved by periodic sleeping. This paper proposes adaptive listening, which largely reduces such cost for energy savings. It enables each node to adaptively switch mode according to the traffic in the network [40].

According to the author, Huan Pham, A new adaptive mobility-aware Sensor MAC protocol (MS-MAC) for mobile sensor applications. In S-MAC protocol, a node detects its neighbor's mobility based on a change in its received signal level from the neighbor, or a loss of connection with this neighbor after a timeout period. By propagating mobility presence information, and distance from nearest border node, each node learns its relative distance from the nearest mobile node and from nearest border node. Depending on the mobile node movement direction, the distances from mobile and border nodes, a node may trigger its neighbor search mechanism to quicken the connection setup time [12].

The author critically evaluated the topology changes and presents a mobility-adaptive, collision-free MAC protocol for mobile sensor networks. MMAC caters for both weak mobility (e.g. topology changes, node joins and node failures) and strong mobility (e.g. concurrent node joins and failures, and physical mobility of nodes). Finally authors point out that this protocol adapts the time frame, transmission slots, and random-access slots according to mobility [26].

The author Zhiwei Zhao et al. states that at present, most MAC protocols use the same transmission power when sensor nodes send packets. However, the deployment of the sensor nodes is asymmetrical in wireless sensor networks, which will bring more energy consumption and unnecessary collisions. This paper, proposed a transmission power control protocol for WSNs based on SMAC protocol. Power control at the MAC layer selects the minimum amount of transmitting energy needed to exchange messages between any pair of neighboring nodes. The simulation results show that, compared with SMAC protocol, proposed protocol has improved a lot in the delay of packets, reception rate, energy consumption and throughput of the networks [44]. R.Yadav and S. Verma present the challenges in the design of the energy efficient medium access control (MAC) protocols for the wireless sensor network. Authors describe several MAC protocols for the WSNs emphasizing their strength and weakness wherever possible. Finally, discuss the future research directions in the MAC protocol design [33].

Authors [43] have provided some good comparisons on some of the prominent protocols that use power management mechanism topology control. The key idea of power control is that, instead of transmitting using the maximum power, nodes in a

WSN collaboratively determine their transmission power while preserving some required properties. The basic idea of sleep scheduling is to save energy by putting redundant nodes into the sleeping mode.

A. Roy and N. Sharma critically evaluate the different parameter for saving the energy in wireless sensor network. Wireless sensor networks have been widely used in many important fields such as target detection and tracking, environmental monitoring, industrial process monitoring, and tactical systems. As nodes in wireless sensor networks typically operate unattended with a limited power source, energy efficient operations of the nodes are very important. Although energy conservation in communication can be performed in different layers of the TCP/IP protocol suit, energy conservation at MAC layer is found to be the most effective one due to its ability to control the radio directly. In this author have investigate the available energy-efficient MAC protocols for sensor networks emphasizing their energy saving methods [1]. In this they give the energy model in MAC Layer of wireless sensor network and different parameter used in the ns 2 simulation Trace file.

The literature survey gives the details about the research area and briefly discusses the paper with its conclusion. After doing the literature survey we came to know about the major sources of energy wastages and different routing protocols. It is found that there are many existing MAC protocols for wireless sensor networks including S-SMAC. S-MAC is the most popularly used MAC protocols for wireless sensor networks. In this still there is open issues discuss by the researcher even found in the survey about the energy and their causes like idle listening, collusion , overhearing , problem in synchronization and message passing. So, this paper would take S-SMAC as the problem area and we will discuss in details.

3 S-Mac Protocols

S-MAC protocol was proposed by SCADDS project group at USC/ISI (Scadds, available online). S-MAC is most popularly used protocol designed specifically for WSN. S-MAC is designed aiming at the requirement of saving energy of WSN according to 802.11 MAC. The main goal of S- MAC protocol is to reduce energy consumption, while supporting good scalability and collision avoidance [39]. This protocol tries to reduce energy consumption from all the sources that have been identified to cause energy waste, i.e., idle listening, collision, overhearing and control overhead.

3.1 Periodic Listen and Sleep

As stated above, in many sensor network applications, nodes are idle for long time if no sensing event happens. Given the fact that the data rate is very low during this period, it is not necessary to keep nodes listening all the time. S-MAC reduces the listen time by putting nodes into periodic sleep state. The basic scheme is shown in Fig.2. Each node sleeps for some time, and then wakes up and listens to see if any other node wants to talk to it. During sleeping, the node turns off its radio, and sets a timer to awake itself later. All nodes are free to choose their own listen/sleep schedules.

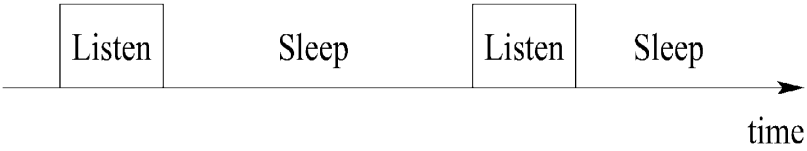


Fig. 2. Periodic listens and sleeps

3.2 Collision Avoidance

If multiple neighbors want to talk to a node at the same time, they will try to send when the node starts listening. In this case, they need to contend for the medium. Among contention protocols, the 802.11 does a very good job on collision avoidance. S-MAC follows similar procedures, including virtual and physical carrier sense, and the RTS/CTS exchange for the hidden terminal problem [37].There is a duration field in each transmitted packet that indicates how long the remaining transmission will be. If a node receives a packet destined to another node, it knows how long to keep silent from this field. The node records this value in a variable called the Network Allocation Vector (NAV) [42].

Carrier sense time is randomized within a contention window to avoid collisions and starvations. The medium is determined as free if both virtual and physical carrier sense indicates that it is free. All senders perform carrier sense before initiating a transmission. If a node fails to get the medium, it goes to sleep and wakes up when the receiver is free and listening again. Broadcast packets are sent without using RTS/CTS. Unicast packets follow the sequence of RTS/CTS/DATA/ACK between the sender and the receiver. After the successful exchange of RTS and CTS, the two nodes will use their normal sleep time for data packet transmission. They do not follow their sleep schedules until they finish the transmission.

3.3 Adaptive Listening

The scheme of periodic listen and sleep is able to significantly reduce the time spent on idle listening when traffic load is light [39]. However, when a sensing event indeed happens, it is desirable that the sensing data can be passed through the network without too much delay. When each node strictly follows its sleep schedule, there is a potential delay on each hop, whose average value is proportional to the length of the frame.

S-MAC follows an important technique, called adaptive listen [40] to improve the latency caused by the periodic sleep of each node in a multi-hop network. The basic idea is to let the node who overhears its neighbor's transmissions (ideally only RTS or CTS) wake up for a short period of time at the end of the transmission. In this way, if the node is the next-hop node, its neighbor is able to immediately pass the data to it instead of waiting for its scheduled listen time. If the node does not receive anything during the adaptive listening, it will go back to sleep until its next scheduled listen time.

3.4 Message Passing

This section describes how to efficiently transmit a long message in both energy and latency. A message is the collection of meaningful, interrelated units of data. The receiver usually needs to obtain all the data units before it can perform in network data processing or aggregation. The disadvantages of transmitting a long message as a single packet are the high cost of re-transmitting the long packet if only a few bits have been corrupted in the first transmission. However, if we fragment the long message into many independent small packets, we have to pay the penalty of large control overhead and longer delay. It is so because the RTS and CTS packets are used in contention for each independent packet. This protocol fragments the long message into many small fragments, and transmits them in a burst. Only one RTS and one CTS are used. They reserve the medium for transmitting all the fragments. Every time a data fragment is transmitted, the sender waits for an ACK from the receiver. If it fails to receive the ACK, it will extend the reserved transmission time for one more fragment, and re-transmit the current fragment immediately.

4 Research Methodologies

The simulation methodology used to perform our experiment work on Network Simulator Ns2-2.34. This work would analyze the performance of MAC and S-MAC protocol on different parameters with the NS-2 (Network Simulator-2) version 2.34 is chosen for simulation purpose.

4.1 Proposed Work

The proposed work would compared the performance of IEEE 802.11 MAC protocol with S-MAC protocol on different parameters that is without periodic sleep and with periodic sleep on different performance metrics like remaining Energy vs Time, Energy consumption vs Global packet id and Average End to End Delay vs Time. And determine the length of sleep period in a frame using duty cycle parameter to show the fundamental tradeoffs on energy and latency. Finally, investigate how the performance of S-MAC protocol improves on the basis of duty-cycle and different routing protocol to save more energy, through the simulation studies.

5 Results and Discussion

5.1 Simulation Environment

This section presents the topology and different parameters used in the simulation process. Fig.3 shows the topology which is having 11 nodes with one source and one sink. The first node is the source and the last node is the sink which is static and of 10 hop linear network.

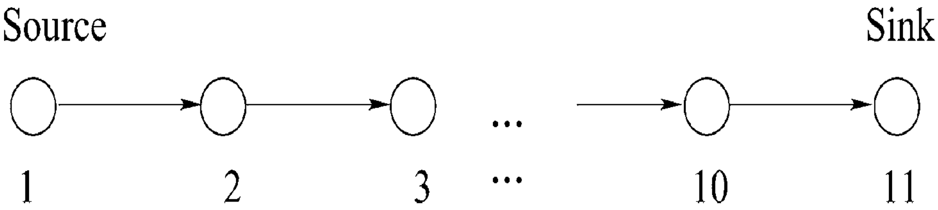


Fig. 3. 10-Hop linear network with one source and one sink

This simulation process considered a wireless network of 11 static nodes which are placed within a 500m x 500m area. CBR (constant bit rate) traffic is generated among the nodes. The simulation runs for 100 Seconds. Table 1 shows the important simulation parameters used in the simulation process.

Table 1. Important Simulation Parameters

Parameters	Values
Simulation time	100 Sec
Simulation area	500m x 500m
Antenna	Omni antenna
No. of nodes	11
Packet size	512 Bytes
Max queue length	50
Traffic	CBR
Routing protocol	AODV
Energy	100j
Idle Power	1j
Rx Power	1j
Tx Power	1j
SMAC duty cycle	10 %

5.2 Experimented Results

Remaining Energy

The Value of MAC protocol Trace from Tr. file generated by the Ns2 simulator. We have analyzed and evaluated the trace file and trace the value of remaining energy and compare the energy of MAC and S-MAC without sleep and with sleep with same time approximately. Fig. 4 shows the measured remaining energy at the node and energy consumption in network with time changing. In this case, the graph shows the remaining energy in the network at the node. In this we have experimented that 802.11 MAC uses more than twice the energy used by S-MAC with periodic sleep. Since idle listening rarely happens, energy savings from periodic sleeping is very limited. S-MAC achieves energy savings mainly by avoiding overhearing and efficiently transmitting long messages.

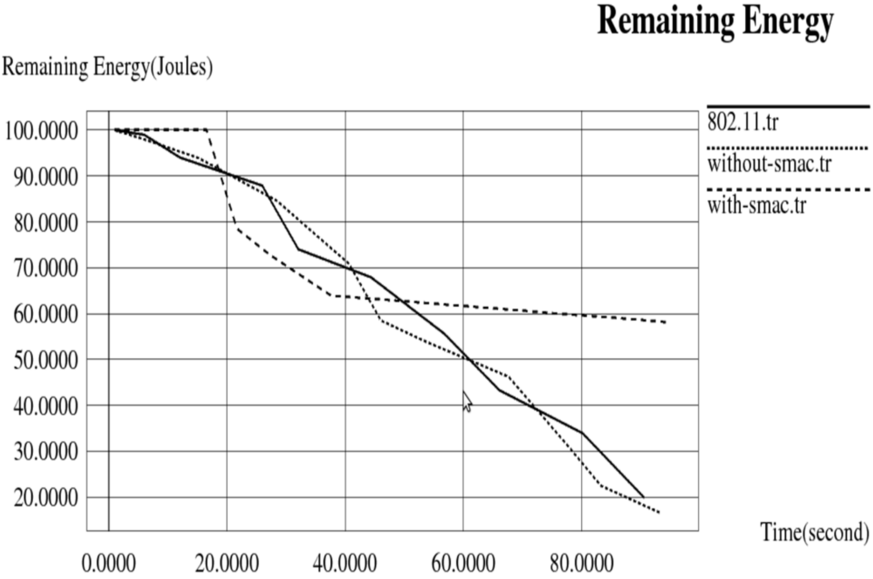


Fig. 4. Remaining Energy vs Time in Network

The S-MAC protocol with periodic sleep has the best energy performance than 802.11 MAC. S-MAC without periodic sleep also performs better than 802.11MAC. However, as shown in the figure, when idle listening dominates the total energy consumption, the periodic sleep plays a key role for energy savings.

Energy Consumption at Source Node

We change the traffic load by varying the inter-arrival period of messages. If the message inter-arrival period is 10s, a message is generated every 10s by each source node. In this experiment, the message inter-arrival period varies from 1 to 10s. Duty cycle parameter for S-MAC with periodic sleep mode is kept 10%. For each traffic pattern, we have done five independent tests when using different MAC protocols. In fig.5 we have shown the consumed energy from the trace file of MAC and SMAC without sleep and with sleep. So, from the figure it has been concluded that SMAC consume less energy than the MAC protocol which prolong the sensor node life for long time. Fig.5 shows the measured average energy consumption at source node in network. In above case, the S-MAC with periodic sleep protocol has the best energy performance, and far outperforms IEEE-802.11 MAC and S-MAC without periodic sleep.

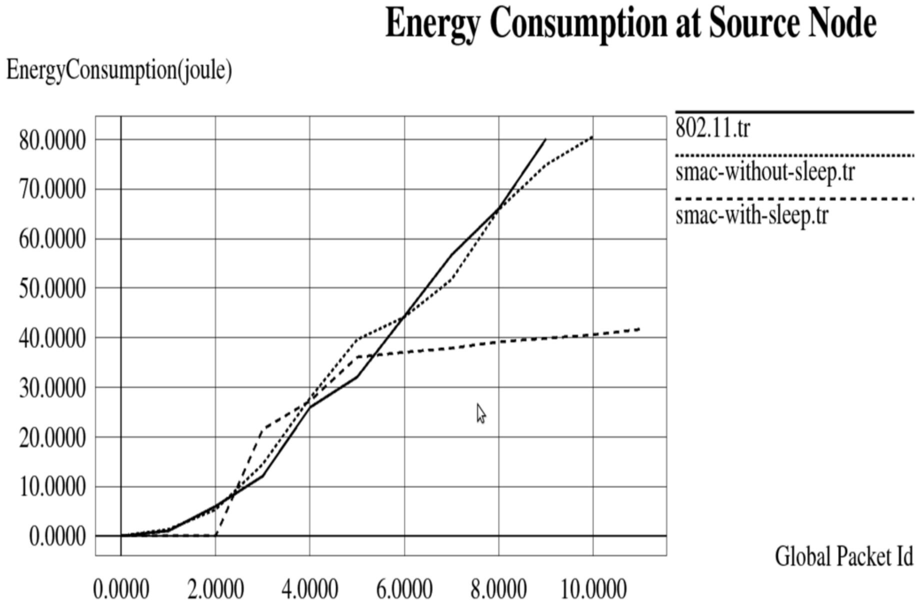


Fig. 5. Energy Consumption on Radios at Source Node

Measurement of Average End to End Delay

In fig.6 we have evaluated delay result from trace file. Since S-MAC makes the tradeoff of latency for energy savings, we expect that it can have longer latency in a multi-hop network due to the periodic sleep on each node. To quantify latency and measure the benefits of S-MAC, we use the same ten-hop network topology. In fig.6 we measure latency of IEEE-802.11 and S-MAC protocols.

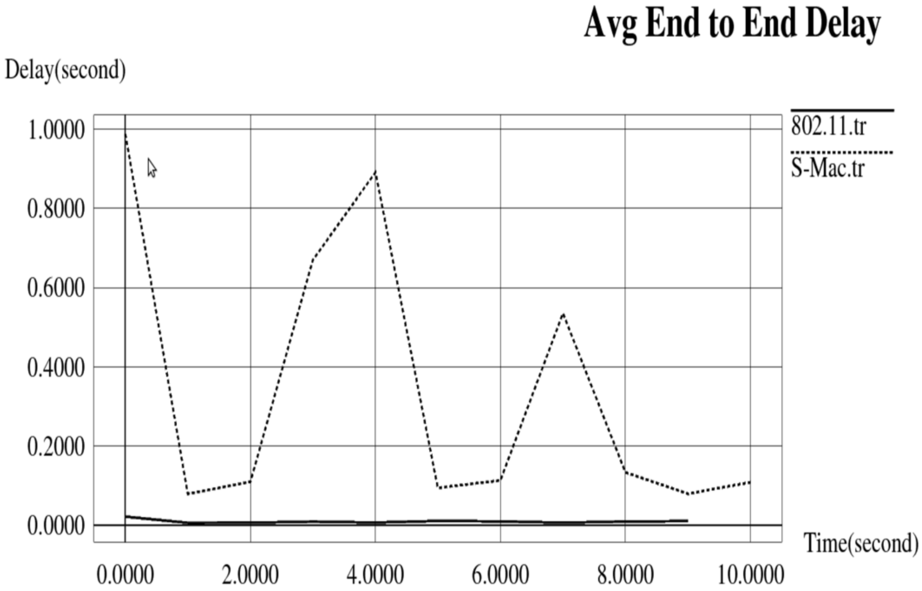


Fig. 6. Average Delay vs Time

5.3 Modification for Heavy Traffic Load With 20% Duty Cycle

This section presents the topology and different parameters used in the simulation scenario in Table 2. In this, the topology is having of two nodes with one source and one sink. This simulation process considered a wireless network of two static nodes, CBR (constant bit rate) traffic is generated among the nodes. The simulation runs for 50 seconds, used for heavy traffic load with 20% duty cycle. Table 2 shows the important simulation parameters used in the simulation process. In this we have changed the different parameter and analyze the trace file value and its performance with the duty cycle.

Table 2. Important Simulation Parameters with 20%Duty Cycle

Parameters	Values
Simulation time	50 Sec
Simulation area	500m x 500m
Antenna	Omni antenna
No. of nodes	02
Packet size	512 Bytes
Max queue length	50

Table 3. (*Continued*)

Parameters	Values
Traffic	CBR
Routing protocol	AODV
Energy	50j
Idle Power	1j
Rx Power	1j
Tx Power	1j
SMAC duty cycle	20 %

Remaining Energy with 20% Duty Cycle

We evaluate the value with the last second of the simulation that how much energy remains in the sensor node. Fig.7 shows the measured remaining energy in network with time changing. In this case, 802.11 MAC uses more than twice the energy used by S-MAC with periodic sleep. Since idle listening rarely happens, energy savings from periodic sleeping is very limited. S-MAC achieves energy savings mainly by

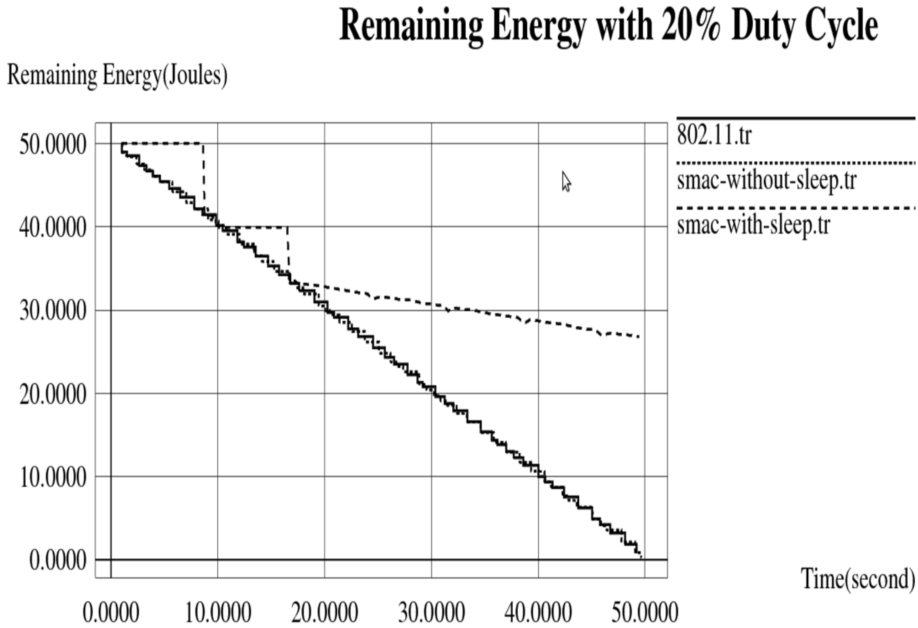


Fig. 7. Remaining Energy vs Time in Network

avoiding overhearing and efficiently transmitting long messages. The graph below show the remaining energy with 20% duty cycle which shows that S-Mac with sleep state consume less energy and save energy for long life of the sensor node. 802.11 MAC protocol and S-MAC without sleep state shows the common remaining energy.

Performance Evaluation Using Different Routing Protocol with SMAC Protocol
The energy consumption is also depending on the different routing protocol. In this simulation we have used the routing protocol AODV,DSDV and DSR with S-MAC.

Remaining Energy

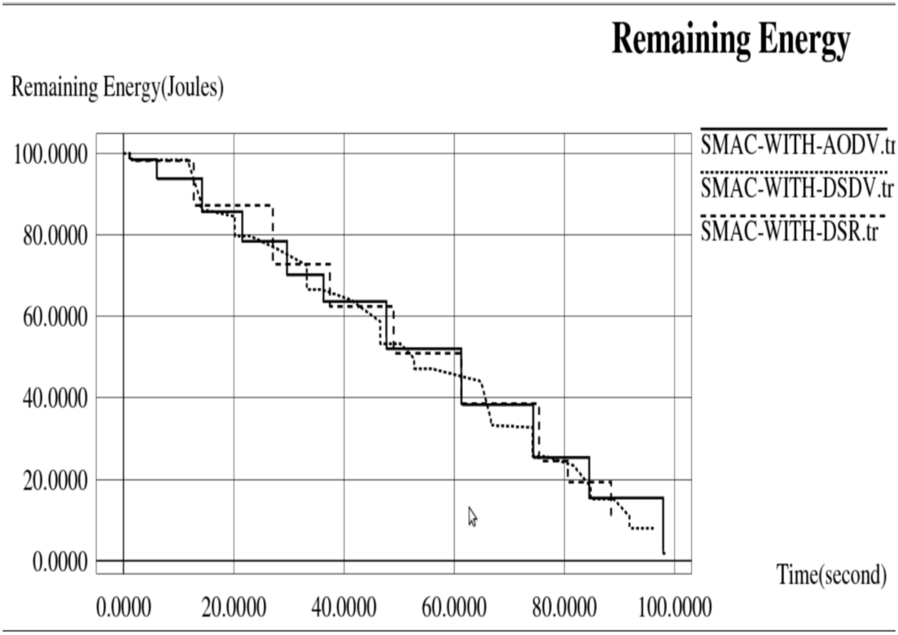


Fig. 8. Remaining Energy with Time between AODV, DSR and DSDV routing protocol

In the above simulated graph we analyze the remaining energy in the sensor node with different routing protocol. In this we have simulated SMAC protocol with AODV routing protocol. Then we analyze the S-MAC protocol with DSDV routing protocol and we analyze that the remaining energy with DSR routing which result is slightly different from AODV on different simulation time. Finally we analyze that the S-MAC protocol with DSDV and AODV which gives better energy saving.

Form fig.8, we can conclude that SMAC with DSR and SMAC with DSDV give the similar result with minor difference. The remaining energy of SMAC with AODV remains the more energy for long life of the sensor node.

6 Conclusion

The energy is most important to run the wireless sensor nodes in network for long time. In this paper we have discussed the different issues of energy wastage and energy consumption parameter. We discuss how the wireless network and wireless sensor networks consume energy in transferring the message. The main problem is collision in S-MAC protocols which have been removed with the mechanism sleep state for some time and wake up when it sense the message or data. S-MAC and MAC protocol discuss in details, it has been concluded that S-MAC is the most popularly used MAC protocol for wireless sensor networks. In our work we have compared the performance of IEEE 802.11 MAC protocol with S-MAC protocol on energy consumption in network. S-MAC obtains significant energy savings compared with 802.11 IEEE protocol without sleeping, which is clearly discussed in the experimented result section.

It has been found that at the end of the simulation time we see 5% to 15% energy saving and some time it more or less. It depends on the design of the network scenario. The simulation conclusion shows that the S-MAC with periodic sleep obtains more energy savings compared with IEEE-802.11 protocol and S-MAC without periodic sleeping protocol. The duty cycle parameter plays an important energy saving mechanism which is variable and change up to 20%, which determines the length of sleep period in a frame. It is able to greatly prolong the network lifetime, which is critical for real world sensor network applications. Periodic sleeping provides excellent energy performance at light traffic load. It makes S-MAC with periodic sleep and adaptive listening ideal for sensor networks where traffic is intermittent. Finally, we compare the S-MAC protocol with different routing protocols like DSR, DSDV and AODV.

References

1. Roy, A., Sarma, N.: Energy Saving in MAC Layer of Wireless SensorNetworks. In: National Workshop in Design and Analysis of Algorithm (NWDAA), Tezpur University, India (2010)
2. Sinha, A., Chandrakasan, A.: Dymamic power management in wireless sensor networks. IEEE Design Test Computers 18(2), 62–74 (2001)
3. Woo, A., Culler, D.: A Transmission Control Scheme for Media Access in Sensor Networks. In: Proc. ACM MobiCom 7th Annual International Conference on Mobile Computing and Networking, New York (2001)
4. Goldsmith, A.J., Wicker, S.B.: Design challenges for energy-constrained ad hoc wireless networks. IEEE Wireless Communication (2002)
5. Autonomic computing, `http://w3.ibm.com/autonomic/index.shtml`

6. Yahya, B., Ben-Othman, J.: Towards a classification of energy aware MAC protocols for wireless sensor networks. Wireless Communications and Mobile Computing 9(12), 1572–1607 (2009)
7. Murthy, C.S.R., Manoj, B.S.: Ad Hoc Wireless Networks: Architectures and Protocols. Prentice Hall PTR, New York (2004)
8. Suh, C., Ko, Y.B.: A traffic aware, energy efficient MAC protocol for wireless sensor networks. In: The Proceeding. of the IEEE International Symposium on Circuits and Systems, pp. 2975–2978 (2005)
9. Estrin, D., Govindan, R., Heidemann, J.: Embedding the Internet. Commun. ACM 43(5), 39–41 (2006)
10. Demirkol, I., Ersoy, C., Alagoz, F.: MAC protocols for wireless sensor networks: a survey. IEEE Communications Magazine (4), 115–121 (2006)
11. Tolle, G., Culler, D., Hong, W.: A macroscope in the redwoods. In: Proceedings of the 3rd International Conference on Embedded Networked Sensor System, San Diego (2005)
12. H. Pham and S. Jha.: Addressing Mobility in Wireless Sensor Media Access Protocol. In Proc. IEEE Intelligent Sensors, Sensor Networks and Information Processing Conference, Melbourne, pp.113-115 (2004), `http://op.inria.fr/mistral/personnel/Eitan.Altman/COURS-NS/n3.pdf`
13. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on Sensor Networks. IEEE Communications Magazine 40(8), 102–114 (2002), `http://citeseer.ist.psu.edu/akyildiz02survey.html`
14. Downard, I.T.: Simulating Sensor Network. Naval Research laboratory (2004)
15. Vasilescu, I., Kotay, K., Rus, D., Dunbabin, M., Corke, P.: Data collection, storage, retrieval with an underwater sensor network. In: Proceedings of the Third International Conference on Embedded Networked Sensor Systems (Sensys), San Diego (2005)
16. Akyildiz, I.F., Stuntebeck, E.P.: Wireless underground sensor networks: research challenges. Ad-Hoc Networks 4, 669–686 (2006)
17. Akyildiz, I.F., Melodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Computer Networks 51, 921–960 (2007)
18. Hill, J., Culler, D.: A wireless embedded sensor architecture for system level optimization. University of California Berkeley Technical Report (2002)
19. Heidemann, J., Li, Y., Syed, A., Wills, J., Ye, W.: Underwater sensor networking: research challenges and potential applications. In: Proceedings of the Technical Report ISI-TR-2005-603. USC/ Information Sciences Institute (2005)
20. Xiao, J., Yu, F.: An Efficient Transmission Power Control Algorithm in Wireless Sensor Networks. In: Proc. IEEE International Conference on Wireless Communications, Networking and Mobile Computing WiCom 2007, pp. 2767–2770 (2007)
21. Langendoen, K.: Medium access control in wireless sensor networks. In: Medium Access Control in Wireless Networks, vol. 2, Nova Science Publishers, Huntington (2007)
22. Chhari, L., Kamoun, L.: Wireless sensors networks MAC protocols analysis. Journal of Telecommunications (1), 42–48 (2010)
23. LAN-MAN Standards Committee of the IEEE Computer Society. Wireless LAN medium access control (MAC) and physical layer (PHY) specification, IEEE Std. 802.11-1997 edn. IEEE, New York (1997)
24. Deliang, L., Fei, P.: Energy-efficient MAC protocols for Wireless Sensor Networks. Beihang University, Beijing (2009)
25. Al Ameen, M., Riazul Islam, S.M., KyungsupKwak: Energy saving mechanism for Mac protocol in wireless sensor network. International Journal of Sensor Distributed Sensor Network (2010)

26. Ali, M., Uzmi, Z.A.: Medium access control with mobility-adaptive mechanisms for wireless sensor networks. Int. J. Sensor Networks, 134–142 (2006)
27. Li, M., Yang, B.: A survey on topology issues in wireless network. In: Proceedings of the International Conference on Wireless Networks (ICWN 2006), Las Vegas, Nev, USA (2006)
28. Papageorgiou, P.: Literature Survey on Wireless Sensor Networks (2003), `http://www.cs.umd.edu/users/pavlos/papers/unpublished/papageorgiou03sensors.pdf`
29. Sun, P., Zhang, X., Dong, Z., Zhang, Y.: A Novel Energy Efficient Wireless Sensor MAC Protocol. In: Proc. IEEE Fourth International Conference on Networked Computing and Advanced Information Management NCM 2008, Gyeongju, pp. 68–72 (2008)
30. Sivaraman, R., Sharma, V.R., Aarthy, V., Kavitha, K.: Energy comparison for cluster based environment in wireless sensor network. International Journal Trends in Engineering 2 (2009)
31. Yadav, R., Verma, S., Malaviya, N.: A Survey of MAC Protocols for Wireless Sensor Network. Procedingof UBCI Journal, 827–832 (2009)
32. Kumar, S., Raghavan, V.S., Deng, J.: Medium access control protocols for ad hoc wireless networks: a survey. Ad Hoc Networks 4(3), 326–358 (2006)
33. Wen-miao, S., Yan-ming, L., Research, Z.: on SMAC protocol for WSN. In: Proc. IEEE 4th International Conference on Wireless Communications, Networking and Mobile Computing WiCOM, pp. 1–4 (2008)
34. SCADDS: Scalable Coordination Architectures for Deeply Distributed Systems, `http://www.isi.edu/scadds/projects/S-MAC`
35. Sinha, Chandrakasan, A.: Dymamic power management in wireless sensor networks. IEEE Design Test Computers 18(2), 62–74 (2001)
36. Rajendran, V., Obraczka, K., Garcia-Luna-Aceves, J.J.: Energy Efficient Collision Free Medium Access Control for Wireless Sensor Networks. In: ACM International Conference on Embedded Networked Sensor Systems (SenSys), pp. 181–192 (2003)
37. Bharghavan, V., Demers, A., Shenker, S., Zhang, L.: MACAW: A media access protocol for wireless lans. In: Proc. ACM SIGCOMM, London, U.K. (1994)
38. Liao, W.H., Wang, H.H.: An adaptive energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of the1st International Conference on Embedded Networked Sensor Systems, pp. 171–180 (2003)
39. Ye, W., Heidemann, J., Estrin, D.: An energy-efficient Mac protocolfor wireless sensor networks. In: Proc. IEEE INFOCOM, New York (2002)
40. Ye, W., Heidemann, J., Estrin, D.: Medium Access Control With Coordinated Adaptive Sleeping for Wireless Sensor Networks. IEEE/ACM Transactions on Networking, 493–506 (2004)
41. Wang, L.C., Wang, C.W., Liu, C.M.: An adaptive contention window-based cluster head election mechanism for wireless sensor networks. In: Proc. of the IEEE Vehicular Technology Conference (2005)
42. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11 (1999)
43. Zhang, X., Ding, X., Lu, S., Chen, G.: Principles for energy-efficient topology control in wireless sensor networks. In: Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (2009)
44. Zhao, Z.W., Zhang, X.M., Sun, P., Liu, P.X.: A transmission power control MAC protocol for wireless sensor networks. In: Proc. IEEE 6th International Conference on Networking, Martinique, pp. 5–5 (2007)

SEP-E (RCH): Enhanced Stable Election Protocol Based on Redundant Cluster Head Selection for HWSNs

Raju Pal[1,*], Ritu Sindhu[1], and Ajay K Sharma[2]

[1] Dept. of Computer Science and Engineering, Galgotias University, Gr. Noida
[2] Dept of Computer Science and Engineering, NIT, Jaladhar
raju.pal@galgotiasuniversity.edu.in

Abstract. In this paper, an improved redundant cluster head selection mechanism for SEP-E has been proposed to balance the network load and to extend the network life time. In order to select reasonable cluster heads, our scheme first select an initial cluster head and a tentative or redundant cluster head in every cluster at each round. The node which has higher remaining energy and minimum mean distance is elected as cluster head of current round. Simulation results illustrate that SEP-E (RCH) prolongs the network life effectively, the time which first node dies, half of the node dies and last node dies is longer than SEP-E.

Keywords: Wireless sensor network, clustering, heterogeneity, Energy Efficient.

1 Introduction

With the advances in the technology of micro-electromechanical system (MEMS), developments in wireless communications and wireless sensor networks have also emerged [1]. The last few years have seen an increased interest in the potential use of wireless sensor networks (WSNs) in various fields like disaster management, battle field surveillance, and border security surveillance [2,3,4]. In such applications, a large number of sensor nodes are deployed, which are often unattended and work autonomously. Sensor nodes are typically powered by batteries with a limited lifetime, and in most cases, the batteries cannot be recharged. The energy problem in WSNs remains as one of the major barriers that prevent the complete exploitation of this technology. Clustering is a key technique used to extend the lifetime of a sensor network by reducing energy consumption [5]. It can also increase network scalability.

In recent years, Most studies of Clustered topology [6] schedule in Wireless Sensor networks (WSNs) such as LEACH[7], MPTC[8], AToM[9], GENSEN[10] aimed at the homogeneous sensor networks (the nodes of the sensor network equipped with the same amount of energy). Even for the wireless sensor networks composed of same type of nodes, the new nodes are arranged on the basis of the old ones, in order to

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 104–114, 2013.

prolong the networks life. The nodes newly added have more energy than the old ones. On the other hand, it is impossible for every node to use its own energy uniformly, because of the failure of wireless link or other accidents. It is necessary to design clustered topology suited to heterogeneous sensor networks. In this paper, we propose a redundant and energy efficient cluster head selection protocol that significantly increases the lifetime and reliability of the network. The selection criteria for cluster heads are weighted by remaining energy of a node and the minimum mean distance from all the ordinary nodes to the cluster heads in the network. Our simulation results justified that our algorithm provides longer stability period than SEP-E.

The rest of the paper is organized as follows. We briefly review related work in section 2. We then describe the radio model and network model used by our protocol in section 3. In section 4, we present our SEP-E (RCH) protocol. Section 5 describe the performance matrices over with network is analyzed. Section 6 provides simulation results and analysis. We conclude the paper in section 7.

2 Related Work

Clustering techniques have been employed to deal with energy management in WSNs. In [7], Low Energy Adaptive Clustering Hierarchy (LEACH), a clustering based protocol that utilizes randomized rotation of local cluster base station (cluster-heads) to evenly distribute the energy load among the sensors in the network was proposed. These sensors organize themselves into clusters using a probabilistic approach to randomly elect themselves as heads in an epoch. However, LEACH protocol is not heterogeneity-aware, in the sense that when there is an energy imbalance between these nodes in the network, the sensors die out faster than they normally should have if they were to maintain their energy uniformly. In real life situation it is difficult for the sensors to maintain their energy uniformly, thus, introducing energy imbalances. LEACH assumes that the energy usage of each node with respect to the overall energy of the system or network is homogeneous. Conventional protocols such as Minimum Transmission Energy (MTE) and Direct Transmission (DT) [13] do not also assure a balanced and uniformly use of the sensor's respective energies as the network evolves.

Stable Election Protocol (SEP) [11] was proposed for the two-level heterogeneous wireless sensor networks, which is composed of two types of nodes according to the initial energy. The advance nodes are equipped with more energy than the normal nodes at the beginning. SEP prolongs the stability period, which is defined as the time interval before the death of the first node. In [14], an extension of SEP i.e. Enhanced Stable Election Protocol (SEP-E) is presented to properly distribute energy and ensure maximum network life time. It operates in a WSN under three-level heterogeneity.

The proposed algorithm uses conditional probability and minimum mean distance to choose cluster head and redundant cluster head. On one hand cluster heads are distributed in a more reasonable manner, the energy consumption is decreased and the network life time is extended. On other hand, this protocol solves the problem of cluster head failure and damage by attackers. The reliability and security has thus been improved.

3 Radio Energy and WSN Model

We have used the energy model as in [7, 14]. In this energy model, E_{elec}^{Tx} and E_{elec}^{Rx} are defined as the energy being dissipated to run the transmitter's or receiver's circuitry, respectively, to send or receive one bit of the data packet. ε_{amp} represents the energy dissipation of the transmission amplifier to convey one bit of the data packet to the receiver node with a distance of d. As such, transmit (E_{Tx}) and receive (E_{Rx}) energies are calculated as follows [7]:

$$E_{Tx}(l,d) = \begin{Bmatrix} l.E_{elec}^{Tx} + l.\varepsilon_{amp}\,.d^2 & if\ d < do \\ l.E_{elec}^{Tx} + l.\varepsilon_{amp}\,.d^4 & if\ d \geq do \end{Bmatrix} \tag{1}$$

$$E_{Rx}(l,d) = l.E_{elec}^{Rx} \tag{2}$$

Where l is the length of the transmitted/received message in bits, d represents the distance over which the data is communicated and d_o is the distance threshold for swapping amplification models, which can be calculated as $d_o = \sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. As it can be seen, the transmitter expends energy to run the radio electronics and power amplifier, while the receiver only expends energy to run the radio electronics.

In this paper, we consider both free space (ε_{amp} =ε_{fs}) and two-ray multipath (ε_{amp} =ε_{mp}) models to approximate signal attenuation as a function of the distance between transmitters and receivers.

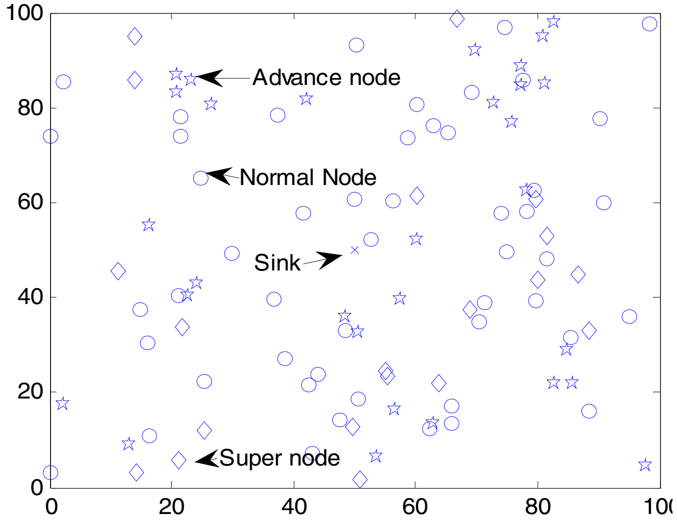


Fig. 1. Random Wireless Sensor Network

We describe our model of a wireless sensor network with heterogeneous nodes. Assume that there are n sensor nodes, which are uniformly dispersed within an M x M m^2square region (Figure. 1). The nodes always have the data to transmit to base station (sink). The heterogeneous settings and network model we have used is same as in [15]. Suppose E_o is the initial energy of each normal node. The energy of each super node is then $E_o(1+\beta)$ and each advanced node is then E_o(1+α).

4 SEP-E (RCH)

SEP-E (RCH) is a redundant and energy efficient cluster head selection method for enhanced stable election protocol. In this we first select initial cluster heads and initial redundant cluster head, and then compare the two nodes. The node which has better performance will be elected as cluster head, and other node will be as redundant cluster head in current round.

4.1 Selection of Initial Cluster Head

SEP-E (RCH) assigns a weight to the optimal probability p_{opt}.This weight must be equal to the initial energy of each node divided by the initial energy of the normal node. Let us define p_{nrm} the weighted election probability for normal nodes, p_{adv} the weighted election probability for advanced nodes, and p_{sup} the weighted election probability for super nodes [15].

$$p_{nrm} = \frac{p_{opt}}{1+m*(\alpha+m_o*\beta)} \tag{3}$$

$$p_{adv} = \frac{p_{opt}}{1+m*(\alpha+m_o*\beta)} * (1+\alpha) \tag{4}$$

$$p_{sup} = \frac{p_{opt}}{1+m*(\alpha+m_o*\beta)} * (1+\beta) \tag{5}$$

Where m is the proportion of advanced nodes to the total number of nodes n with energy more than the rest of the nodes and m_o is the proportion of super nodes. The threshold $T(nrm)$, T(sup), $T(adv)$ for normal, advanced and super nodes respectively remains the same as in [14].

For a node s , s generates a random number *rand*(s) from 0-1, *rand*(s) multiply a factor representing the remaining energy level of a node be as the new random number *rand* '(s):

$$rand'(s) = rand(s) * \left(E_{max} - \frac{E_{current}}{E_{max}}\right) \tag{6}$$

More remaining energy of a node, smaller the new random number produced, the probability is greater if *rand* '(s) is less than the threshold T(s), the node is more likely to become a cluster head. So if the cluster node in the cluster which generates random number *rand* '(s) less than threshold T(s), it will be selected as initial cluster head, and then broadcasts a message to the base station and other nodes that it was elected as initial cluster head. If *rand* '(s) is bigger than threshold T(s), node s will be the ordinary node. The ordinary nodes send join-request to the nearest cluster head.

4.2 Selection of Redundant Cluster Head

In section-V, part A, the improved cluster head selection algorithm only considered its remaining energy of cluster head, but not considered the location of cluster head and energy consumption in data transmission from other nodes to cluster head. As an example, some cluster head may be located at the edge of the cluster. The other nodes require lots of energy in sending data to them compared to the cluster heads which are located at the center of the cluster.

We select a node in the cluster that has more energy than other nodes except the initial cluster head in the cluster as the initial redundant cluster head. Initial cluster head and initial redundant cluster head both have chance to be cluster head. We can choose one which has optimal performance as the cluster head at this round, and the rest node will become the redundant cluster head. In the next section, we will describe the detail process of cluster head determination.

4.3 Determination of Cluster Head

After selecting initial cluster head (ICH) and redundant cluster head (RCH), we should determine which node is elected as cluster head. The specific steps are as follows:

- ICH and RCH both use flooding way to broadcast status information.
- When ordinary nodes receive the status information they broadcast their own status information and forward the other nodes status information which they received.
- ICH and RCH calculates mean distance to all other nodes in the cluster. Both of the nodes send their minimum mean distance K_{ICH} and K_{RCH} respectively, and residual energy to the base station.
- Base station receive the information and calculate the weight as follows:

$$C_{K_{ICH}} = \frac{K_{ICH}}{K_{ICH}+K_{RCH}}$$

$$C_{K_{RCH}} = \frac{K_{RCH}}{K_{ICH}+K_{RCH}}$$

$$C_{e_{ICH}} = \frac{E_{Curr-ICH}}{E_{Curr-ICH}+E_{Curr-RCH}}$$

$$C_{e_{RCH}} = \frac{E_{Curr-RCH}}{E_{Curr-ICH}+E_{Curr-RCH}}$$

- Compute C_I and C_R :

$$C_I = C_{K_{ICH}} \times C_{e_{ICH}}$$

$$C_R = C_{K_{RCH}} \times C_{e_{RCH}}$$

- Compare C_I with C_R,

```
If $C_I >= C_R$
        Node ICH is the cluster head in this round
Else
        Node IRCH is the cluster head.
End if.
```

This cluster head selection will be more reasonable and the energy consumption of the entire network will be less than enhanced stable election protocol.

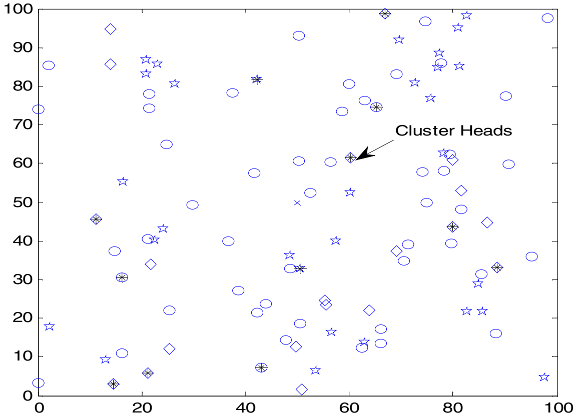


Fig. 2. Cluster head nodes in one round

4.4 Data Transmission Phase

In this phase, cluster members will transmit *l* bits to CH with probability *p* (0< *p* <1) by multiple frames. In each frame, each cluster member will transmit its data during its allocated transmission slot specified by the TDMA schedule in Cluster formation phase, and then sleep in other slots to save energy. After a CH receives data frames from its cluster members, it will perform data aggregation to remove the redundancy in the data. Then CH will transmit the aggregated data bits to the base station.

5 Performance Matrices

In this section the measures used to evaluate the performance of clustering protocols have been defined.

- *Stability Period:* is the time interval from the start of network operation until the death of the first sensor node. We also refer to this period as “stable region.”
- *Instability Period:* is the time interval from the death of the first node until the death of the last sensor node. We also refer to this period as “unstable region.”

- *Network Residual Energy:* It measures the instantaneous amount of energy being consumed in the network per round. This is simply the energy difference from the beginning till the end of the round
- *Number of alive per round:* This instantaneous measure reflects the total number of nodes and that of each type that has not yet expended all of their energy.
- *Data Packets received at base station:* It is total number of data packets or messages that are received by the base station. This measure varies linearly for all protocols.

Clearly, the larger the stable region and unstable region, the better the reliability of clustering process of sensor network is.

6 Simulation Results and Discussions

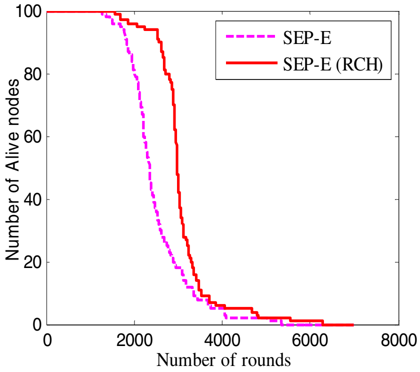
In order to validate the performance of the SEP-E (RCH) scheme, we compare it with SEP-E in the same heterogeneous setting, where the extra initial energy of advanced nodes and super nodes is uniformly distributed over the sensor field. We use $100m \times 100m$ region of 100 sensor nodes. MATLAB is used to implement the simulation. The key parameters [7, 13, 11] have been listed in Table 1. The simulation is performed to evaluate the performance of SEP-E (RCH).

Table 1. System parameters value

Parameters	**Value**
Network Size	100×100 meter^2
Sink	(50,50)
Number of Nodes	100
Initial Energy of Nodes	0.3 J
M	0.2
m_0	0.3
A	2
B	1
E_{elec}	50 nJ/bit
E_{fs}	10 pJ/bit/m^2
E_{mp}	0.0013 pJ/bit/m^4
E_{DA}	5 nJ/bit/message
Data Packet	4000 bits

Figure 3 indicate that the total number of alive nodes per round in SEP-E (RCH) is greater than SEP-E. This is due to fact that cluster heads moves to the center of the cluster at each round because we choose minimum mean distance as one of the cluster head selection criteria.

This leads to less consumption of energy for sending data to cluster head by the cluster members, hence increases the life time of the network. The stability period and instability period of SEP-E (RCH) was prolonged than SEP-E as demonstrated by Figure 3.



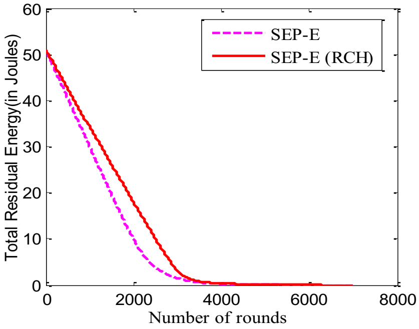


Fig. 3. Numbers of Nodes Alive

Fig. 4. Total residual energy of the per round

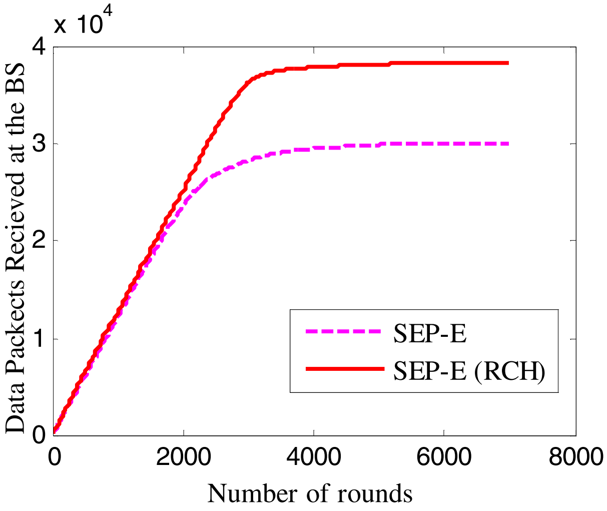


Fig. 5. Data packets received at the base station

Figure 4 depicts that the total residual energy of the network in SPE-E (RCH) at each round is greater than SEP-E because there are more number of alive nodes per round in SEP-E (RCH). The energy consumption per round is very less in SPE-E (RCH) as compared to SEP-E.

The initial energy of each node is 0.3 J.The energy consumption of SEP-E (RCH) is reduced by 14.5% as compared to SEP-E within 1000 rounds. Table 2 depicts the total residual energy of the network at different rounds as 1st, 1000th, 2000th and 4000th for both of the protocol. SEP-E (RCH) always has higher residual energy as compared to SEP-E.

Table 2. Comparison of total residual energy (in Joules)

Number of rounds	**Total Residual Energy (in Joules)**	
	SEP-E (RCH)	**SEP-E**
1	50.9820	50.9820
1000	34.3299	29.9742
2000	17.9360	9.9038
4000	0.2716	0.0610

Figure 5 demonstrate the number of data packets received at the base station. The results show that for both the protocols it goes linearly for around 200 rounds and after that the difference can be seen. It is clear SPE-E (RCH) has more numbers of data packets received at base station in comparison to SEP-E. Then we compare the both of the protocols based on the death of the nodes i.e. the time when first node dies (Stability Region), 25% nodes dies, 50% node dies and all node dies. Figure 6 shows the comparison using bar chart.

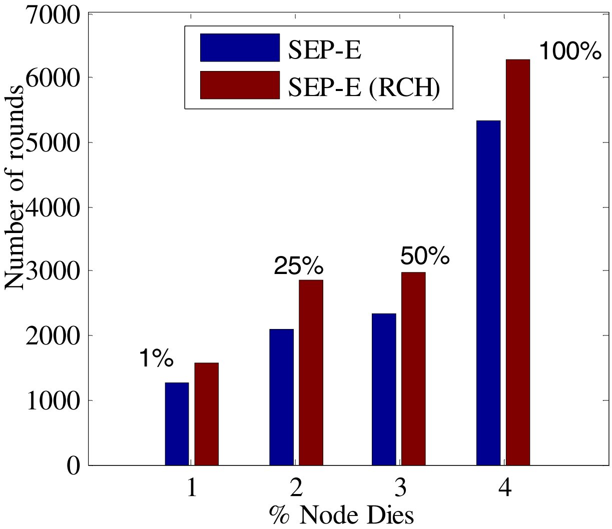


Fig. 6. Comparison between % node die

In SEP-E, the death of first node is encountered at 1273th round while in SPE-E (RCH) it is happened at 1571th round. Hence the stability region is significantly increased by 23.4%. Similarly, 25% of the node died at 2092th round and 2872th round and 50% of the node died at 2349th round and 2988th round in SEP-E and SPE-E (RCH) respectively (Table 3.2). At 5335th round and 6289th round the all of the nodes are goes down for SEP-E and SEP-E (RCH) respectively. Hence the instability period is also increased by 17.8%.

7 Conclusion

This paper designs a SEP-E (RCH) scheme for the redundant and energy efficient cluster head selection to balance the network load and to extend the network life time. In order to select reasonable cluster heads, our scheme first select an initial cluster head and a tentative or redundant cluster head in every cluster at each round. The node which has higher remaining energy and minimum mean distance is elected as cluster head of current round.

Simulation results shows that this method of electing cluster heads is more robust, energy efficient and increases the life time of the network than SEP-E protocol.

References

[1] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38(4), 393–422 (2002)

[2] Bokareva, T., Hu, W., Kanhere, S., Ristic, B., Gordon, N., Bessell, T., Rutten, M., Jha, S.: Wireless sensor networks for battlefield surveillance (2006)

[3] Dudek, D., Haas, C., Kuntz, A., Zitterbart, M., Krüger, D., Rothenpieler, P., Pfisterer, D., Fischer, S.: A wireless sensor network for border surveillance. In: Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems, pp. 303–304 (2009)

[4] Hart, J.K., Martinez, K.: Environmental Sensor Networks: A evolution in the earth system science. Earth Science Reviews 78, 177–191 (2006)

[5] Gupta, G., Younis, M.: Load balanced clustering in wireless sensor networks. In: Proceedings of the International Conference on Communication (ICC) (2003)

[6] Yong, S., Bo, J., Zong-lin, Z.: Survey on Energy Efficiency of Clustered Routing in wireless Sensor Network. Journal of Electronics & Information Technology, 2262–2264 (2007)

[7] Heinzelman, W.R., Chandraksan, A.P., Balakrishnan, H.: An application-specific protocol architectures for wireless microsensor networks. IEEE Trans. on Wireless Commnuications 1(4), 660–670 (2002)

[8] Xian, Z., Li, Y., Zhao, W.: MPTC – A minimum-energy path-preserving topology control algorithm for wireless sensor networks. 10th IFIP/IEEE International Conference on Management of Multimedia and Mobile Networks and Services Proceedings, 77–182 (2007)

[9] Shen, S., O'Hare, G.M.P., Marsh, D., Diamond, D., O'Kane, D.: AToM: Atomic topology management of wireless sensor networks. In: Gaiti, D., Pujolle, G., Al-Shaer, E.S., Calvert, K.L., Dobson, S., Leduc, G., Martikainen, O. (eds.) AN 2006. LNCS, vol. 4195, pp. 243–254. Springer, Heidelberg (2006)

[10] Camilo, T., Silva, J.S., Rodrigues, A., Boavida, F.: GENSEN: A topology generator for real wireless sensor networks deployment. In: Obermaisser, R., Nah, Y., Puschner, P., Rammig, F.J. (eds.) SEUS 2007. LNCS, vol. 4761, pp. 436–445. Springer, Heidelberg (2007)

[11] Smaragdakis, G., Matta, I., Bestavros, A.: SEP: A stable election protocol for clustered heterogeneous wireless sensor networks. In: Proc. of the Int'l Workshop on SANPA, pp. 251–261 (2004)

[12] Shepard, T.J.: A channel access scheme for large dense packet radio networks. In: Proccedings of ACM SIGCOMM, pp. 219–230 (1996)

[13] Heinzelman, W., Chandrakasan, A., Balakrishnan, H.: Energy-efficient routing protocols for wireless microsensor networks. In: Proc. 33rd Hawaii Int. Conf. System Sciences (HICSS), Maui, HI (2000)

[14] Aderohunmu, F.A., Deng, J.D.: An Enhanced Stabel Election Protocol (SEP) for Clustered Heterogeneous WSN. Discussion Paper Series, No. 2009/07. Department of Information Science, University of Ontago (2010) ISSN: 1177-455X

[15] Kumar, D., Aseri, T., Patel, R.B.: EEHC:Energy efficient heterogeneous clustered scheme for wireless sensor networks. Computer Communications 32, 662–667 (2009)

Mobility Based Energy Efficient Coverage Hole Maintenance for Wireless Sensor Network

Anil Kumar Sagar and D.K. Lobiyal

School of Computer and Systems Sciences
Jawaharlal Nehru University, New Delhi-110067
aksagar22@rediffmail.com, lobiyal@gmail.com

Abstract. Wireless Sensor Networks (WSNs) are a special class of wireless networks where randomly and densely distributed sensor nodes take local measurements of a phenomenon. Coverage is a fundamental measure of QoS which represent how well the Area of interest is monitored. Considering the limited battery power of sensors, this paper presents an Energy Efficient Coverage Hole Maintenance (EECHM) algorithm where redundant sensor nodes move towards coverage hole. In EECHM we have used probabilistic method to calculate the direction and magnitude of nodes that helps to migrate the redundant sensor nodes. Further, it also reduces the consumption of energy and thus enhances the network lifetime.

Keywords: Wireless Sensor Networks, Coverage hole, Energy, EECHM, Network lifetime.

1 Introduction

Wireless sensor network have attracted enormous research curiosity due to their wide range of applications such as temperature monitoring, environment monitoring, military surveillance, habitat monitoring, health care, etc. Each sensor node consists of three basic unit; sensing unit, processing unit and communication unit. The sensing unit senses the phenomenon of interest in the target area; processing unit processes the sensed data; and transmission unit sends the data to the base station. A WSN consists of a large number of sensor nodes, but they have limited memory, computational speed and battery power.

These limitations of sensor nodes affect the QoS [1].For example; the Art Gallery Problem [2] determines the minimum number of cameras necessary for art gallery area such that every point is observed by at least one camera. Sensor nodes can be deployed randomly or deterministically to collect the information. A well-planned deployment can help to maximize the sensing coverage area but it is impossible to deploy the sensors deterministically in hostile environments. Since sensors may be spread in an arbitrary manner, therefore, coverage becomes a fundamental research issue in WSN. Generally coverage reflects how well the deployed sensors monitor the area of interest. The coverage of WSNs could be classified into three types - Area

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 115–127, 2013.

Coverage, Point Coverage and Barrier Coverage. Area Coverage finds minimum number of sensors such that each physical point in the area is monitored by at least a working sensor. Point Coverage is used to cover a set of given targets. It only monitors a finite number of discrete points in target area. In Barrier Coverage sensors form a barrier for intrusion detection. There are mobile sensors such as MICAbot [3] and Robomote [4] which provide better coverage result as compared to static sensors. Random deployment or failure of sensor nodes may cause coverage holes that reduce the performance of the network. Therefore, it is important to identify these coverage holes and fill it with other redundant neighbor nodes.

The rest of the paper is organized as follows. Related work is given in section 2. Section 3 introduces problem description and proposed model. Section 4 describes the coverage hole maintenance of network by considering energy constraint of sensor nodes. Performance and Result analysis of our algorithms are done in Section 5. The work is concluded in section 6.

2 Related Work

Archana Sekhar et al. in [5] proposed a dynamic coverage maintenance scheme with limited mobility such that the neighboring node moves towards the dead node as much as possible and also the existing coverage region is not affected. In[6] authors uses the concept of voronoi diagram to find out the coverage holes and move mobile sensors from densely deployed areas to sparsely deployed areas. In [7], authors proposed algorithm to maintain the coverage and connectivity of the WSNs with limited mobility. Only one hop neighbors of the dead node are allowed to move towards coverage hole without disturbing existing communication and coverage. In [8] authors power off the redundant nodes and find the minimal set of sensors to maintain the desired level of coverage. Amitabha Ghosh in [9] uses Voronoi diagram to estimate the exact number of coverage holes and also proposes a collaborative algorithm to calculate the number of additional mobile sensor nodes to be deployed and relocated to the coverage hole area. Yu-Chen Kuo et al. In [10] have proposed a fast sensor relocation algorithm to arrange redundant nodes without GPS. Prasan Kumar Sahoo et al. In [11] use the concept of distributed algorithm and vector method to calculate the magnitude and direction of movement of mobile nodes for coverage hole recovery in wireless sensor network. To minimize the energy consumption sensor nodes with only one hop neighbors of coverage hole are allowed to move. Authors in [16] proposed a protocol (Co-Fi) for coverage hole maintenance with the help of mobile sensor nodes. In this protocol high energy nodes move towards the coverage hole without losing its existing coverage. Kazi Sakib et al. In [18] give three policies to repair coverage hole. In Directed Furthest Node First (DFNF) policy an active node selects it's one of deactivated neighbor based on distance from coverage hole. In Weighted Directed Furthest Node First (WDFNF) selects and replace deactivated node by considering both the distance and direction of a node from coverage hole. In Best Fit Node (BFN) all the active neighbors of a dead node participate and make decision to identify the appropriate replacement. Authors in [19] uses tracking and a robot repairing mechanism for maintaining coverage hole. Tracking mechanism calculates the movement trajectory information of the robot then robot repairing algorithm constructs the shortest path for repairing multiple coverage hole regions. This algorithm takes minimal time and energy to recover from coverage hole.

3 Definitions, Problem Statement and Network Model

3.1 Definitions

Definition 1. The sensing range R_S of a sensor is an area in which occurrence of an incident is sensed by the sensors.

Definition 2. The communication range R_C of a sensor is an area in which sensors can communicate with one another directly.

Definition 3. The sensing neighbor set of a node *S*, is a set of nodes located in the sensing range R_S of *S*.

Definition 4. The communication neighbor set of a node *S*, is a set of nodes located in the communication range R_C of *S*.

3.2 Problem Statement

The unmonitored parts of a coverage area are called coverage holes, which may exist due to the initial random deployment of the sensor nodes unable to cover the whole area. Coverage holes can also occur due to dead nodes. One of the prime reasons for a node to become dead is energy exhaustion of sensor nodes or due to environmental factors such as fire, storm etc. These coverage holes results in the ineffectiveness of the whole network, while most of the sensors are still working normally. If the coverage holes can be detected and removed with other redundant nodes, the whole network will work again effectively and efficiently.

3.3 Network Model

For large scale sensor network applications, deterministic deployments of sensor nodes is not feasible, therefore Sensors nodes are randomly and densely deployed into a 2-dimensional field as shown in fig.1. Each node has the same sensing and transmission range, and is assumed to be perfect disk. Each sensor node knows its location with the help of positioning system such as GPS or with some other localization technique [12]. Each node collects the location of its neighbors as soon as the deployment process is over. Each node is mobile can travel at a constant speed.

When a mobile sensor moves from one location to another, it consumes energy according to the distance traveled. Initially, each node has the same amount of energy. It is assumed that there are no obstacles on their way and there are multiple coverage holes after the deployment of sensor nodes.

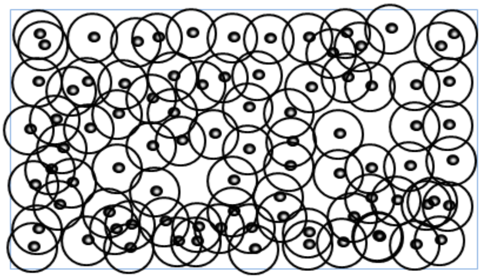


Fig. 1. Random Deployment of Sensor Nodes

4 Proposed Solution

In this section we have explained the algorithm devised for maintenance of coverage holes occurring due to uneven distribution of sensor nodes or due to failure of sensor nodes. The main objective of our work is to recover from these coverage holes by minimizing energy consumption. As sensor nodes are randomly and densely deployed over a square area, coverage area of many nodes may overlap with each other due to this uneven distribution of sensor nodes. Our main goal is to select and move high energy redundant nodes towards coverage hole area. In the proposed work, we calculate the residual energy E_{res} of each neighboring node around the coverage holes.

If the residual energy E_{res} of a node is more than threshold, it is allowed to move and fill the coverage hole. We calculate the magnitude and direction between a coverage hole and sensor node with the help of probabilistic method.

4.1 Random Node Distribution

Nodes are distributed in a square region. For our analysis, the sensor nodes are considered randomly distributed over two-dimensional geographical region. In many remote hostile environment sensors may be scattered with the help of aircraft. Since there are N sensor nodes uniformly deployed in region R_A, we have average node density $\lambda=N/R_A$. The number of sensors located in the region R_A, is $N\ (R_A)$ that is considered as parameter $\lambda \| R_A \|$of a Poisson process, where $\| R_A \|$ is the area of the region

$$P(N(R_A)=x)=\frac{e^{-\lambda\|R_A\|}(\lambda\|R_A\|)^x}{x!}.$$

4.2 Common Sensing Region

When there is random distribution of sensor nodes, it may happen that two or more sensor nodes can monitor the common sensing region. Let the Euclidean distance between two sensor nodes is d$_o$. From fig. 2 the overlap area [15] of two sensor nodes Φ is given by

$$\Phi = 2(\frac{1}{2}r^2\alpha - \frac{1}{2}r^2 \sin\alpha). \tag{1}$$

The value of angle α is calculated as

$$\cos\left(\frac{\alpha}{2}\right) = \frac{d_o/2}{r}.$$

$$\left(\frac{\alpha}{2}\right) = \cos^{-1}\left(\frac{d_o}{2r}\right).$$

$$\alpha = 2\cos^{-1}\left(\frac{d_o}{2r}\right).$$

Substituting the value of α in equation (1) we can find the overlap area of two sensor nodes as

$$\Phi = r^2\{\ 2\cos^{-1}\left(\frac{d_o}{2r}\right) - \sin(2\cos^{-1}\left(\frac{d_o}{2r}\right)\ \}.$$

This overlapping area can be reduced with increase in the Euclidean distance between the nodes.

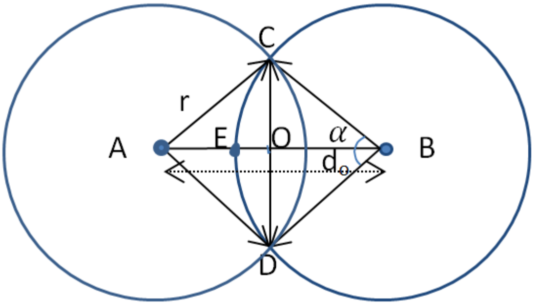


Fig. 2. Overlapped Coverage area of two Sensor Nodes

4.3 Probabilistic Method to Calculate Mobility of Redundant Nodes

In Fig.3 one hop redundant sensor nodes can move from its current position to dead node's position so that the coverage hole area is eliminated.

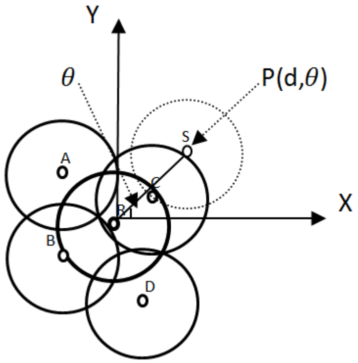


Fig. 3. Migration of Redundant Node

Let the coordinate of a dead node *S* is *(X_S, Y_S)*. Suppose one hop redundant neighbor of dead node *S* is *R* and its coordinate are *(X_R, Y_R)*. *A, B, C, & D* are neighbor nodes of redundant node *R*. Now, node *R* moves towards *S* with angle θ such that the coverage hole created by the dead node *S* can be filled. The distance d_m between node *R* and *S* is calculated as

$$d_m = \sqrt{(X_R - X_S)^2 + (Y_R - Y_S)^2} \quad . \tag{2}$$

This is the distance which a sensor is required to move for coverage hole maintenance. Now we can find angle θ by solving the following equation:

$$\tan\theta = \frac{Y_s}{X_s}. \tag{3}$$

$$\theta = \tan^{-1}\frac{Y_s}{X_s}. \qquad \text{If} \quad -\frac{\pi}{2} < \theta < \frac{\pi}{2} \tag{4}$$

$$\theta = \tan^{-1}\frac{Y_s}{X_s} + \pi. \qquad \text{If} \quad \frac{\pi}{2} < \theta \le \frac{3\pi}{2} \tag{5}$$

In our network model nodes are deployed with homogeneous Poisson process. The probability of coverage hole is given by 1-P (no other node in the area of πx^2) and x is the radius of a coverage hole.

$$F(x) = 1 - \frac{\| \pi x^2 \|^k \; e^{-\lambda \| \pi x^2 \|}}{k!}.$$

$$F(x) = 1 - e^{-\lambda \pi x^2}.$$

Now we can compute the probability density function by differentiating the F(x)

$$f(x) = \frac{d}{dx} F(x) = 2\lambda\pi \times e^{-\pi x^2}.$$

The probability of covering a coverage hole of radius x is given as

$$f(x) = \frac{d}{dx} F(x) = 2\pi \times e^{-\pi x^2}, \; for \; \lambda = 1.$$

4.4 Residual Energy Calculation

Residual Energy information of a sensor node is the total consumed energy by a sensor node. Generally, the main consumption of energy occurs during transmitting, receiving, sensing and amplifying and migration process. Let the total initial energy of a sensor node is denoted by E_{tot} .We use the first order radio model of [13] and according to this model, energy required to transmit *n* bit long message over distance *d* is

$$E_{TX}(n,d) = E_{diss} n + \Sigma_{amp} n d^x, where (x \geq 2). \tag{6}$$

Where, E_{diss} is the energy required by the electronic component for transmitting and receiving, and $\sum_{\text{amp}}$ is the transmitting amplifier energy. For receiving n bit long message energy required by a receiver is given by

$$E_{RX}(n) = E_{diss} n. \tag{7}$$

Similarly according to [14] energy consumption in sensing an event of interest, is given by

$$E_{sx} = V_s I_{sx} T_{sx}. \tag{8}$$

Where V_s is the voltage supply, I_{sx} is the current supplied and T_{sx} is the time desirable to sense the activity.

Energy Consumption during Sensor Node Movement: Energy consumption by a mobile sensor to move distance d_m is calculated as [17]

$$E(d_m) = \int_0^T p_{mov}(t) dt.$$

Where T is the time for migration of a sensor node. Let T be a function of ω and d_m Then

$$E(d_m) = \int_0^{f_T(\omega, d_m)} p_{mov}(t)\,dt. \tag{9}$$

Where,
$$P_{mov}(t) = V_{mov}(t) I_{mov}(t).$$

$$V_{mov}(t) = R I_{mov}(t) + K_e \omega(t).$$

$$I_{mov} = \frac{1}{K_t}[(J_m)\frac{d\omega(t)}{dt} + T_f + D\omega(t)].$$

Therefore, the residual energy of a sensor node is calculated as

$$E_{res} = (E_{tot} - (E_{TX}(n,d) + E_{RX}(n) + E_{SX} + E(d_m))). \tag{10}$$

A sensor node with residual energy higher than threshold value is selected for coverage hole maintenance.

Table 1. Notations

Symbol	Description
N	Total no of sensor nodes deployed
n	Number of bits in message
R_A	Deployment region
λ	Average node density
d_m	Distance between sensor node and coverage hole
E_{res}	Residual energy of a sensor node
E_{tot}	Total initial energy of a sensor node
E_{TX}	Energy required to transmit message
E_{diss}	Energy required by electronic component
Σ_{amp}	Transmitting amplifier energy
E_{RX}	Energy required for receiving a message
E_{RX}	Energy required for receiving a message
E_{sx}	Energy required for sensing an event
Φ	Overlap Area of two sensor nodes
$E(_{dm})$	Energy consumption during motion
$P(t)$	Power consumption by motor
$V(t)$	Voltage supply to motor
$I\ (t)$	Current flow to motor
R	Armature resistance
T	Time for movement
$\omega\ (t)$	Angular velocity
J_m	Inertia of motor
K_T	Torque constant of motor
D	Viscous damping
T_f	Friction torque

Algorithm. Energy Efficient Coverage Hole Maintenance

```
Step1. Select one hop redundant neighbors in the coverage hole and assign them to set
       $N_n$
Step2. Calculate the residual energy of each node in the set $N_n$
        Step2.1  Select a node $N_i$ from set $N_n$
        Step2.2  Initially $N_{TE} := \infty$;
        Step2.3  for all neighbor node $N_i \in N_n$ if the residual Energy
            $E_{res} > E_{th}$ (threshold)
                    $N_{TE} = N_i$;
                    $i = i + 1$;
            End if;
Return to step 2.1
Step3.Calculate the required distance and direction for movement by each node from
       set $N_{TE}$
Step3.1 Select a node from set $N_{TE}$.
Step3.2 Calculate distance of selected node from node S and assign it to set
        $N_{diss}$.
        Step3.3 Initially $N_{diss} :=\propto$
Step3.4 Repeat until $N_{TE} :=\propto$;
Step4. Find node with minimum distance and Migrate it with angle $\theta$ towards S
```

5 Result Analysis and Performance Evaluation

We have simulated and tested the proposed protocol. The simulation parameters are given in the following table 2. Initial deployment of the sensor node is random. The results obtained through simulation show the performance of the proposed model.

In the Fig. 4 shows that the overlapped sensing area of two nodes decreases with increase of Euclidean distance between two sensor nodes. Initially when the Euclidean distance between two sensor nodes is zero, overlapping area is maximum. It is evident from the results that for sensing range of *5m* and *10m*, if the Euclidean distance between two nodes becomes *10 m* and *20m*, respectively, the overlapping area is minimum.

Table 2. Simulation Parameter

Simulation parameter	Values
Simulation time	140s
R_A	500m×500m
E_{diss}	50 nj/bit
E_{tot}	3j
$_{amp}$	100pj/bit/m^2
V_s	3v
I_{sx}	25mA
T_{sx}	2ms
R	22•
J_m	3.824*10^{-6} z-in-sec^2
K_e	0.682 mv/rpm
K_T	0.923 oz-in/A
T_f	0.014 oz-in
D	4.85 * 10^{-8}Nm/(rad/s)

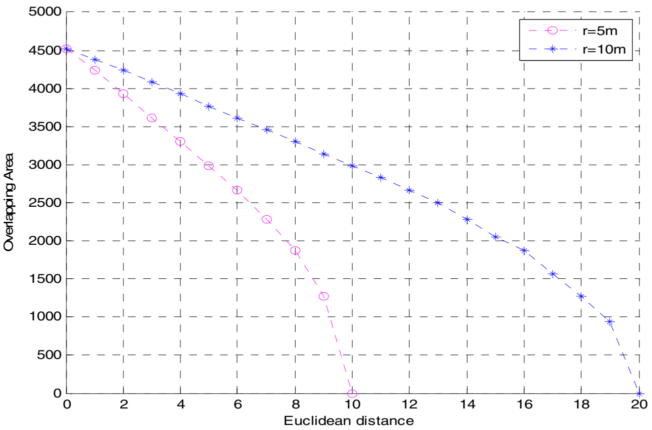


Fig. 4. Overlapping Coverage Area

Fig.5 shows that the energy consumption by sensor node is linear to the moving distance. When the moving distance of a sensor node is zero, energy consumption is greater than zero. This shows that there is always some energy consumed during startup.

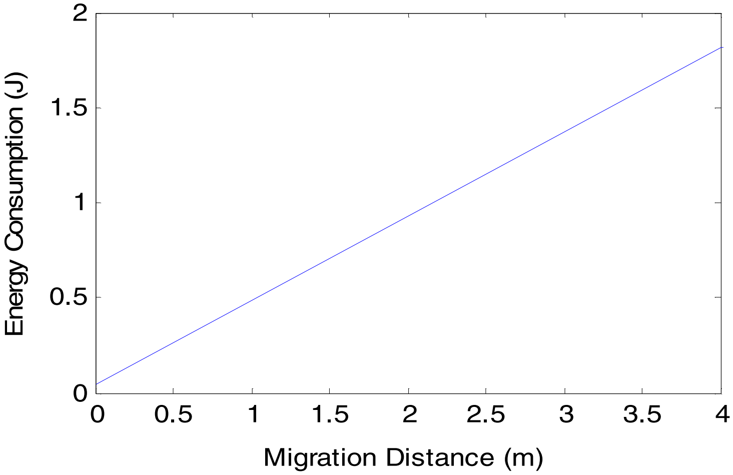


Fig. 5. Energy Consumption of Sensor Node during Mobility

Fig.6 shows relationship between migration of redundant sensor node and coverage probability. As the migration distance towards the coverage hole increases, the coverage probability also increases. It is evident from the figure that initially when the migration distance is 1m, the coverage probability is 0.72 but as the migration distance increases to 1.7m, the coverage probability becomes 1 since the coverage holes are covered by redundant sensor nodes.

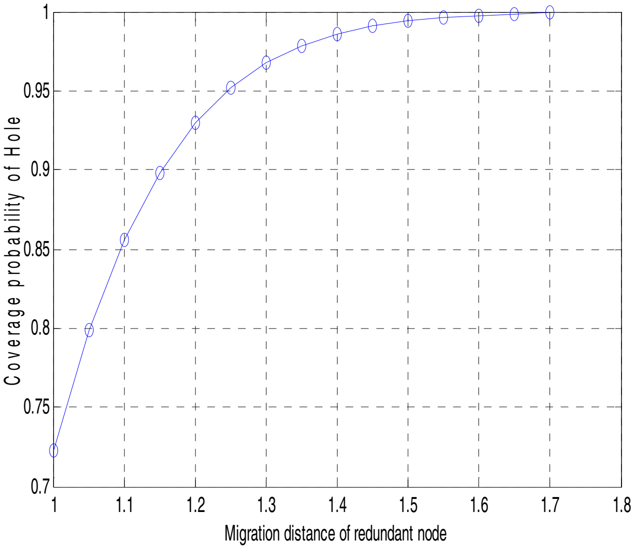


Fig. 6. Coverage Probability of Hole after Migration of Redundant Sensor Nodes

As shown in Fig.7, we have compared our algorithm with Co-Fi and DCM in terms of residual energy and time for different communication range. We know that communication consume more energy as compared to sensing activity of sensor nodes. The communication range of EECHM is 10m which is just half as compared to Co-Fi and DCM, therefore we get better result in terms of residual energy of sensor nodes.

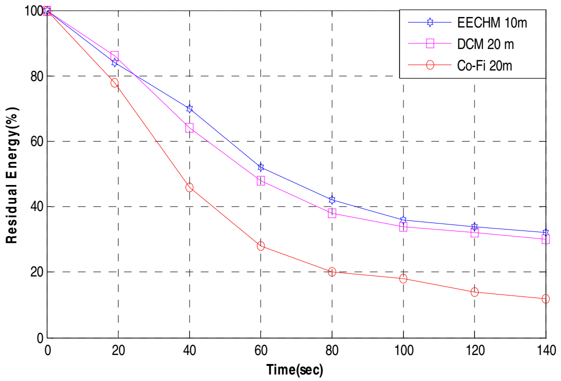


Fig. 7. Percentage remaining energy of sensor node for different time slots

6 Conclusion

In this paper we have proposed a solution for coverage hole maintenance problem. This solution is suitable for networks where sensor nodes distribution is non-uniform or failure of sensor nodes occur due to battery exhaustion, fire, storm etc. It is applicable for a network with restricted node mobility. Decision of distance and direction of mobility considers only one hop redundant neighbor nodes of coverage hole area. Due to limited mobility and least communication range, the energy consumption of sensor nodes is small and therefore, lifetime of the network is increased.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38(4), 393–422 (2002)
2. Marengoni, M., Draper, B.A., Hanson, A., Sitaraman, R.A.: A System to Place Observers on a Polyhedral Terrain in Polynomial. Time Image and Vision Computting 18(10), 773–780 (2000)
3. McMickell, M.B., Goodwine, B., Montestruque, L.A.: MICAbot: A Robotic Platform for Large-Scale Distributed Robotics. In: ICRA 2003 IEEE International Conference on Robotics and Automation, pp. 1600–1605. IEEE Press, Taiwan (2003)

4. Sibley, G.T., Rahimi, M.H., Sukhatme, G.S.: Robomote: A Tiny Mobile Robot Platform for Large-Scale ad-hoc Sensor Networks. In: ICRA 2002 IEEE International Conference on Robotics and Automation, pp. 1143–1148. IEEE Press, Los Angeles (2002)
5. Sekhar, A., Manoj, B.S., Murthy, C.S.R.: Dynamic Coverage Maintenance Algorithms for Sensor Networks with Limited Mobility. In: 3rd IEEE International Conference on Pervasive Computing and Communications, pp. 51–60. IEEE Press, Chennai (2005)
6. Wang, G., Cao, G., Porta, T.L.: Movement-assisted sensor deployment. IEEE Transactions on Mobile Computing 5(6), 640–652 (2006)
7. Sahoo, P.K., Sheu, J.P., Lin, W.S.: Dynamic Coverage and Connectivity Maintenance Algorithm for Wireless Sensor Networks. In: 2rd IEEE International Conference on Communication Systems Software and Middleware, pp. 1–9. IEEE Press, chungli (2007)
8. Slijepcevic, S., Potkonjak, M.: Power efficient organization of Wireless Sensor Networks. In: IEEE International Conference on Communications, ICC 2001, pp. 472–476. IEEE Press, Los Angeles (2001)
9. Ghosh, A.: Estimating Coverage Holes and Enhancing Coverage in Mixed Sensor networks. In: 29th IEEE International Conference on Local Computer Networks, pp. 68–76. IEEE Press, Bangolore (2004)
10. Kuo, Y.-C., Lin, S.-C.: A Fast Sensor Relocation Algorithm in Wireless Sensor Networks. World Academy of Science Engineering and Technology 56 (2009)
11. Sahoo, P.K., Tsai, J.-Z., Ke, H.-L.: Vector Method based Coverage Hole Recovery in Wireless Sensor Networks. In: 2nd IEEE International Conference on Communication System and Networks, pp. 1–9. IEEE Press, Taiwan (2010)
12. Savvides, A., Han, C.C., Srivastava, M.B.: Dynamic fine-grained localization in ad-hoc networks of sensors. In: 7th ACM International Conference on Mobile Computing and Networking, pp. 166–179. ACM Press, USA (2001)
13. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: 33rd Annual Hawaii International Conference on System Sciences, pp. 1–10. IEEE Press, USA (2000)
14. Halgamuge, M.N., Zukerman, M., Ramamohanarao, K.: An Estimation of Sensor Energy Consumption. Progress In Electromagnetics Research B 12, 259–295 (2009)
15. Sudip, M., Kumar, M.P., Obaidat, M.S.: Connectivity preserving localized coverage algorithm for area monitoring using wireless sensor networks. Computer Communications 34, 1484–1496 (2011)
16. Ganeriwal, S., Kansal, A., Srivastava, M.B.: Self Aware Actuation for Fault Repair in Sensor Networks. In: ICRA 4th IEEE, International Conference on Robotics & Automation. IEEE Press, Los Angeles (2004)
17. Electro-Craft Corporation.: DC Motors, Speed Controls, Servo Systems, An Engineering Handbook. Pergamon Press (1977)
18. Sakib, K., Tari, Z., Bertok, P.: Failed Node Replacement Policies for Maximising Sensor Network Lifetime. In: 6th International Symposium on Wireless and Pervasive Computing, pp. 1–6. IEEE Press, Australia (2011)
19. Chang, C.-Y., Lin, C.Y., Yu, G.-J., Kuo, C.H.: An energy-efficient hole-healing mechanism for wireless sensor networks with obstacles. Wireless Communications and Mobile Computing, 1530–8677 (2011)

HASL: High-Speed AUV-Based Silent Localization for Underwater Sensor Networks

Tamoghna Ojha and Sudip Misra

School of Information Technology, Indian Institute of Technology, Kharagpur,
West Bengal, India
{tojha,smisra}@sit.iitkgp.ernet.in

Abstract. The existing solutions that have been proposed to address the localization problem for mobile Underwater Sensor Networks (UWSNs) exhibit performance challenges such as high message overhead, localization error, and cost. Few Autonomous Underwater Vehicle (AUV) based methods were introduced to utilize the flexibility of movement of an AUV. In this paper, we propose a distributed, 3-dimensional, energy-efficient localization scheme, named High-Speed AUV-Based Silent Localization (HASL), for large-scale mobile UWSNs. Three AUVs are used to provide beacon messages to localize the mobile sensor nodes 'silently'. Therefore, with the use of high-speed AUV and 'silent' listening, we design an efficient scheme capable of addressing some of the above mentioned challenges with the existing solutions. We evaluated our proposed scheme in NS-3 simulator. Simulation results show that HASL achieves more than 90% localization coverage with localization error in the order of 2-7 *meters*.

Keywords: AUV based localization, Underwater Acoustic Sensor Networks, Silent Localization

1 Introduction

In event-driven networks such as UWSNs, tagging of sensed data with location information is fundamentally important, specifically for applications such as target tracking [1], and environmental monitoring. Also, node localization is applied in the design of underwater simulators [2]. Moreover, the performance of medium access and routing protocols can be significantly increased by providing location-aware information. However, UWSNs exhibit some unique challenges, e.g., passive node mobility, acoustic communication, high energy consumption and limited bandwidth [3], [4], [5]. Due to passive node mobility, the sensor nodes do not remain in the same position over time, and, thus, it is difficult to deploy fixed anchor nodes underwater. Therefore, the localization schemes which consider sensor nodes and anchor nodes to be static inside water, do not work efficiently in mobile UWSN. Acoustic communication consumes more energy than radio frequency modems, and transmission of signal consumes at least 10 times more power than reception. For example, the acoustic modem proposed by [6]

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 128–140, 2013.

consumes 0.203 *watts*, 0.024 *watts* and 3×10^{-6} *watts* for transmit, receive, and sleep, respectively. Hence, it is not energy-efficient to use frequent message exchange schemes in underwater localization. Also, the limited network bandwidth can be saved if the communication overhead is reduced.

Few localization protocols are proposed for 3D, mobile UWSNs in [7], [8], [9], [10]. However, these methods either introduce communication overhead, and implementation cost in the network, or require direct communication between the surface sinks and the underwater anchor nodes. The AUV-based methods [11], [12], [13], [14] use AUVs instead of anchor nodes as the beacon provider. Some of these methods introduce communication overhead, high localization delay (for example, [11], [14]), and assume that the sensor nodes are stationary or cabled with anchors and buoys.

The above mentioned challenges motivate us to propose an energy-efficient localization scheme, specifically for large-scale mobile UWSNs. To utilize the flexibility of movement of an AUV inside water, we use three high-speed AUVs as the location beacon provider. At the same time, the sensor nodes employ 'silent' listening of beacon messages, which reduces the communication overhead, and energy consumption significantly. Initially, the AUVs remain at the water surface, and receive GPS coordinates while floating. Subsequently, they descend below the water surface, and follow a predefined trajectory to cover the whole deployment area. Each underwater node silently localizes itself after receiving beacon messages from three different AUV.

The rest of the paper is organized as follows. In Section 2, we briefly present the related works, and in Section 3, we describe the proposed localization scheme, HASL, in detail. We evaluate HASL and present the simulation results in Section 4. Finally, we conclude the paper with suggestions for few future research directions in Section 5.

2 Related Works

A large number of research works exist in the literature for UWSN localization. Few excellent survey papers [15], [16] also exist summarizing the existing work done on this problem. The early works on UWSN localization [17], [18], [19] concentrated mainly on small scale network localization, and are impaired by huge communication cost, and low convergence.

Erol *et al.* proposed a localization method, named Dive'n'Rise Localization (DNRL) [7], with "Dive'n'Rise" (DNR) mobile anchor nodes, which dive and rise along the water column. The anchor nodes get their coordinates using the GPS receiver affixed with them, and then they dive into water to announce their coordinate through beacon messages. The ordinary nodes listen to the beacon messages and localize themselves 'silently'. The advantages of DNRL are that it is 'silent' and energy-efficient. The disadvantage of this method is, the requirement of large number of mobile anchors for localization coverage. Also, the implementation cost increases with the use of large number of mobile anchors. The nodes placed deep inside water are localized later than the nodes placed near the surface. Moreover, nodes should be time synchronized to estimate range from Time-of-Arrival (ToA) measurements. However, as these DNR beacons are slow, the position estimation of the sensor nodes are greatly affected by node mobility. In [11], AUVs were used instead of the DNR beacons. The method in this work does not require any fixed network infrastructure, and the nodes do not require to be time-synchronized. However, this method is not energy-efficient as the nodes do 'active' message exchange between themselves, and the localization delay in this method is about 2 *hours*.

DNRL was extended in Multi Stage Localization (MSL) [9], where the nodes once localized are used as reference nodes for the rest of the nodes. The use of iterative localization increases the overall communication cost, and, thus, additional energy gets consumed. Also, the position estimation error in the first iteration propagates to rest of the nodes, while they are localized using those estimated coordinates. This method also needs time synchronization between nodes. Another multi-stage iterative localization method was proposed in [14], where unlocalized nodes are initially localized with the help of an AUV. Thereafter, the rest of the nodes localize themselves iteratively. However, this scheme is limited only to 2D static UWSNs.

Mirza *et al.* proposed a localization scheme [8], where the effect of propagation delay, and node mobility in distance estimation were considered. The centralized algorithm iteratively estimates the distance throughout the execution of the application. Therefore, this method does not require anchor nodes. However, the method was criticized for the lack of time synchronization algorithm [15]. Another centralized prediction based localization scheme, called Collaborative Localization, was proposed in [20]. In this method, a specific application scenario was considered, where two types of nodes, 'profilers' and 'followers', descend under the water. The 'profilers' descend faster, and they predict the future locations of the 'followers' by measuring the distance from 'followers' using ToA technique. However, lack of synchronization between the nodes may disrupt the performance of the algorithm. Moreover, the algorithm specifically suits limited applications.

Using Directional Beacons for Localization (UDB) [12], the authors used a directional antenna powered AUV as the location beacon provider. It was extended for 3D UWSNs in Localization with Directional Beacons (LDB) [13]. Both of these methods are energy-efficient, as they use 'silent' localization. However, the authors assumed that the nodes cannot move freely, as they are restricted to move by the elastic anchor chain's pull force, and buoy's floating force. Moreover, to cover the whole network, an AUV needs to traverse the network more than once, as it is able to send beacon to one direction only.

3 High-Speed AUV-Based Silent Localization

3.1 Assumptions

We consider a 3-dimensional UWSN deployed in a large area, which is affected by passive node mobility. The sensor nodes are equipped with pressure sensors, using which they are able to calculate their depth. We also consider that three AUVs move along the middle of the network maintaining their trajectory using dead-reckoning [21], [22]. The scenario model is shown in Figure 1. We also assume that the AUVs are time synchronized and move together maintaining the same speed. Therefore, the AUVs broadcast beacons together and the sensor nodes receive three beacons from three different AUVs at nearly the same time. Also, the displacement of a sensor node between the time interval of the reception of three beacons is negligible. The deployment area may be along an ocean coast, which is bounded by large value of area length with comparatively less breadth and depth.

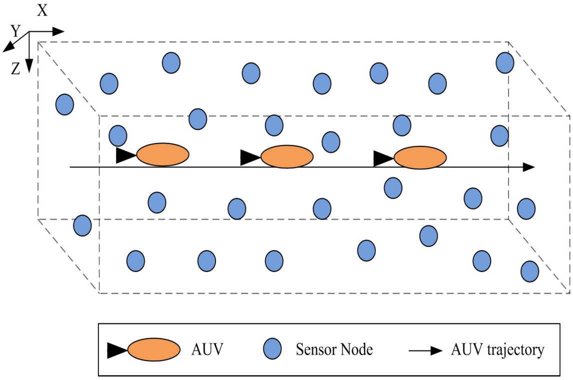


Fig. 1. Deployment of sensor nodes and the AUVs

3.2 Features

In this Section, we summarize the salient features of HASL, which differentiates it from the existing localization protocols.

a. *It is applicable to large-scale mobile UWSNs.* The HASL protocol can localize nodes deployed over a vast area. We consider that the nodes are displaced with the effect of passive node mobility. The effect of passive node mobility does not affect the performance of the protocol.
b. *It is energy-efficient.* We make use of the ‘silent’ beacon message receiving method, and thus, the nodes consume energy only for beacon listening.
c. *Time synchronization between sensor nodes is not needed.* In HASL, the sensor nodes passively listen to the AUVs. Thus, there is no additional requirement of time synchronization between the sensor nodes. However, the AUVs are time synchronized.
d. *No anchor nodes required.* There is no need of deploying any static anchor node inside or above the water surface. Therefore, HASL is free from the complexity of deploying static anchor node inside water.

e. *Protocol overhead is very low.* Only three beacon messages are required for localizing an unlocalized sensor node.

f. *Localization time is low.* The time required to localize the sensor nodes $t \propto \frac{l}{v}$, where l is the length of the network, and v is the average velocity of the AUVs.

3.3 Procedure

Beacon Sending. The AUVs are the only location beacon providers in the proposed HASL scheme. Initially, the AUVs collect their coordinates from the GPS receiver attached with them. Then they descend till the middle of the deployment region's depth. This is the starting position of the AUV's trajectory. It maintains its predefined trajectory along the middle of the network. Each AUV broadcasts beacon messages starting from the first position, with constant time interval between two beacon messages. Each beacon message sent from the AUV contains the AUV ID, and the present location of that AUV with current time-stamp. One set of 'effective beacon messages' is formed with beacon messages from three different AUVs, with the same time-stamp value. Using the time-stamp value, the sensor nodes differentiate between two different sets of beacon messages. The beacon message format is shown in Figure 2.

AUV id	X_{AUV}	Y_{AUV}	Z_{AUV}	time

Fig. 2. A beacon message format

Sensor Node Localization. Each sensor node 'silently' listens to beacon messages, and measures its distance (d_i) from the corresponding AUV using Received Signal Strength Intensity (RSSI). The sensor nodes are able to calculate their depth using the pressure sensor equipped with them. Therefore, a sensor node needs to calculate its x, and y coordinate values only. Let, at t_1 time, the position of the three AUVs at those time be (x_1, y_1, z_1), (x_2, y_2, z_2), and (x_3, y_3, z_3), respectively. This is one set of 'effective beacon messages'.

For each of the three beacon messages,

$$(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2 = d_i^2, \text{where, } i = 1, 2, 3 \tag{1}$$

$$\text{or, } (x - x_i)^2 + (y - y_i)^2 = d_i^2 - h^2, \text{where, } h = z - z_i$$

Here, $i = 1, 2, 3$ is the three beacon messages from the three AUVs. From these three beacon messages, a sensor node can successfully localize itself using Equation 1.

4 Performance Evaluation

4.1 Simulation Settings

We evaluated the performance of HASL through simulation based experiments performed in NS-3 simulator [23]. The simulation area considered was 1000 m × 200 m × 200 m, and in each of these simulation regions, we deployed 100, 200, 300, 400 and 500 nodes. In two different scenarios, the transmission range of the AUVs, as well as the sensors, are taken to be 100 m, and 150 m. Each AUV moves with constant velocity of 15 $Knots$ (7.7166 m/s), maintaining its trajectory, as shown in Figure 1. This type of high-speed AUV, called *VT High Speed AUV (VT HSAUV)* designed by Virginia Tech Autonomous Systems and Controls Laboratory, the specifications can be found at [25]. Other simulation parameters are shown in Table 1.

Table 1. Simulation Parameters

Parameter	Value
Inter AUV distance	10-30 m
Node mobility	0.5-2 m/s
Node mobility model	Meandering Current Mobility model [24]
Transmission power	0.203 $watts$ [6]
Receive & Idle power	0.024 $watts$ [6]
Sleep power	3×10^{-6} $watts$ [6]
Initial energy of a node	150 J

4.2 Performance Metrics

We evaluated the performance of our algorithm using the following metrics:

- *Localization Error:* Localization error is the Euclidean distance between the sensor node's estimated location, and the original location.
- *Localization Coverage:* It is defined as the number of localized nodes to the total number of nodes. A node is considered to be localized if the localization error is less than the error threshold.

4.3 Benchmark

We compare the performance of the proposed protocol, HASL, with LDB [13]. LDB is an localization algorithm for 3D static UWSNs. In LDB, an AUV equipped with directional antenna is the location beacon provider for the sensor nodes. The sensor nodes localize themselves using *the first-heard beacon point*, and *the last-heard beacon point* of beacon message from the AUV.

LDB uses 'silent' localization, which provides an energy-efficient localization scheme. Also, it is free from range estimation. Therefore, we choose LDB to compare the performance of HASL.

4.4 Results and Analysis

Effect of Node Mobility. In Figure 3, we plot the localization error for the localized sensor nodes. The transmission range of both AUVs, and the nodes are set to 100 m. The error threshold value is set to be 10 m. The inter-AUV-distance for this experiment was set to 20 m. We plot the results for node mobility value of 0.5 m/s, 1.0 m/s, and 2.0 m/s. The localization error increases with higher node mobility. However, LDB results in more localization error than HASL. In LDB, nodes estimate their position by the first-heard, and the last-heard beacon positions. However, LDB does not include the displacement of sensor nodes, during this time period, in the estimation.

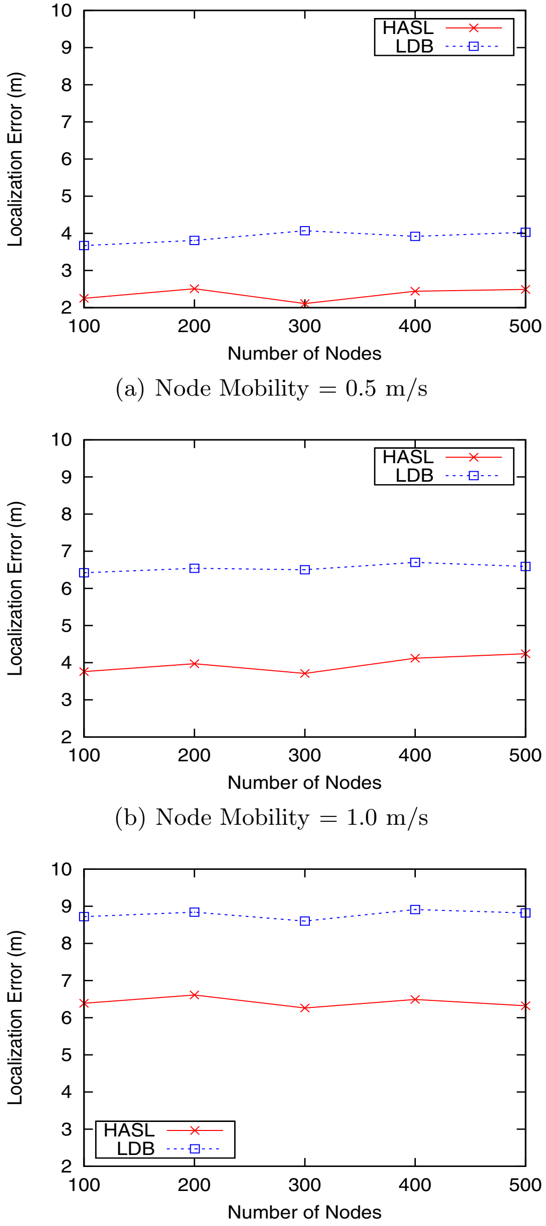
In Figure 4, we plot the localization error for HASL, and LDB. Here, the transmission range of both AUVs, and the nodes are set to 150 m. With the increase of transmission range, the farthest nodes are also localized. However, the transmission delay of beacon messages let the nodes displace more from their original position. With this amount of added delay, LDB results in more localization error.

Effect of Transmission Range. We simulated HASL by changing the transmission range of the sensor nodes and the AUVs. Increasing the transmission range allows more number of nodes to be present in the communication zone of the AUVs. Therefore, more number of sensor nodes can receive beacon messages, which results in better localization coverage. We show the results of this experiment in Figure 5.

LDB also exhibited the same type of performance characteristics for localization coverage in both the scenarios. However, the time required to localize the sensor nodes is double in case of LDB compared to HASL. This is because of the directional antenna used in LDB, the AUV need to traverse the deployment area twice.

In Figure 6, we present the effect of transmission range on the localization coverage. The more the number of nodes present in the transmission range of the AUVs, the more is the localization coverage. The solid circle represents the zone covered by transmission range r of the AUV, and the dotted circle represents the zone covered when the transmission range is R. The maximum coverage is attained for $R = d \times \sqrt{2}$, when d is the side of the square.

Effect of Inter-AUV-Distance. We study the effect of inter-AUV-distance (d_{aa}) on the localization error, and the results are plotted in Figure 7. In this experiment, we varied d_{aa} from 10-30 m with the transmission range of the AUVs, and the sensor nodes were set to 100 m. It is found that, when the AUVs are closely spaced, the localization estimation is coarse. However, increasing

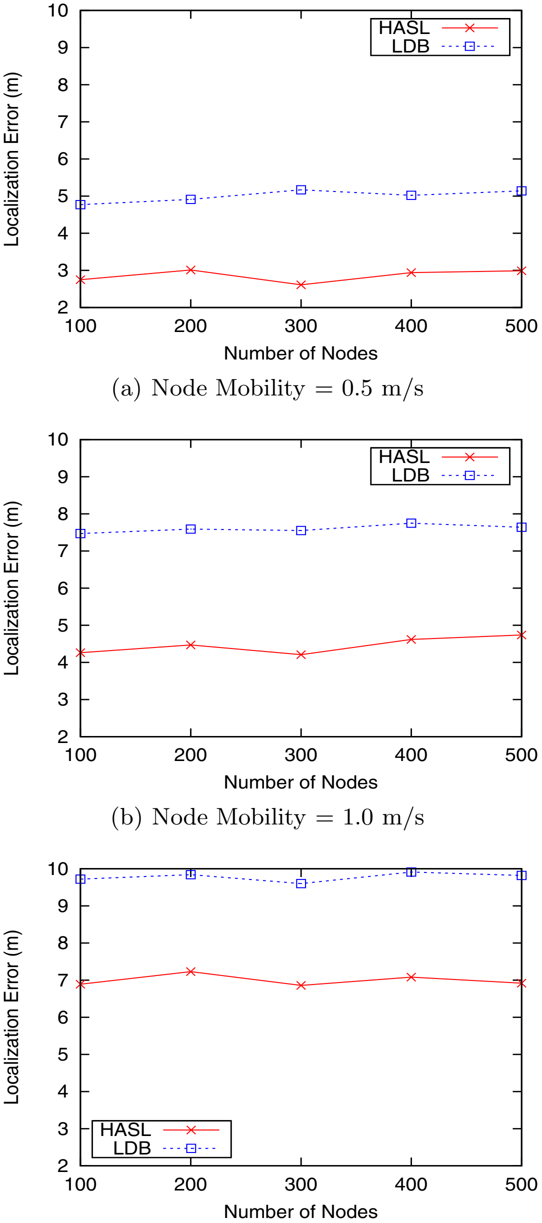


(a) Node Mobility = 0.5 m/s

(b) Node Mobility = 1.0 m/s

(c) Node Mobility = 2.0 m/s

Fig. 3. The effect of node mobility on localization error (transmission range = 100 m)



(a) Node Mobility = 0.5 m/s

(b) Node Mobility = 1.0 m/s

(c) Node Mobility = 2.0 m/s

Fig. 4. The effect of node mobility on localization error (transmission range = 150 m)

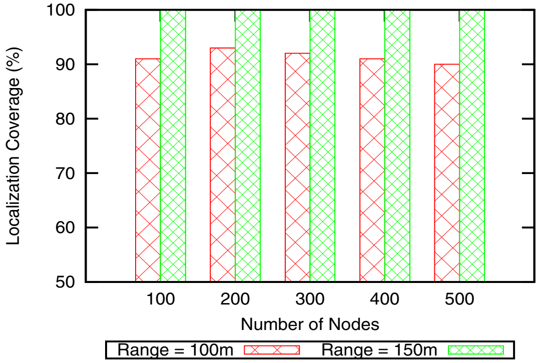


Fig. 5. Localization Coverage

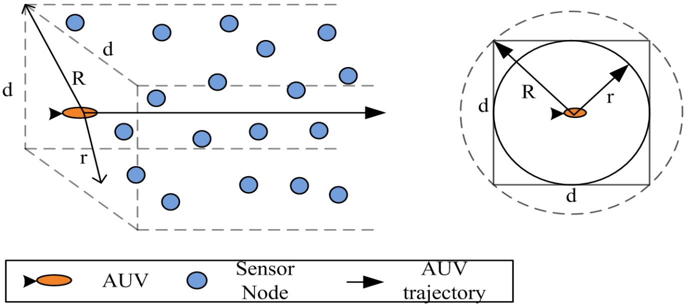


Fig. 6. Effect of Transmission Range on Localization Coverage

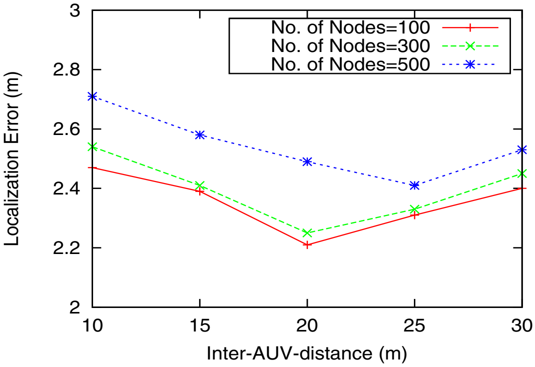


Fig. 7. Effect of inter-AUV-distance on Localization Error (Node mobility = 0.5 m/s)

the d_{aa} parameter further exceeding $d_{aa} = 20\ m$, the localization error again increases. This behavior can be explained with the help of propagation delay. For higher inter-AUV-distance, the three beacons take different time to reach a sensor node. During this time, the sensor nodes change their position, thereby introducing increased error in localization.

5 Conclusion

In this paper, we proposed HASL, an energy-efficient localization scheme for large-scale mobile UWSNs. Three high-speed AUVs are used as location beacon provider. As the mobile sensor nodes employ ‘silent’ beacon listening, nodes do not have any extra communication overhead. Therefore, HASL is, indeed, energy-efficient. Sensor nodes localize themselves in less time, which is equal to one trip travel time of the AUVs. The time required to localize the sensor nodes increases with increase of network dimension, and decreases with increase of AUV speed.

The simulation results show that with increasing effect of passive node mobility, HASL localizes nodes with less localization error than LDB. We showed how the communication range of an AUV affect the localization coverage.

In the future works, we would like to study: 1) different AUV trajectories to increase the localization coverage in different types of node deployment, and 2) the performance of the proposed protocol under the various underwater issues, such as, jamming [26], [27], wormhole attack [28], variable sound speed [29], and shadow zone.

Acknowledgement. This work is partially supported by a grant from the Department of Electronics and Information Technology, Government of India, Grant No. 13(10)/2009-CC-BT, which the authors gratefully acknowledge.

References

1. Misra, S., Singh, S.: Localized policy-based target tracking using wireless sensor networks. ACM Transactions on Sensor Networks 8, 27 (2012)
2. Dhurandher, S.K., Misra, S., Obaidat, M.S., Khairwal, S.: UWSim: A Simulator for Underwater Sensor Networks. Simulation 84, 327–338 (2008)
3. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater Acoustic Sensor Networks: Research Challenges. Ad Hoc Networks 3, 257–279 (2005)
4. Heidemann, J., Li, Y., Syed, A., Wills, J., Ye, W.: Research Challenges and Applications for Underwater Sensor Networking. In: Proceedings of IEEE Wireless Communication and Networking Conference, pp. 228–235 (2006)
5. Cui, J.-H., Kong, J., Gerla, M., Zhou, S.: Challenges: Building Scalable Mobile Underwater Wireless Sensor Networks for Aquatic Applications. IEEE Network 20, 12–18 (2006)
6. Sanchez, A., Blanc, S., Yuste, P., Serrano, J.J.: A low cost and high efficient acoustic modem for Underwater Sensor Networks. In: Proceedings of IEEE OCEANS, pp. 1–10 (2011)

7. Erol, M., Vieira, L.F.M., Gerla, M.: Localization with Dive NRise (DNR) Beacons for Underwater Acoustic Sensor Networks. In: Proceedings of Workshop on Underwater Networks (WUWNet), pp. 97–100 (2007)
8. Mirza, D., Schurgers, C.: Motion-Aware Self-Localization for Underwater Networks. In: Proceedings of Workshop on Underwater Networks (WUWNet), pp. 51–58 (2008)
9. Erol, M., Vieira, L.F.M., Caruso, A., Paparella, F., Gerla, M., Oktug, S.: Multi Stage Underwater Sensor Localization using Mobile Beacons. In: Proceedings of Sensor Technologies and Applications (SENSORCOMM), pp. 710–714 (2008)
10. Zhou, Z., Peng, Z., Cui, J.-H., Shi, Z., Bagtzoglou, A.C.: Scalable Localization with Mobility Prediction for Underwater Sensor Networks. IEEE Transaction on Mobile Computing 10, 335–348 (2011)
11. Erol, M., Vieira, L.F.M., Gerla, M.: AUV-Aided Localization for Underwater Sensor Networks. In: Proceedings of Wireless Algorithms, Systems and Applications, pp. 44–54 (2007)
12. Luo, H., Zhao, Y., Guo, Z., Liu, S., Chen, P., Ni, L.M.: UDB: Using directional beacons for localization in Underwater Sensor Networks. In: Proceedings of Parallel and Distributed Systems (ICPADS), pp. 551–558 (2008)
13. Luo, H., Guo, Z., Dong, W.: LDB: Localization with Directional Beacons for Sparse 3D Underwater Acoustic Sensor Networks. J. of Networks 5, 28–38 (2010)
14. Waldmeyer, M., Tan, H.-P., Seah, W.K.G.: Multi-stage AUV-aided Localization for Underwater Wireless Sensor Networks. In: Proceedings of Advanced Information Networking and Applications, pp. 908–913 (2011)
15. Erol-Kantarci, M., Mouftah, H.T., Oktug, S.: A Survey of Architectures and Localization Techniques for Underwater Acoustic Sensor Networks. IEEE Communications Surveys and Tutorials 13, 487–502 (2011)
16. Tan, H.-P., Diamant, R., Seah, W.K.G., Waldmeyer, M.: A survey of techniques and challenges in underwater localization. Ocean Engineering 38, 1663–1676 (2011)
17. Austin, T., Stokey, R., Sharp, K.: PARADIGM: A Buoy-Based System for AUV Navigation and Tracking. In: Proceedings of MTS/IEEE Oceans, pp. 935–938 (2000)
18. Bechaz, C., Thomas, H.: GIB System: The Underwater GPS Solution. In: Proceedings of Europe Conference on Underwater Acoustics (ECUA) (2000)
19. Garcia, J.E.: Ad Hoc Positioning for Sensors in Underwater Acoustic Networks. In: Proceedings of MTS/IEEE Oceans, pp. 2338–2340 (2004)
20. Mirza, D., Schurgers, C.: Collaborative localization for fleets of underwater drifters. In: Proceedings of MTS/IEEE OCEANS, pp. 1–6 (2007)
21. Fallon, M.F., Papadopoulos, G., Leonard, J.J., Patrikalakis, N.M.: Cooperative AUV Navigation using a Single Maneuvering Surface Craft. International Journal of Robotics Research 29, 1461–1474 (2010)
22. Woithe, H., Boehm, D., Kremer, U.: Improving Slocum Glider Dead Reckoning Using a Doppler Velocity Log. In: Proceedings of MTS/IEEE OCEANS 2011, pp. 1–5 (2011)
23. NS-3 Simulator, `http://www.nsnam.org/` (accessed December 15, 2012)
24. Caruso, A., Paparella, F., Vieira, L.F.M., Erol, M., Gerla, M.: The Meandering Current Mobility Model and its Impact on Underwater Mobile Sensor Networks. In: Proceedings of IEEE INFOCOM, pp. 221–225 (2008)
25. Autonomous Undersea Vehicle Applications Centre, `http://auvac.org/configurations/view/106` (accessed December 15, 2012)

26. Misra, S., Dash, S., Khatua, M., Vasilakos, A.V., Obaidat, M.S.: Jamming in Underwater Sensor Networks: Detection and Mitigation. IET Communications 6, 2178–2188 (2012)
27. Khatua, M., Misra, S.: Exploiting Partial-Packet Information for Reactive Jamming Detection: Studies in UWSN Environment. In: Frey, D., Raynal, M., Sarkar, S., Shyamasundar, R.K., Sinha, P. (eds.) ICDCN 2013. LNCS, vol. 7730, pp. 118–132. Springer, Heidelberg (2013)
28. Wang, W., Kong, J., Bhargava, B., Gerla, M.: Visualisation of Wormholes in Underwater Sensor Networks: a Distributed Approach. International Journal of Security and Networks 3, 10–23 (2008)
29. Misra, S., Ghosh, A.: The effects of variable sound speed on localization in Underwater Sensor Networks. In: Proceedings of Australasian Telecommunication Networks and Applications Conference (ATNAC), pp. 1–4 (2011)

Clusterhead Selection Using Multiple Attribute Decision Making (MADM) Approach in Wireless Sensor Networks

Puneet Azad[1,2] and Vidushi Sharma[2]

[1] Department of Electronics & Communication, Maharaja Surajmal Institute of Technology,
C-4, Janak Puri, New Delhi 110058, India
[2] School of Information & Communication Technology, Gautam Buddha University,
Yamuna Expressway, Greater Noida, Gautam Budh Nagar, Uttar Pradesh 201308, India
puneet_azad@yahoo.com, vidushi@gbu.ac.in

Abstract. Cluster head (CH) plays an important role in aggregating and forwarding data in a wireless sensor networks (WSNs). The major challenge in WSNs is an appropriate selection of cluster heads for gathering data from nodes. In this paper, we present a multi-criterion approach for the selection of cluster heads (CHs) using Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Three attributes are considered for the selection of CHs, namely residual energy, number of neighbors and distance from the base station. The simulation results demonstrate that the present approach is more effective than another Low-energy Adaptive Cluster Hierarchy (LEACH) protocol in prolonging the network lifetime.

Keywords: Wireless Sensor Networks, Clustering, TOPSIS, Optimum.

1 Introduction

The WSNs have recently attracted widespread attention as they can be deployed without the need of existing communication infrastructure. They have a wide range of applications in forest fire detection [1], surveillance [2], audio and video retrieval [3], healthcare [4] etc. These networks consist of nodes having sensing, data processing, and communicating components [5] for data collection from the remote or inaccessible areas. One of the stringent requirements of these nodes is the efficient use of available energy as it is difficult to recharge or replace their batteries once they are deployed. A variety of algorithms have been designed to cater the need of conservation of energy in WSNs. Clustering is a statistical technique used for grouping the sensor nodes into clusters based on several attributes like location, residual energy, distance from the base station, signal strength or connectivity etc. The nodes present in each cluster are responsible for sensing the physical phenomenon under consideration. Each cluster has a coordinator called cluster head, which is responsible for gathering the data from the nodes present in the cluster. Once the CH drains its entire energy, there is a need to replace the cluster head. Thus, in each cycle of data transmission, re-clustering can be done to rotate the position of

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 141–154, 2013.

cluster head to enhance the overall network lifetime of the system. The selection of cluster heads may either be chosen random or it can be based on one or more criteria. A systematic approach to cluster head selection process is necessary in order to select the optimum clusters for the WSN application.

In the present study, the sensor nodes are first screened using Pareto-optimal solution [6] and further TOPSIS is used to select the best cluster heads based on three different criteria. After the cluster heads are selected, clusters are formed using minimum Euclidean distance of nodes from the base station. Multiple criteria approaches have been used in a variety of applications using including MOECS [7] and TOPSIS [8].

The rest of the paper is organized as follows. Section 2 highlights the background and related work, section 3 briefly illustrates the system model used in our protocol, section 4 presents the optimum number of cluster head, section 5 presents the MADM techniques, section 6 presents the empirical illustrations, sections 7 presents the discussions and simulations results and finally section 8 presents the conclusion.

2 Background and Related Work

Several algorithms have been proposed for the selection of cluster heads but very few of them are based on multi attribute decision making approach in wireless sensor networks.

The Low-energy Adaptive Cluster Hierarchy (LEACH) [9] is the most popular routing protocol in WSN based on randomized rotation of the CHs. Each node elects itself as a CH based on a probabilistic scheme and broadcast its availability to all the sensor nodes present in the area. The communication between different nodes is based on the received signal strength and clusters are formed based on the minimum communication cost. The CH present in each cluster performs aggregation of the packets received from all the nodes present in its cluster. Also all the nodes are given a chance to become the CH to balance the over all energy consumption across the network. Although the complexity of LEACH is low, the algorithm is not energy efficient due to irregular distribution of the CHs.

EEHC [10] is another protocol which works in heterogeneous environment in which a percentage of nodes are equipped with more energy than others. The nodes play the role of a cluster head based on weighted election probabilities according to the residual energy. Though the concept of heterogeneity is introduced, this protocol does not consider different parameters for the selection of CHs. The Hybrid Energy Efficient Distributed Protocol (HEED) [11] is another single-hop clustering protocol in which CHs are selected based on a hybrid metric consisting of residual energy and neighbors proximity. Nodes having high residual energy and operates under low communication cost can become CHs. Multiple CHs are used for transferring the data to the base station if a particular CH is far apart using the concept of multi-hop communication. But HEED cannot guarantee the optimum number of elected CHs.

Another algorithm based on AHP (Analytical Hierarchy Process) [12] is a centralized CH selection scheme using Multi Criteria Decision Making Approach approach to select appropriate CHs. The factors contributing to the network lifetime are Residual energy, mobility and the distance to the involved cluster centroid. CHs are selected in each cycle based on the mobility and the remaining energy of the nodes. It is shown that the AHP approach can improve the network lifetime remarkably, especially for differentiated initial energy of nodes.

EECS [13] is a multi criteria approach for the selection of CHs and the formation of clusters based on three factors including residue to dual energy, distance between node the CH and distance between CH and BS. Thus a cost factor is calculated for associating nodes with CHs. An overhead of this scheme is the wastage of energy due to sending of messages by all the nodes in the network. This technique is further modified in MOECS [7], which considers a multi-criterion optimization for the formation of clusters only. An optimum number of cluster heads are chosen and nodes join a cluster based on the multiple attributes such as residual energy and distances thereby utilizing the local information only.

In the proposed study, the concept of multi criterion optimization is used in the selection of cluster heads unlike in the case of MOECS, where it is done for the formation of clusters. A Pareto optimal approach is used to find out a set of cluster heads which are further ranked in each cycle using TOPSIS. An optimum number of the best clusters heads are selected after the ranking and clusters are formed by attracting nodes with maximum received signal strength.

3 System Model

The following assumptions are made for our network;

1. Nodes are distributed randomly in a 100x100 square region following a uniform distribution.
2. The initial number of clusters is fixed by taking the optimum value (discussed in section 4) and keeps on varying with the node density once the nodes starts dying. The smaller clusters merge with the bigger ones.
3. The BS is a node who is responsible for gathering the entire data from all the CHs and has no energy constraint.
4. A simple radio energy dissipation model [9] in transmitting a *k bit* message over a distance *d* to achieve an acceptable Signal-to-noise ratio (SNR) is used. Energy consumed for transmission is given by

$$E_{TX} = \begin{matrix} k * E_{elec} + k * \varepsilon_{fs} * d^2 \ \ if \ \ d \le do \\ k * E_{elec} + k * \varepsilon_{mp} * d^4 \ \ if \ \ d \ge do \end{matrix}, \tag{1}$$

where E_{elec} is the energy dissipated per bit to run the transmitter or the receiver circuit, ε_{fs} and ε_{mp} are the energy consumed in the amplifier and depend on the amplifier model. By equating the two expressions at $d=d_o$, we have,

$$d_o = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}}$$

The energy consumed while reception is

$$E_{RX} = k * E_{elec} \tag{2}$$

4 The Selection Procedure for Optimum Number of Cluster Heads

In each cycle, it is very important to decide the numbers of clusters present in the area for maximizing the energy efficiency. Each cluster has a cluster head that is responsible for the data aggregation of the data received from its cluster members and does not take part in the sensing operation. For our experiment, two ranges of distances between nodes and base station are observed when the base station is placed firstly in the centre of the field and secondly far away from the field: 2 m < $\text{dist}_{\text{toBS}}$ < 70 m and 75 m < $\text{dist}_{\text{toBS}}$ < 182 m. An estimate for the optimum number of clusters, k_{opt} [14] is given by

$$k_{opt} = \sqrt{\frac{\varepsilon_{fs}}{\pi(\varepsilon_{mp} d_{toBS}^{4} - E_{elec})}}.M\sqrt{N} \tag{3}$$

Using the above equation, we calculate the optimum number of clusters to be 9< k_{opt} <11 when the base station is placed away from the field. Also in Fig. 1, we

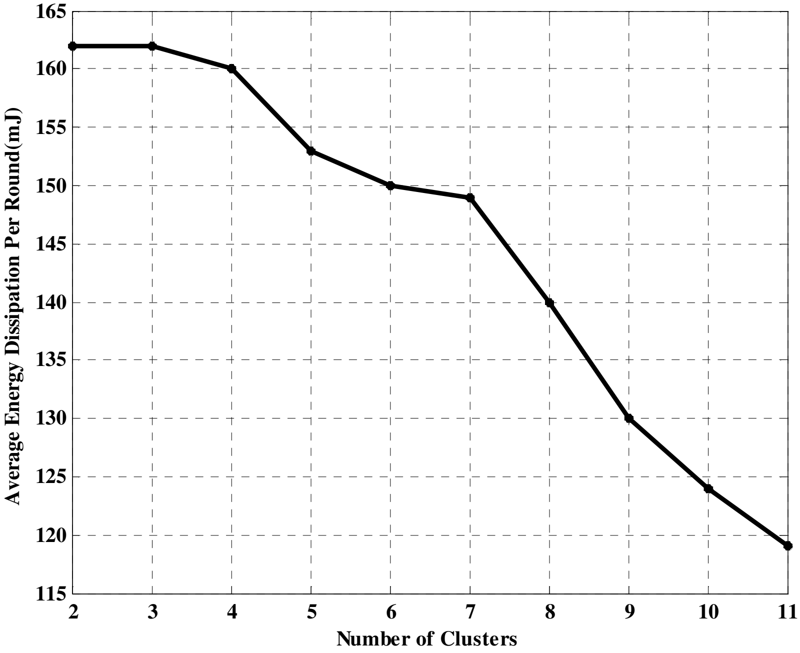


Fig. 1. Average Energy per cycle versus number of clusters in TOPSIS

have shown the average energy dissipated per cycle as a function of the number of clusters by two MADM approaches. It is observed that as the number of clusters increases the average energy dissipation decreases. The simulation results of energy dissipation cover the range as obtained through analytical results of equation (3). Thus for this algorithms, we set the number of clusters to be 10.

5 The MADM Techniques

The Multiple Attribute Decision Making (MADM) methods have been widely used to solve a variety of uncertainty problems. MADM models are capable of selecting the best alternative out of a given list of alternatives based on their prioritized attributes. There is always some extent of uncertainty in these clustered algorithms as which nodes should be chosen as CHs and what criteria should be adopted? In the present study, a well known MADM technique (TOPSIS) is used rank the cluster heads. The three attributes which are residual energy, number of neighbors and distance of nodes from the base station are chosen in some percentage for the selection of CHs. Nodes with higher residual energy, maximum number of neighbors and lesser distance from the base stations are given priority to become a CH. In each cycle, the entire network is re-clustered with the fresh selection of CHs on the basis of ranking done by TOPSIS. Once CHs are selected, each node will join a particular CH based on minimum distance and clusters are formed. Further the algorithm is divided into cycles composed of setup and data transmission phases. In setup phase, cluster heads are selected using a MADM technique and clusters are formed. In the data transmission phase, all the nodes send the data to their respective cluster heads, which is further transferred to the base station by the CH after data aggregation.

5.1 Initialization

Initially each node sends their location information of co-ordinates (location), residual energy and distance from the base station to the base station. The information received from all the nodes is processed and stored as separate records by the base station along with their status of being dead or alive. If a node has expired all of its energy after few rounds of data transmission, then it is declared as dead. Further the number of neighbors within the close vicinity of each node is also estimated and stored by the base station.

5.2 Cluster Head Selection Techniques

A set of cluster heads are selected using Pareto optimal theory using three criteria (mentioned earlier). Since the number of cluster heads given by Pareto front is not fixed in each cycle, they are further ranked using TOPSIS and an optimum number of top ranked CHs are chosen for clustering.

5.2.1 Pareto-Optimal Solution

The Pareto-optimal solutions [15,16] are non-dominated in a given solution space as described by economist Vilfredo Pareto. In multi objective decision making problems, the solution space is defined as a region consisting of all possible solutions (real and otherwise). Solution space can be classified into three sets namely a) Completely dominated, b) Neither dominated nor dominating and c) Non-dominated. In a completely dominated solution there exists at least one (real) alternative which completely overshadows all the properties of all the alternatives in a desirable manner. In the second type of set the alternatives have properties some of which are dominated by the others while the rest are dominating, thus, they are also not ideal for application. Non-dominated solutions are the alternatives that have the best trade-off between properties and are not dominated by any other alternative in the solution space.

In one-dimensional problem, there exists only one such alternative that satisfies the Pareto-optimality test. However, most of the engineering problems are multi-dimensional in nature. Various multi-objective evolutionary algorithms (MOEAs) are extensively investigated for Pareto-optimal solution in multi-objective decision making problems. In the present study, the cluster heads are obtained through Pareto-optimal solutions.

The Pareto solution for the selection of cluster based on three criteria residual energy of the node, minimum distance from the base station and number of neighbors is shown in Fig 2. It is observed that cluster heads (shown in red dots in Fig 2) are selected based on three criteria with maximum energy and neighbors and minimum distance from the base station. In this paper, we have used three criteria for the better selection of cluster heads.

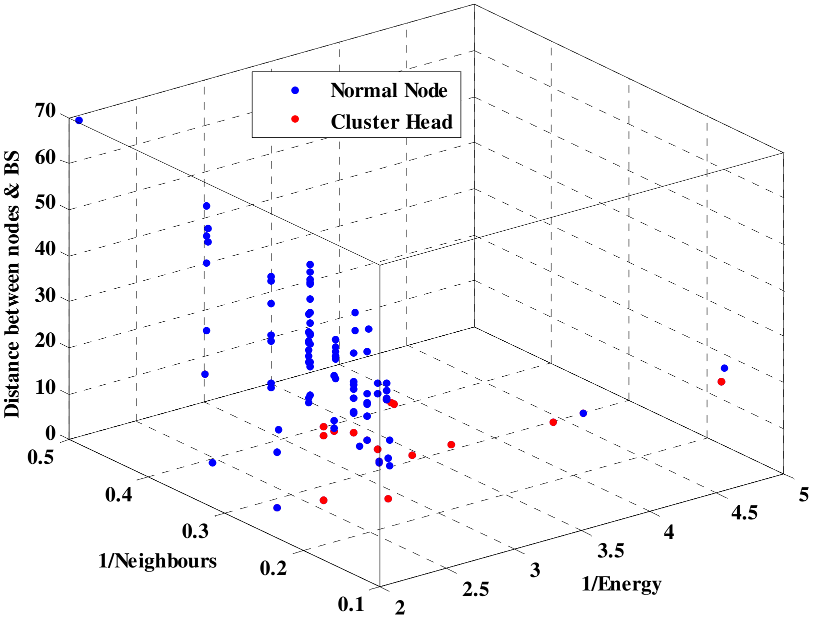


Fig. 2. Pareto-optimal solutions using three criteria

5.2.2 Ranking Using MADM Approaches

The number of CHs selected using Pareto optimal solution is different for each cycle. These CHs are further ranked using the TOPSIS and the top CHs with highest indexes are selected in an optimum number described in section 4. The three MADM techniques are described in details as follows.

TOPSIS Model

Hwang and Yoon first suggested TOPSIS in 1981 in which a decision matrix having '*m*' alternatives and '*n*' attributes can be assumed to be problem of '*n*' dimensional hyperplane having '*m*' points whose location is given by the value of their attributes. The chosen alternative should have the farthest distance from the positive ideal solution (best possible case) and the shortest distance from the negative ideal solution (worst possible case) respectively. This technique has been widely applied in various research applications [17,18,19]. The TOPSIS method involves the following steps:

Step 1: Obtain the normalized decision matrix;

$$r_{ij} = \frac{X_{ij}}{\sqrt{\left(\sum_{i=1}^{m} X_{ij}^{2}\right)}}\ ;\ \forall j \tag{4}$$

Where r_{ij} indicates the normalised value of alternative A_i w.r.t. criterion C_j.
Step 2: Obtain the weighted normalised decision matrix;

$$V_{ij} = [r_{ij}]_{mxn} * [W_j] \tag{5}$$

where W_j is the weight of the j_{th} criterion and $\sum_{j=1}^{n} W_j = 1$.

Step 3: The selection of an alternative using TOPSIS method is based on the shortest distance from the positive ideal solution (A^+) and the farthest from the negative-ideal solution (A^-), which are defined as;

$$A^{+} = \left[(\max(V_{ij}), j \in J_1), (\min(V_{ij}), j \in J_2), i = 1,2,3,....m\right]; \forall j \tag{6}$$

$$A^{-} = \left[(\min(V_{ij}), j \in J_1), (\max(V_{ij}), j \in J_2), i = 1,2,3,....m\right]; \forall j \tag{7}$$

where, J_1 corresponds to benefit criteria and J_2 corresponds to cost criteria.
Step 4: Separation measures which are measured using Euclidean distance from the positive and negative ideal solutions are;

$$S_i^{+} = \left[\sum_{j=1}^{m} \left(V_{ij} - V_j^{+}\right)^{2}\right]^{0.5},\ \text{i=1,2,3...} \tag{8}$$

$$S_i^{-} = \left[\sum_{j=1}^{m} \left(V_{ij} - V_j^{-}\right)^{2}\right]^{0.5},\ \text{i=1,2,3,...} \tag{9}$$

Step 5: Obtain relative closeness of alternatives to the ideal solution;

$$C_i^+ = \frac{S_i^-}{S_i^- + S_i^+} \tag{10}$$

In each cycle of data transmission, C_i^+ is calculated for all the cluster heads obtained through Pareto theory, where larger value indicates better performance of the alternative. Thus all the CHs are ranked in decreasing order of C_i^+ and the best CHs are chosen from the top.

5.3 Cluster Formation and Data Transmission

All the selected CHs now send advertisement messages in the network declaring their presence as cluster heads. Each node now measures distance from all the cluster heads, form a vector having ten entries (optimum value for number of CHs is taken as 10 here). The node joins the CH with minimum distance and sends a message to the nearest cluster head. If the distance between the node and the CH is more than its distance to the BS, the node will communicate with the BS directly. Otherwise it joins cluster based on the nearest distance (Euclidean distance), thereby forming clusters. The nodes are re-clustered based on the distance with the selected cluster head using a distance matrix, DM (m x n) given as follows;

$$DM = \begin{bmatrix} d_{CH1,x1} & d_{CH1,x2} & .. & d_{CH1,xn} \\ d_{CH2,x1} & d_{CH2,x2} & .. & d_{CH2,xn} \\ \vdots & \vdots & \vdots & \\ d_{CHm,x1} & d_{CHm,x2} & d_{CHm,x3} & d_{CHm,xn} \end{bmatrix} \tag{11}$$

where *d* is the Euclidean distance between CH and a node based on its location information. If *y* and *z* represent the location of two nodes p and q, then the Euclidean distance is

$$d_{p,q} = [(p_x - q_x)^2 + (p_y - q_y)^2]^{1/2} \tag{12}$$

Each element $d_{i,j}$ in the distance matrix represents the distance between the i^{th} clusterhead and j^{th} node. The column containing the minimum value represents the cluster number to be joined by the corresponding node. For example, if $d_{CH2,x1}$ is the minimum value in the first column, in this situation the node x_1 gets associate with the second cluster where CH2 is cluster head.

Once the clusters are formed, the CH assigns a time slot for each member after receiving all CH_join messages from all the nodes. Each cluster head is responsible for gathering the data from all the nodes in the cluster. When a frame of data from all the members is received, the CH send the frame to the base station after applying data aggregation. The CH must remain in active state while the member nodes can go to sleep mode from time to time. It is to be noted that the re-clustering methodology is also adopted in LEACH protocol where CHs are elected by using the probabilistic approach rather than deterministic technique. The operation of re-clustering and data transmission continues for many cycles until the death of all the nodes. If the size of the cluster is smaller than the predefined threshold, the cluster merges with the neighboring clusters. With the start of the death of nodes, it is found that there are a lesser number of nodes present in each cluster now. Thus as the number of alive nodes starts decreasing with cycles, the number of clusters also decreases and the decrease in the number of alive nodes eventually results in the reduction in number of clusters. The amount of information also decreases with the fewer nodes left in the physical area.

6 Empirical Illustrations

TOPSIS Model

In each cycle of algorithm, new CHs are selected and clusters are formed. We consider a specific cycle (cycle no-02) for the empirical analysis of Topsis method in which 14 cluster heads are short listed by Pareto optimal solutions and the data is given in Table 1. Firstly the first two factors Eo and n are taken in reciprocal so that all the three criteria can be used in decreasing order in Pareto solutions given in Table 2. Based on the first step of the TOPSIS procedure, each element is normalized by Eq. (4). The resulting normalized decision matrix for the TOPSIS analysis is shown as Table 3. Table 4 finally shows the final results of step 3, 4 & 5, which calculates positive (**A**$^+$) and negative (**A**$^-$) ideal solutions, distances of each CH from them and

Table 1. Decision Matrix for TOPSIS Analysis in second Cycle

Cluster Head No.	Residual Energy, Eo (Joules)(C1)	Number of Neighbors, n (C2)	Distance from Sink, d (C3)
CH1	0.4998	7	21.5830
CH2	0.4998	9	24.2745
CH3	0.4887	8	20.9972
CH4	0.4894	10	39.3231
CH5	0.4998	6	23.2092
CH6	0.4998	10	39.8408
CH7	0.4998	3	7.6944
CH8	0.4988	4	4.5873
CH9	0.4998	9	25.5698
CH10	0.4947	6	10.6745
CH11	0.4964	5	9.8442
CH12	0.4998	4	16.6420
CH13	0.4919	9	24.2008
CH14	0.4998	6	17.6095

Table 2. First level normalized decision matrix for TOPSIS Analysis

Cluster Head No.	Reciprocal of Eo, 1/Eo (Joules)	Reciprocal of n	Distance from Sink, d
CH1	2.0009	0.1429	21.5830
CH2	2.0009	0.1111	24.2745
CH3	2.0462	0.1250	20.9972
CH4	2.0434	0.1000	39.3231
CH5	2.0008	0.1667	23.2092
CH6	2.0008	0.1000	39.8408
CH7	2.0009	0.3333	7.6944
CH8	2.0049	0.2500	4.5873
CH9	2.0008	0.1111	25.5698
CH10	2.0213	0.1667	10.6745
CH11	2.0147	0.2000	9.8442
CH12	2.0009	0.2500	16.6420
CH13	2.0327	0.1111	24.2008
CH14	2.0008	0.1667	17.6095

Table 3. Second level normalized decision matrix for TOPSIS Analysis using Eq 4

Cluster Head No.	C1	C2	C3
CH1	0.2658	0.2124	0.2529
CH2	0.2658	0.1652	0.2844
CH3	0.2718	0.1858	0.2460
CH4	0.2714	0.1487	0.4607
CH5	0.2658	0.2478	0.2719
CH6	0.2658	0.1487	0.4668
CH7	0.2658	0.4955	0.0901
CH8	0.2663	0.3716	0.0537
CH9	0.2658	0.1652	0.2996
CH10	0.2685	0.2478	0.1251
CH11	0.2676	0.2973	0.1153
CH12	0.2658	0.3716	0.1950
CH13	0.2700	0.1652	0.2835
CH14	0.2658	0.2478	0.2063
Weight	0.5	0.25	0.25

relative closeness (C_i^+) given in Table 5. The CH with higher C_i^+ in TOPSIS are chosen and given the ranks as given below-

CH10 > CH11 > CH14 > CH3 > CH8 > CH1 > CH13 > CH2 > CH9 > CH5 > CH12 > CH7 > CH4 > CH6

Our attention should focus on the top few optimum choices as derived in section 4, according to which the top ten CHs should be picked and ten clusters are formed. This scenario will be repeated in every cycle during the setup phase where CHs are selected and clusters are formed.

Table 4. TOPSIS Analysis Results

Cluster Head No.	C1	C2	C3	S_i^+	S_i^-	$C_i^+ = \frac{S_i^-}{S_i^- + S_i^+}$
CH1	0.1329	0.0531	0.0632	0.0523	0.0	0.6294
CH2	0.1329	0.0413	0.0711	0.0578	0.0	0.6202
CH3	0.1359	0.0465	0.0615	0.049	0.0	0.6597
CH4	0.1357	0.0372	0.1152	0.1018	0.0	0.4601
CH5	0.1329	0.0619	0.0680	0.0599	0.0	0.5683
CH6	0.1329	0.0372	0.1167	0.1033	0.0	0.4566
CH7	0.1329	0.1239	0.0225	0.0872	0.0	0.5193
CH8	0.1331	0.0929	0.0134	0.0557	0.1	0.6592
CH9	0.1329	0.0413	0.0749	0.0616	0.0	0.6006
CH10	0.1342	0.0619	0.0313	0.0306	0.1	0.7755
CH11	0.1338	0.0743	0.0288	0.0402	0.1	0.7149
CH12	0.1329	0.0929	0.0487	0.066	0.0	0.5311
CH13	0.1350	0.0413	0.0709	0.0576	0.0	0.6210
CH14	0.1329	0.0619	0.0516	0.0455	0.0	0.6641
A^+	0.1329	0.0372	0.0134			
A^-	0.1359	0.1239	0.1167			

7 Discussions and Simulations Results

The simulator is developed in MATLAB in which TOPSIS is used for the selection of cluster heads and clusters are formed. A network model similar to [9] is used in which operation progresses in cycles. Table 5 provides the simulation parameters used in our experiments. Each cycle consists of clustering and data transmission phase. In clustering phase, the top ten CHs according to the three criteria are summarized in Table 6.

Table 5. Simulation parameters for transmission

Description	Symbol	Value
Number of nodes in the system	N	100
BS Location	-	(50, 175)
Size of the data packet	-	500 bytes
Hello / Broadcast / CH_Join message	-	25 bytes
Energy consumed by the amplifier to transmit at a short distance	ε_{fs}	10 pJ/bit/m^2
Energy consumed by the amplifier to transmit at a longer distance	ε_{mp}	0.0013 pJ/bit/m^4
Energy consumed in the electronics circuit to transmit or receive the signal	E_{elec}	50 nJ/bit

Table 6. Top ten CHs in 2nd cycle from TOPSIS

Rank	Cluster Head
1	CH10
2	CH11
3	CH14
4	CH3
5	CH8
6	CH11
7	CH13
8	CH2
9	CH9
10	CH5

In each cycle of algorithm, re-clustering is done to select the best optimum number of cluster heads as shown in Table 7 for the second cycle and the same process is repeated for every cycle till all the nodes expire their entire energy. The three chosen attributes for the selection of CHs are taken in appropriate proportions and the average results of network lifetime are considered.

In the present study, a MADM approach called TOPSIS is simulated taking the three attributes into account and compared with LEACH protocol. The base station is placed far away from the field. The lifetime of the network is measured in terms of number of cycles until the first node in the network runs out of its entire energy. Fig 3

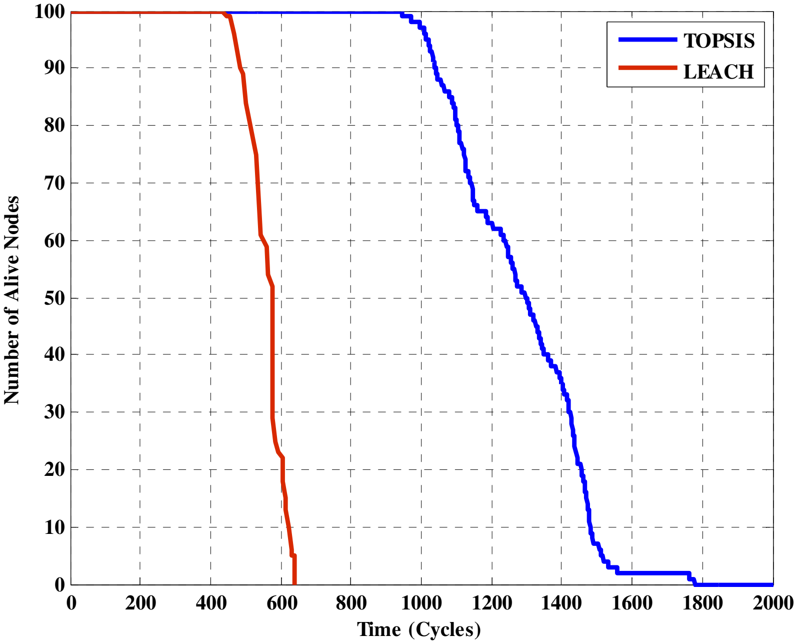


Fig. 3. Network Lifetime in cycles

shows the results of the experiment, where sensor nodes are deployed randomly on a square area of 100x100 m and network lifetime is plotted, which shows the number of alive nodes over the time in cycles. All results are expressed in averages taken over 20 random independent experiments. It is shown that the network lifetime (when first node dies) of TOPSIS (946 cycles) is 117 % higher than LEACH. It can be observed that TOPSIS outperforms Leach by a good margin in terms of network lifetime. Also the stability region remains highest in TOPSIS as compared to Leach.

8 Conclusions

In this paper, we have presented a method to select and rank cluster heads using TOPSIS in WSNs. Further it is compared with LEACH in terms of network lifetime. The optimum numbers of cluster heads are selected by using the ranking done by TOPSIS and clusters are formed. Simulations results demonstrate that TOPSIS achieves significant energy savings and enhances network lifetime compared to LEACH. In the proposed methodology, we have considered three attributes in different proportions of weights and average results are plotted. In future, we will consider more attributes for the cluster head selection.

References

1. Aslan, Y.E., Korpeoglu, I., Ulusoy, O.: A framework for use of wireless sensor networks in forest fire detection and monitoring. Computer, Environment and Urban Systems 36, 614–625 (2012)
2. Komar, C., Donmez, M.Y., Ersoy, C.: Detection quality of border surveillance wireless sensor networks in the existence of trespassers' favorite paths. Computer Communications 35, 1185–1199 (2012)
3. Rahimi, M., Baer, R., Iroezi, O., Garcia, J., Warrior, J., Estrin, D., Srivastava, M.: Cyclops: in situ image sensing and interpretation in wireless sensor networks. In: Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys), San Diego, CA (2005)
4. Corchado, J.M., Bajo, J., Tapia, D.I., Abraham, A., Abraham, A.: Using heterogeneous wireless sensor networks in a telemonitoring system for healthcare. IEEE Transactions on Information Technology in Biomedicine 14(2), 234–240 (2010)
5. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: A survey. Computer Networks 38, 393–422 (2002)
6. Chaudhry, S.B., Hung, V.C., Guha, R.K., Stanley, K.O.: Pareto-based evolutionary computational approach for wireless sensor placement. Engineering Applications of Artificial Intelligence 24, 409–425 (2011)
7. Aslam, N., Phillips, W., Robertson, W., Sivakumar, S.: A multi-criterion optimization technique for energy efficient cluster formation in wireless sensor networks. Information Fusion 12, 202–212 (2011)
8. Soltanpanah, H., Farughi, H., Golabi, M.: Utilization and comparison of multiple attribute decision techniques to rank countries upon human development rate. Int Res. J. Finance Econ. 60, 175–188 (2010)

9. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1(4), 660–670 (2002)
10. Kumar, D., Aseri, T.C., Patel, R.B.: EEHC: Energy efficient heterogeneous clustered scheme for wireless sensor networks. Computer Communications 32, 662–667 (2009)
11. Younis, O., Fahmy, S.: HEED:A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Transactions on Mobile Computing 3(4), 366–379 (2004)
12. Yin, Y.Y., Shi, J.W., Li, Y.N., Zhang, P.: Cluster head selection using analytical hierarchy process for wireless sensor networks. In: IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC (2006)
13. Ye, M., Li, C.F., Chen, G.H., Wu, J.: EECS: an energy efficient clustering scheme in wireless sensor networks. In: IEEE International Performance Computing and Communications Conference (IPCCC), pp. 535–540 (2005)
14. Comeau, F., Sivakumar, S.C., Robertson, W., Phillips, W.J.: Energy conserving architectures and algorithms for wireless sensor networks. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, vol. 9 (2006)
15. Kasprzak, E.M., Lewis, K.E.: Pareto Analysis in multiobjective optimization using the colinearity theorem and Scaling Method. Structural and Multidisciplinary Optimization 22, 208–218 (2001)
16. Chaudhry, S.B., Hung, V.C., Guha, R.K., Stanley, K.O.: Pareto-based evolutionary computational approach for wireless sensor placement. Engineering Applications of Artificial Intelligence 24, 409–425 (2011)
17. Chauhan, A., Vaish, R.: Magnetic material selection using multiple attribute decision making approach. Materials and Design 36, 1–5 (2012)
18. Rathod, M.K., Kanzaria, H.V.: A methodological concept for phase change material selection based on multiple criteria decision analysis with and without fuzzy environment. Materials and Design 32, 3578–3585 (2011)
19. Yang, T., Hung, C.: Multiple-attribute decision making methods for plant layout design problem. Robotics and Computer-Integrated Manufacturing 23, 126–137 (2007)

Key Pre-distribution in a Non-uniform Network Using Combinatorial Design

Sarbari Mitra and Sourav Mukhopadhyay

Indian Institute of Technology, Kharagpur, India
{sarbari,sourav}@maths.iitkgp.ernet.in

Abstract. In this paper we propose a key pre-distribution scheme using combinatorial design. The network is assumed to be heterogeneous and the rectangular grid structure of the network is non-uniform. During key distribution phase, nodes are placed in the rectangular grid to form a virtual network. The actual network consists of the nodes along with their location in the target region. However, the deployment strategy and the key distribution technique indicate high connectivity of the network. Our scheme demonstrates superior performance compared to the existing similar schemes.

Keywords: key pre-distribution, projective planes, pairwise connectivity, resilience.

1 Introduction

Wireless Sensor network [WSN] is a collection of spatially distributed small, battery-powered, low-cost devices, with limited constraints to transmit data within a specified radio frequency range, known as wireless sensor nodes. Initially the evolution of WSN were motivated by the military applications but now-a-days it plays an active role in industrial application areas, healthcare machines, traffic control etc. Sensor nodes are densely distributed in the intended region for monitoring physical and environmental conditions. They gather information from the environment and actively transmit the collected data to the desired location through the network by communicating among themselves. The location of the sensor nodes in the network is not predetermined as they are deployed from air crafts; hence keys are assigned first to them, before deployment. The method of assigning secret keys to the nodes prior to their deployment in the target region is termed as key pre-distribution. The salient features of a good key pre-distribution scheme (KPS) includes less memory, less computation, greater connectivity and high robustness of the network against node capture.

Random Key Pre-distribution Scheme (KPS) was introduced by Eschenauer and Gligor in 2002 [6]. Combinatorial designs have become one of the most useful mathematical tools for KPS. *Projective planes* are used in [3]. Transversal design based schemes were proposed by Lee and Stinson in [9, 10], which were extended in [4] by merging blocks to construct nodes. Partially Balanced Incomplete Blocks Designs were used in [12] and Codes in [13] for KPS. For details of other combinatorial design based KPS we refer to the surveys [5, 11].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 155–170, 2013.

A storage-efficient key pre-distribution scheme for a non-uniform rectangular grid structured network is presented in this paper. The nodes are placed at the intersection of the rows and the columns of the grid. Keys are distributed in such a manner that all the nodes on a row and a few columns form projective planes. There exists at least one path of length less than or equal to three, between any two nodes, provided they lie within radio frequency range. However, the key-path between any two nodes is not unique, a sufficient number of paths (of equal or larger length) exists between them. This increases the probability of the two nodes being connected, even after a number of nodes are compromised. With small storage the scheme induces a network of sufficiently large size. We emphasize that apart from storage efficiency, our design also provides better resilience and reasonable connectivity as compared to other existing schemes based on combinatorial designs. Moreover, given the size of the network, the number of rows and columns can be suitably chosen so as to get a desirable trade-off between the evaluation parameters, e.g., storage, connectivity and resilience.

2 Preliminaries

Definition 2.1. A *set-system* is defined as a pair (X, A) such that
(i) X is a set of points or elements,
(ii) A is a subset of the power set of X (i.e. collection of non-empty subsets or blocks of X).
The *degree* (denoted by r) of $x \in X$ is the number of blocks of A containing x; the *rank* (denoted by k) is the size of the largest block in A.
(X, A) is said to be *regular* and *uniform* if all the points in X have the same degree and all the blocks in A have the same size respectively. A regular, uniform set-system with $|X| = v$, $|A| = b$ is known as a (v, b, r, k)*-design* .

Definition 2.2. A (v, b, r, k)-design in which any set of t points is contained in exactly λ blocks, is known as a t - (v, b, r, k, λ)*-design* which is often denoted as t - (v, k, λ)*-design.*

Definition 2.3. A symmetric 2 - $(n^2 + n + 1, n^2 + n + 1, n + 1, n + 1, 1)$-design is known as a *finite symmetric projective plane of order* n. Precisely, it is a pair of a set of $\left(n^2 + n + 1\right)$ points and a set of $\left(n^2 + n + 1\right)$ lines, where each line contains $(n + 1)$ points and each point occurs in $(n + 1)$ lines.

3 Proposed Scheme

3.1 Protocol Requirements

- The nodes are arranged in a virtual rectangular grid during key pre-distribution phase. After key distribution, the nodes are deployed in the target region that the nodes sharing common keys are placed together, so that they lie within radio frequency range.
- It is assumed that all the nodes are not identical. One-third of the total number of nodes are more powerful than the rest. Comparatively "more powerful" nodes have higher radio frequency range and more storage capacity.

3.2 Terminologies and Notations

- Two nodes in the network are said to be *key-connected* or *physically-connected* when they share a common key or lie within radio range of each other respectively. A pair of nodes is said to be connected if it is both key-connected and physically-connected.
- Two nodes are said to be *pairwise connected*, when the common key between them is not assigned to any other node in the network.
- *Combinatorially Complete Graph*: As the name suggests, here we shall make use of a combinatorial design, namely projective planes, to form a complete graph. A projective plane of order p is used in such a way that the graph representing the network containing n nodes is complete with only $p+1$ keys assigned to each of the nodes, where $n \approx p^2+p+1$, for some prime power p, as we discussed in the previous section. The graph thus obtained is referred as a combinatorially complete graph. We say a set of nodes is combinatorially complete, if any two nodes of the set are key-connected.

The distance function $d(\cdot,\cdot)$ is different from the conventional distance function. the distance between two nodes depends only on the keys stored at them, not on their physical location. The distance function, on the virtual network, is defined as follows:

Define a graph G=(V,E), where V=$\{$Sensor nodes$\}$=$\{N_1, N_2, \cdots, N_N\}$, say, with $\mid V \mid$=N. $(N_1, N_2) \in E$ if N_1 and N_2 are key-connected.
Define $d(N_1, N_2) = l$, (in other words, there exists an l-hop path between N_1 and N_2) if $\exists$ a path $N_1 N_{u_1} N_{u_2} \cdots N_{u_{l-1}} N_2$ in G where $(N_1, N_{u_1}) \in E, (N_{u_{l-1}}, N_2) \in E$ and $(N_{u_i}, N_{u_{i+1}}) \in E$, for $i = 1, 2, \cdots, l-2$.

We use the following notations throughout the paper.

r	Number of rows in the network
c	Number of columns on the network
N	Total number of nodes in the network
k	Average number of keys stored at each node in the network
$N_{i,j}$	The node belonging to the i^{th} row and j^{th} column of the network, where $i \in \{1,2,\ldots,r\}$ and $j \in \{1,2,\ldots,c\}$
R_i	i^{th} row of the grid, for $i \in \{1,2,\ldots r\}$
C_j	j^{th} column of the grid, for $j \in \{1,2,\ldots c\}$
$d(A,B)$	Distance between the nodes A and B in the virtual network
L_i	The number of i-hop paths in the network
s	The number of nodes compromised
$V(s)$	The fraction of nodes that become disconnected
$fail(s)$	Probability that the link between two uncompromised nodes is broken

3.3 Description of the Scheme

1. The whole network comprised of N nodes is distributed into r rows and c columns such that $rc \geq N$ where $c = 3c_1$, for an integer c_1. We choose prime integers p, q such that $r \leq p^2+p+1$, $c \leq q^2+q+1$.

2. Two disjoint key-pools of distinct keys are chosen - one for distributing keys along the rows and the other along the columns. Each node is assigned with keys along the corresponding row and the corresponding column.
3. Each row is made combinatorially complete by considering a projective plane of order q, since there are $c \leq q^2 + q + 1$ nodes on each row. So, any two nodes lying on a row are key-connected.
4. Let C_j be a special column. If $j \equiv 2(\mod 3)$ then C_j is made combinatorially complete by considering a projective plane of order p, since there are $r \leq p^2 + p + 1$ nodes on each column.
5. When $j \neq 2(\mod 3)$, the keys given to the node $N_{i,j}$ (along the column) is $\{j, i(\mod r)\}$ and $\{j, i+1(\mod r)\}$. This implies any two adjacent nodes (and the two lying on the boundary) on this column are key-connected.

For convenience, we refer to the combinatorially complete columns as *Special columns* and the rest of the columns as *ordinary column.* We define two types of nodes in the network

(i) Type A nodes - lie at the intersection of a row and a special column.
(ii) Type B nodes - lie at the intersection of a row and an ordinary column.

It follows from the construction that all the Type A nodes form complete graphs along their corresponding row and column. Whereas, any two adjacent Type B nodes on an ordinary column are pairwise connected. Hence, we assume that Type A nodes are provided with higher transmission range and memory as compared to the Type B nodes.

4 Analysis

The following results are developed in the block graph of the network i.e., the number of multi-hop paths are counted on the basis of the keys stored at each node. Two nodes are key connected in k-hop means that there exists at least one key-path of shortest length k, any two adjacent nodes on that path are key-connected. This path is not unique, i.e., there may exist any other k-hop, even $(k+l)$-hop (for, $l > 0$) paths between those two nodes.

We show the multihop paths from a Type A node in Fig. 1 and from a Type B node in Fig. 2 - Fig. 5 as given below. Diamonds, circles and triangles in Fig. 1 - Fig. 5 respectively denote the nodes who are at a distance of single-hop, 2-hop and 3-hop from $N_{i,j}$.

Theorem 4.1. *Any two nodes in the proposed network are key-connected by at least one path of length at most three.*

Proof. Let the nodes N_{i_1,j_1} and N_{i_2,j_2} wish to communicate. We consider the following cases:

Case (i): Let $i_1 = i_2$ i.e., both the nodes lie on the same row R_{i_1}.
From the construction, N_{i_1,j_1} and N_{i_2,j_2} are directly key-connected, i.e., $d(N_{i_1,j_1}, N_{i_2,j_2}) = 1$.

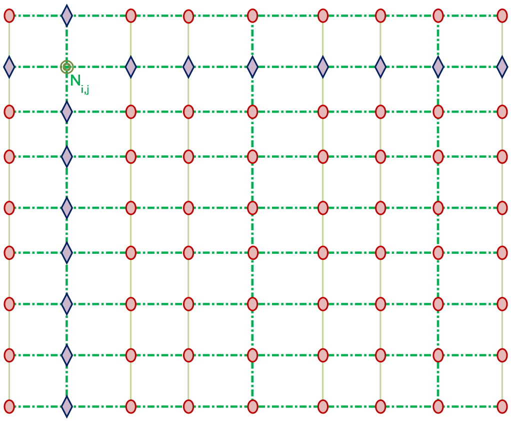


Fig. 1. One-hop and Two-hop paths from a Type A node $N_{i,j}$

Case (ii): Let $i_1 \neq i_2$ and $j_1 = j_2 = 2 \mod 3$ i.e., both the nodes lie on the same special column.

According to our construction, nodes lying on the same special columns form a combinatorially complete graph, hence we must have $d(N_{i_1,j_1}, N_{i_2,j_2}) = 1$.

Case (iii): Let If $i_1 \neq i_2$ and $j_1, j_2 = 2(\mod 3)$ but $j_1 \neq j_2$ i.e., both the nodes lie on the two different special columns.

Now, by case (i) $d(N_{i_1,j_1}, N_{i_1,j_2}) = 1$, and by case (ii) $d(N_{i_1,j_2}, N_{i_2,j_2}) = 1$. Therefore, $d(N_{i_1,j_1}, N_{i_2,j_2}) = d(N_{i_1,j_1}, N_{i_1,j_2}) + d(N_{i_1,j_2}, N_{i_2,j_2}) = 2$.

Case (iv): Let If $i_1 \neq i_2$ and $j_1, j_2 \neq 2(\mod 3)$, $j_1 \neq j_2$.

We consider the following two sub-cases

- Sub-case (a) When $j_2 \equiv 0(\mod 3)$
  We obtain from case (i) $d(N_{i_1,j_1}, N_{i_1,j_2-1}) = 1$, and $d(N_{i_2,j_2-1}, N_{i_2,j_2}) = 1$.
  Since, $j_2 \equiv 0(\mod 3)$, C_{j_2-1} is a special column, and hence by case (ii) we get, $d(N_{i_1,j_2-1}, N_{i_2,j_2-1}) = 1$. Therefore, $d(N_{i_1,j_1}, N_{i_2,j_2}) = d(N_{i_1,j_1}, N_{i_1,j_2-1}) + d(N_{i_1,j_2-1}, N_{i_2,j_2-1}) + d(N_{i_2,j_2-1}, N_{i_2,j_2}) = 3$.
- Sub-case (b) When $j_2 \equiv 1(\mod 3)$
  Hence,C_{j_2+1} is a special column. Proceeding in the similar manner as in the subcase(a), just by replacing $j_2 - 1$ by $j_2 + 1$, we obtain $d(N_{i_1,j_1}, N_{i_2,j_2}) = d(N_{i_1,j_1}, N_{i_1,j_2+1}) + d(N_{i_1,j_2+1}, N_{i_2,j_2+1}) + d(N_{i_2,j_2+1}, N_{i_2,j_2}) = 3$.

In either of the two sub-cases, there exists at leats one 3-hop path between the nodes.

Considering all the case, it is found that there exist at leats one shortest path of length less than or equal to three, between any two randomly chosen nodes, in the block graph of the network. □

Theorem 4.2. *Total number of one-hop paths in the network is* $L_1 = \frac{rc}{6}(r+3c)$.

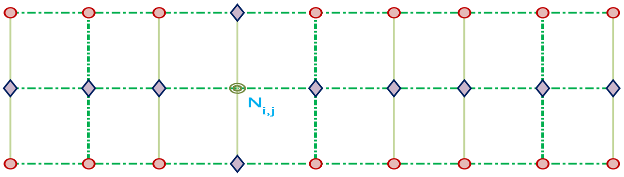


Fig. 2. One-hop and Two-hop paths from a Type B node $N_{i,j}$ when there are exactly three rows in the network

Proof. From Fig. 1 it follows that a Type A node $N_{i,j}$ is key-connected to all the nodes lying on the row R_i (i.e, $N_{i,j'}$ for $j^{'} = 1, 2, \cdots, i-1, i+1, \cdots, r$) and the column C_j (i.e., $N_{i',j}$ for $i^{'} = 1, 2, \cdots, j-1, j+1, c$). Hence, it is key-connected to $r + c - 2$ nodes. Again, a Type B node $N_{i,j}$ is key-connected to all the nodes lying on the row R_i (i.e., $N_{i,j'}$ for $j^{'} = 1, 2, \cdots, i-1, i+1, \cdots, r$) and the two adjacent nodes (i.e., $N_{i-1,j}$ and $N_{i+1,j}$) lying on the column C_j. Which implies that a Type B node is key-connected to $r + c - 2$ nodes.

The total number of Type A and Type B nodes in the network is $\frac{rc}{3}$ and $\frac{2rc}{3}$ respectively. Therefore, the total number of one-hop paths in the network is $\frac{1}{2}\{(r + c - 2)\frac{rc}{2} + (c + 1)\frac{2rc}{3}\}$, the factor 1/2 comes, as each of the single-hop path is counted twice corresponding to each extreme of the path. We obtain the desired result on simplification. □

Theorem 4.3. *The total number of two-hop paths in the network is given by*

$$L_2 = \begin{cases} 0, & \text{if } r < 2; \\ c(c-1), & \text{if } r = 2; \\ 3c(c-1), & \text{if } r = 3; \\ \frac{2c}{3}(\frac{23c}{3} - 5), & \text{if } r = 4; \\ \frac{rc}{6}(\frac{5rc}{3} - r + c + 1), & \text{if } r \geq 5\ . \end{cases}$$

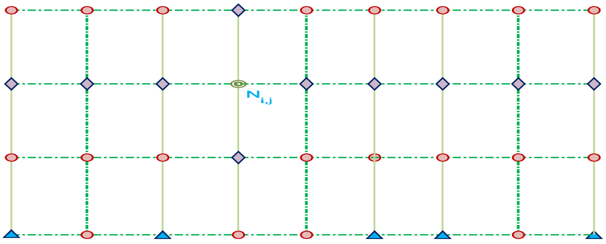


Fig. 3. One-hop, Two-hop and Three-hop paths from a Type B node $N_{i,j}$ when there are exactly four rows in the network

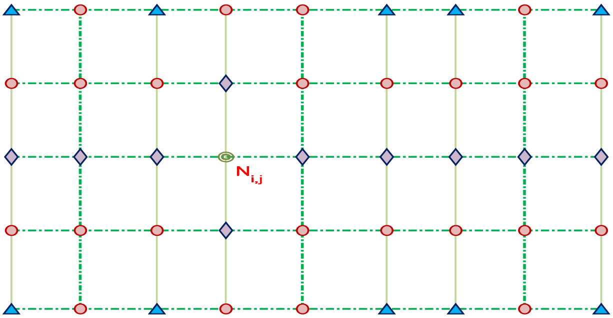


Fig. 4. One-hop, Two-hop and Three-hop paths from a Type B node $N_{i,j}$ when there are exactly five rows in the network

Proof. Let us assume that p_1, p_2 respectively denote the proportion of Type A and Type B nodes, i.e., $p_1 = \frac{rc}{3}$ and $p_2 = \frac{2rc}{3}$. Suppose that the number of nodes to which a Type A and Type B node are connected in 2-hop paths are n_1 and n_2 respectively. Hence we have

$$L_2 = \frac{1}{2}(p_1 n_1 + p_2 n_2) = \frac{rc}{6}(n_1 + 2n_2).$$

From Fig. 1 it follows that a Type A node $N_{i,j}$ is key-connected to all the nodes, in 2-hop path, who lie neither on R_i, nor on C_j. Thus, we have $n_1 = (r-1)(c-1)$. Hence,

$$L_2 = \frac{rc}{6}\{(r-1)(c-1) + 2n_2\}. \tag{1}$$

We observe that the expression for n_2 depends on the number of rows present in the network. We consider the following cases

Case (i) Let $r < 2$

The only possibility is there is only one row in the network. Since each row forms a completely connected graph, all the nodes are key-connected in a direct path, and hence no two-hop path is there. So, $n_2 = 0$.

Case (ii) Let $r = 3$

From Fig. 2 it follows that a Type B node $N_{i,j}$ is at a distance of two, with the nodes given by $N_{i',j'}$ for $i^{'} = \{i+1, i-1\}$ and $j^{'} \in \{1, 2, \cdots, c\} \setminus j$, i.e., $2(c-1)$ nodes. Thus, $n_2 = 2(c-1)$.

Case (iii) Let $r = 4$

From Fig. 3 it follows that a Type B node $N_{i,j}$ is at a distance of two, with the following nodes

1. $N_{i',j'}$ for $i^{'} = \{i+1, i-1\}$ and $j^{'} \in \{1, 2, \cdots, c\} \setminus j$, i.e., $2(c-1)$ nodes
2. $N_{i',j'}$ where $i^{'} = i+2$ (or $i-2$ since, $R_{i-2} = R_{i+2}$ whenever $r = 4$) and $C_{j'}$ is a special column, i.e., $c/3$ nodes.
3. The node $N_{i+2,j}$

Thus, we have $n_2 = 2(c-1) + \frac{c}{3} + 1 = \frac{7c}{3} - 1$.

Case (iv) Let $r \geq 5$

From Fig. 4 and Fig. 5 it follows that a Type B node $N_{i,j}$ is at a distance of two, with the following nodes

1. $N_{i^{'},j^{'}}$ for $i^{'} = \{i+1, i-1\}$ and $j^{'} \in \{1, 2, \cdots, c\} \setminus j$, i.e., $2(c-1)$ nodes
2. $N_{i^{'},j^{'}}$ where $i^{'} \in \{1, 2, \cdots, r\} \setminus \{i-1, i, i+1\}$ and $C_{j^{'}}$ is a special column, i.e., $\frac{c}{3}(r-3)$ nodes
3. The two nodes $N_{i-2,j}$ and $N_{i+2,j}$

Thus, we have $n_2 = 2(c-1) + \frac{c}{3}(r-3) + 2 = (\frac{rc}{3} + c)$.

Substituting the obtained values in each case for n_2 in equation (1) we get the desired expression as given in the statement of Theorem 4.3. □

Theorem 4.4. *The total number of three-hop paths in the network is given by*

$$L_3 = \begin{cases} 0, & \text{if } r \leq 3; \\ \frac{4c}{3}(\frac{2c}{3} - 1), & \text{if } r = 4; \\ \frac{5c}{3}(\frac{4c}{3} - 2), & \text{if } r = 5; \\ \frac{2rc}{3}(\frac{rc}{3} - c - 1), & \text{if } r > 5\ . \end{cases}$$

Proof. In this case, we have $p_1 = \frac{rc}{3}$ and $p_2 = \frac{2rc}{3}$, as the previous theorem. We further assume that the number of nodes to which a Type A and Type B node are connected in 3-hop paths is m_1 and m_2 respectively. Hence we have

$$L_3 = \frac{1}{2}(p_1 m_1 + p_2 m_2) = \frac{rc}{6}(m_1 + 2m_2).$$

Now, from Fig. 1, we notice that any Type A node is key-connected to all the nodes of the network, i.e., a Type A node has no 3-hop path as the smallest

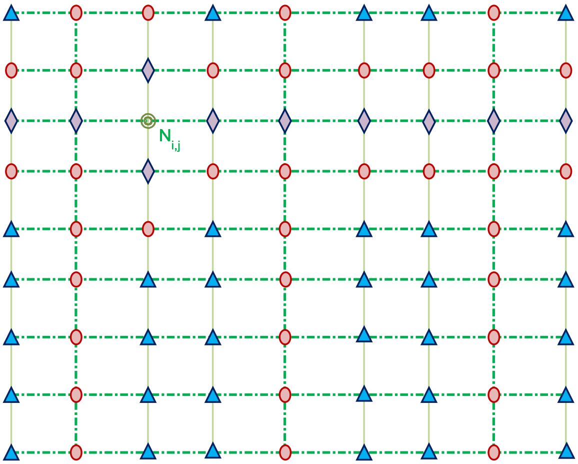


Fig. 5. One-hop, Two-hop and Three-hop paths from a Type B node $N_{i,j}$ when there are more than five rows in the network

path. So, we have $m_1 = 0$. Hence, we obtain

$$L_3 = \frac{rc}{3} m_2. \tag{2}$$

We observe that the expression for m_2 depends on the number of rows present in the network. We consider the following cases

Case (i) Let $r \leq 3$

From Fig. 2, we observe that if there are less than three rows in the network, then there will be no 3-hop path as the smallest path between any two nodes. Thus, in this case $m_2 = 0$.

Case (ii) Let $r = 4$

From Fig. 3 it follows that a Type B node $N_{i,j}$ is at a distance of two, with the nodes given by $N_{i',j'}$ for $i' = i + 2$ (or $i - 2$ since, $R_{i-2} = R_{i+2}$, when $r = 4$) and $C_{j'}$ is an ordinary column but $j' \neq j$. So, $m_2 = \frac{2c}{3} - 1$.

Case (iii) Let $r = 5$

From Fig. 4 we have a Type B node $N_{i,j}$ is at a distance of two, with the nodes given by $N_{i',j'}$ for $i' = i + 2, i - 2$ and $C_{j'}$ is an ordinary column but $j' \neq j$. So, $m_2 = \frac{2c}{3} - 1$. So, $m_2 = 2\left(\frac{2c}{3} - 1\right)$.

Case (iv) Let $r > 5$

From Fig. 5 it follows that a Type B node $N_{i,j}$ is at a distance of two, with the following nodes

1. $N_{i',j'}$ for $i' = \{1, 2, \cdots, r\} \setminus \{i-1, i, i+1\}$ and $C_{j'}$ is an ordinary column but $j' \neq j$, i.e., $(r - 3)(\frac{2c}{3} - 1)$ nodes
2. $N_{i',j}$ where $i' = \{1, 2, \cdots, r\} \setminus \{i - 2, i - 1, i, i + 1, i + 2\}$, i.e., $(r - 5)$ nodes

Thus, we have $m_2 = (r - 3)(\frac{2c}{3} - 1) + (r - 5) = 2(\frac{rc}{3} - c - 1)$.

Substituting the obtained values in each case for m_2 in equation (2) we get the desired expression as given in the statement of Theorem 4.4. □

4.1 KeyPath Establishment

We discuss the key path establishment phase between two randomly chosen nodes, $P : N_{i_1,j_1}$ and $Q : N_{i_2,j_2}$, from the network. The nodes first broadcast their node identifiers : node P broadcasts i_1, j_1 and node Q broadcasts i_2, j_2. Algorithm 4.5 discusses how to find the intermediate node or intermediate key-path between P and Q, if exists.

The expressions (corresponding values for fixed r), obtained in Theorems 4.2, 4.3 and 4.4 adds up to give $\frac{1}{2}rc(rc - 1)$, which is the total number of possible links in the network. This alternatively verifies the validity of Theorem 4.1. It follows from Theorem 4.1 that all the nodes in the network are connected by at least a path of distance at most three. Note that, the one-hop, 2-hop and 3-hop paths between any two nodes are not unique. Moreover, paths of length more than three, also exist on the network. Therefore, when few nodes becomes inactive, and the shortest path between any two communicating nodes do not exist, they look for the alternative paths of same or larger length.

```
Algorithm 4.5
Input: The node identifiers $P : N_{i_1,j_1}$ and $Q : N_{i_2,j_2}$
procedure FindKeyPath
    if ($j_1 \equiv 2 \mod 3$) then
        if (($i_1 = i_2$) || ($j_1 = j_2$))
            print "$P$ and $Q$ are directly connected";
        else if ($j_2 \equiv 2 \mod 3$) then
            $N_{i_1,j_2}$ or $N_{i_2,j_1}$ is an intermediate node;
            else if $N_{i_2,j_1}$ is an intermediate node;
            end if
        end if
    else if ($j_1 \equiv 0 \mod 3$) then
            if ($j_2 \equiv 2 \mod 3$) then
                $N_{i_1,j_2}$ is an intermediate node;
            else
                The key-path is $N_{i_1,j_1} \rightarrow N_{i_1,j_2-1} \rightarrow N_{i_2,j_2-1} \rightarrow N_{i_2,j_2}$;
            end if
    else if ($j_1 \equiv 1 \mod 3$) then
            if ($j_2 \equiv 2 \mod 3$) then
                $N_{i_2,j_1}$ is an intermediate node;
            else
                The key-path is $N_{i_1,j_1} \rightarrow N_{i_1,j_2+1} \rightarrow N_{i_2,j_2+1} \rightarrow N_{i_2,j_2}$;
            end if
    end if
end FindKeyPath
```

5 Overall Performance

In this section we evaluate the efficiency of our scheme on the basis of connectivity, resilience and memory. For convenient comparison, we consider a network consisting of nearly 2000 nodes. We further consider fifteen sets of combinations of the number of rows and columns which lead to network of size almost 2000. We carry out all the computations and comparisons with these fifteen sets of values.

5.1 Memory

It has already been mentioned that storage in each node is limited. Although authors claim that storing even 150 keys per node is permitted [7], it is always better to keep storage (i.e., the average number of keys to be stored per node) as small as possible.

In this section we discuss the memory requirement of the proposed network. Note that the number of keys to be stored by a node depends on the Type of the node. Let us assume that k_A and k_B denote the number of keys to be stored

Table 1. Memory

r	c	N	p	q	k_A	k_B	k
7	285	1995	2	17	21	20	≤ 21
13	156	2028	3	13	18	16	≤ 18
21	96	2016	4	11	17	14	16
29	63	1827	5	8	15	11	≤ 14
31	66	2046	5	8	15	11	≤ 14
37	54	1998	7	7	16	10	14
57	36	2052	7	7	16	10	14
73	27	1971	8	5	15	8	≤ 13
91	21	1911	9	4	15	7	≤ 14
133	15	1995	11	4	17	7	≤ 14
167	12	2004	13	3	18	6	14
222	9	1998	16	3	21	6	16
333	6	1998	19	2	23	5	17
666	3	1998	27	2	31	5	≤ 23

by a Type A and Type B nodes respectively. Let us suppose that k represents the average memory of any randomly chosen node from the network.

In Table 1 we show the memory requirements of a network composed of more or less 2000 nodes, with different choice of the number of rows and columns. From the table it is evident that the memory requirement is very small in our network.

5.2 Connectivity

We assume that the nodes on each row and special columns are deployed together so that the nodes sharing common key also lie within radio frequency range. However, in this section we investigate the connectivity based on the key distribution of the network. For fixed size of the network (i.e., $N \approx 2000$), we provide the percentage of one-hop, two-hop and three-hop paths and the average path-length between any two nodes in the network. The target of our scheme is to minimize the average path-length between any two nodes. Theorems 4.2, 4.3 and 4.4 give the number of single-hop, 2-hop and 3-hop paths in the network. Therefore, the average path-length between two nodes in the network is $d = (L_1 + 2L_2 + 3L_3)/(L_1 + L_2 + L_3)$.

The percentage of one-hop two-hop and 3-hop paths denoted by l_1, l_2 and l_3 respectively, and the average path length d between any two randomly chosen nodes, corresponding to each of the fifteen sets are listed in Table 2.

From Table 2 we note that the average path length has its value in the range (2.10 to 2.40). the network is best connected when the number of rows are very small compared to the number of columns. Note that, the average path-length corresponds to the key-connectivity only.

Table 2. Connectivity

r	c	N	l_1	l_2	l_3	d
7	285	1995	14.41	60.25	25.34	2.109328
13	156	2028	7.91	57.95	34.14	2.262292
21	96	2016	5.11	56.84	38.04	2.329363
29	63	1827	3.98	56.22	39.80	2.358160
31	66	2046	3.73	56.17	40.10	2.363651
37	54	1998	3.32	55.88	40.79	2.374729
57	36	2052	2.68	55.26	42.06	2.393792
73	27	1971	2.61	54.82	42.57	2.399662
91	21	1911	2.69	54.38	42.93	2.402443
133	15	1995	2.98	53.63	43.40	2.404213
167	12	2004	3.38	53.02	43.60	2.402230
222	9	1998	4.16	52.04	43.80	2.396428
333	6	1998	5.86	50.14	44	2.381405
666	3	1998	11.27	44.53	44.20	2.329327

5.3 Resilience

Resilience is one of the most important evaluation parameters, to quantify the efficiency of a network. Resilience measures the robustness of a network under adversarial attack.

We consider the *random node capture* as the attack model. We assume that the adversary can listen and eavesdrop any communication over the channel between two nodes, but cannot tamper it. Under random node capture attack, the adversary also captures a large number of nodes and extracts all the keys stored at them. Now, the remaining nodes cannot further use those keys for communication. We address the measure of resilience in two ways: on the nodes and on the links (direct key-path between two nodes), in the following subsections node disconnection and link failure.

Node Disconnection. A node is said to be disconnected from the network if all the keys stored at the node are known to the adversary. In this section, we find the effect of adversary on the rest of the nodes. We quantify node disconnection by $V(s)$, the fraction of total number of nodes, that becomes disconnected when s nodes are compromised from the network. We obtain an expression for $V(s)$ with the help of the following results.

Theorem 5.1. *Minimum number of nodes to be compromised to disconnect a node $N_{i,j}$ completely from the network is given by*

$$\begin{cases} p+q+2, & \text{if } N_{i,j} \text{ is a Type A node;} \\ p+3, & \text{if } N_{i,j} \text{ is a Type B node.} \end{cases}$$

Proof. A node $N_{i,j}$ gets disconnected from the network if all the connections from $N_{i,j}$ are destroyed, i.e., all the nodes having a common key with $N_{i,j}$ get captured. We consider the following two cases:

Case (i): Let $N_{i,j}$ be a Type A node.

Now, it can be noted that, all the keys, stored at the node $N_{i,j}$, distributed to the nodes in the same column and same row of $N_{i,j}$ should be captured in order to disconnect $N_{i,j}$. There are $(c-1)$ more nodes in the row. From the property of the projective plane of order p, which has $p^2+p+1 \geq c$ nodes, it follows that in order to disconnect one node, at least $p+1$ nodes should be destroyed. According to the assumption C_j is a special column. There are $(r-1)$ more nodes in the column C_j, arranged according to a projective plane of order q, i.e., $q^2+q+1 \geq r$. Similarly at least $q+1$ nodes need to be captured. Therefore, one Type A node $N_{i,j}$ will be disconnected provided $q+1$ nodes along this column C_j and $p+1$ nodes along the row R_i get captured.

Case (ii): Let $N_{i,j}$ be a Type B node.

It can be observed observe that to disconnect $N_{i,j}$, all the nodes along the same row of $N_{i,j}$ and the two adjacent nodes of $N_{i,j}$ along the same column as $N_{i,j}$ should be destroyed. Therefore, total number of nodes to be captured to disconnect $N_{i,j}$ completely is $(p+1)+2 = p+3$ nodes.

This completes the proof of the theorem. □

Corollary 5.2. Average number of nodes disconnected when s nodes are captured is given by $v_1(s) = \frac{3s}{(3p+q+8)}$.

We skip the proof due to page restrictions.

The measure of node disconnection, defined as the fraction of nodes that become disconnected when s nodes are compromised is given by

$$V(s) = \frac{v_1(s)}{N-s} = \frac{3s}{(3p+q+8)(N-s)}.$$

In Table 3, we provide the values $V(s)$ for increasing values of s, the number of compromised nodes. The total number of nodes in the network are assumed to be almost 2000. From the table it follows that we obtain better node disconnections when there are large number of rows and very small number of columns.

Table 3. Node disconnection of the proposed network

r	c	N	$V(100)$	$V(150)$	$V(200)$	$V(250)$	$V(300)$
7	285	1995	0.005107	0.007868	0.010783	0.013864	0.017128
13	156	2028	0.005187	0.007987	0.010941	0.14061	0.017361
21	96	2016	0.005051	0.007779	0.010658	0.013700	0.016919
31	66	2046	0.004973	0.007656	0.010485	0.013471	0.016628
29	63	1827	0.005604	0.008656	0.011896	0.015342	0.019013
37	54	1998	0.004391	0.006764	0.009270	0.011918	0.014723
57	36	2052	0.004269	0.006572	0.008999	0.011561	0.014269
73	27	1971	0.004334	0.006679	0.009157	0.011778	0.014557
91	21	1911	0.004248	0.006552	0.008992	0.011578	0.014325
133	15	1995	0.003518	0.005420	0.007428	0.009551	0.011799
167	12	2004	0.003151	0.004854	0.006652	0.008552	0.010563
222	9	1998	0.002679	0.004127	0.005656	0.007272	0.008984
333	6	1998	0.002359	0.003634	0.004981	0.006404	0.007911
666	3	1998	0.001737	0.002676	0.003667	0.004715	0.005825

Link Failure. A link is said to exist between two nodes if they share a common key. From the construction it follows that there is at most one common key between any two nodes, i.e., in this case, each link corresponds to a key. If the common key between any two key-connected nodes is captured by the adversary, one link is said to be destroyed.

The anti-resilience of a scheme is given by [9]:,

$$fail(s) = 1 - \left(1 - \frac{r^{'} - 2}{N - 2}\right)^{s} \tag{3}$$

where $fail(s)$ denotes the probability that a link between two uncaptured nodes is broken when s nodes are compromised in a network of size N and each key is assigned to $r^{'}$ number of nodes. Our grid-based scheme is neither regular nor uniform, i.e., the number of nodes to which each key is assigned varies in our case. Hence we modify eqn. (3) as follows:

$$fail(s) = 1 - \left(1 - \frac{m - 2}{N - 2}\right)^{s} \tag{4}$$

where m is the average number of nodes to which each key is assigned. We now find an expression for m.

Theorem 5.3. The average number of nodes to which each key is assigned, is

$$m = \frac{p + 3q + 8}{6}.$$

We skip the proof due to page restrictions.

Table 4. Link failure of our scheme

r	c	N	$fail(10)$	$fail(20)$	$fail(30)$	$fail(50)$	$fail(100)$	$fail(200)$	$fail(500)$	$fail(1000)$
7	285	1995	0.015776	0.031303	0.046584	0.076429	0.147016	0.272418	0.548451	0.796104
13	156	2028	0.014709	0.029202	0.043482	0.071414	0.137729	0.256489	0.523329	0.772784
21	96	2016	0.015612	0.030981	0.046110	0.075662	0.145599	0.269999	0.544689	0.792692
29	63	1827	0.017217	0.034137	0.050766	0.083170	0.159422	0.293429	0.580348	0.823892
31	66	2046	0.015385	0.030533	0.045448	0.074594	0.143623	0.266618	0.539399	0.787847
37	54	1998	0.019861	0.039327	0.058406	0.095436	0.181763	0.330489	0.633230	0.865479
57	36	2052	0.019342	0.038309	0.056910	0.093039	0.177422	0.323366	0.623396	0.858170
73	27	1971	0.020961	0.041483	0.061574	0.100503	0.190905	0.345365	0.653264	0.879774
91	21	1911	0.023324	0.046104	0.068353	0.111305	0.210221	0.376250	0.692725	0.905582
133	15	1995	0.027256	0.053769	0.079560	0.129052	0.241449	0.424600	0.748855	0.936926
167	12	2004	0.031189	0.061404	0.090678	0.146514	0.271562	0.469378	0.794902	0.957935
222	9	1998	0.038559	0.075631	0.111274	0.178489	0.325120	0.544538	0.859999	0.980400
333	6	1998	0.044988	0.087952	0.128983	0.205590	0.368913	0.601729	0.899897	0.989979
666	3	1998	0.064041	0.123981	0.180082	0.281736	0.484097	0.733844	0.963454	0.998664

In Table 4, we provide the values $fail(s)$ for increasing values of s, the number of compromised nodes. Table 4 indicates that we obtain better resilience, i.e., smaller values of $fail(s)$ when there are small number of rows and large number of columns.

5.4 Comparison with Existing Schemes

In Figure 6, we provide the comparison of link failure of our scheme with some existing schemes. To keep up N in our scheme comparable with other schemes, we consider a network with 13 rows and 156 columns i.e., $p = 3$ and $q = 13$, where the total number of nodes in the network being 2028.

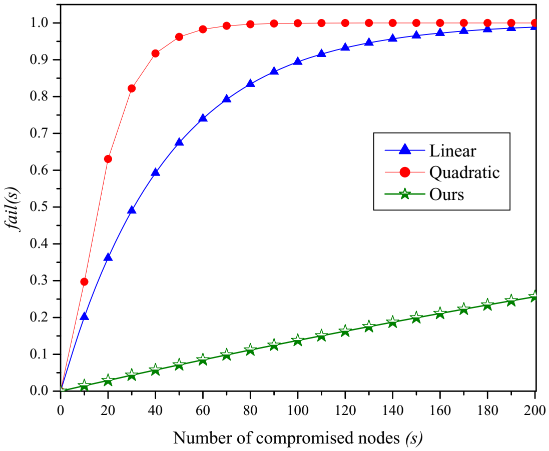


Fig. 6. Comparison of link failure

6 Conclusion

We present a KPS for a non-uniform rectangular grid adapting a deterministic approach. The nodes are assumed to be of two types depending on the resources provided to them. It is seen that the network is well connected, any two nodes can communicate (either directly or via a key-path) whenever they are within radio frequency range. The existence of multiple key-paths (of different lengths) between any two nodes increase the smooth relay of the data throughout the network even under adversarial attack. The results indicate that a large network is supported with a small memory requirement. The obtained results show that the scheme is well-resilient when compared to existing schemes.

References

1. Bag, S., Ruj, S.: Key Distributions in Wireless Sensor Networks Using Finite Affine Plane. In: Workshop of International Conference on Advanced Information Networking and Applications, pp. 436–441. IEEE Computer Scociety (2011)
2. Blackburn, S.R., Etzion, T., Martin, K.M., Paterson, M.B.: Efficient Key Predistribution for Grid-Based Wireless Sensor Networks. In: Safavi-Naini, R. (ed.) ICITS 2008. LNCS, vol. 5155, pp. 54–69. Springer, Heidelberg (2008)
3. Camptepe, S.A., Yener, B.: Combinatorial Design of Key Distribution Mechanisms for Wireless Sensor Networks. ACM Trans. Netw. 15(2), 346–358 (2007)
4. Chakrabarti, D., Maitra, S., Roy, B.: A Key Pre-distribution Scheme for Wireless Sensor Networks: Merging Blocks in Combinatorial Design. In: Zhou, J., López, J., Deng, R.H., Bao, F. (eds.) ISC 2005. LNCS, vol. 3650, pp. 89–103. Springer, Heidelberg (2005)
5. Chen, C.Y., Chao, H.C.: A survey of Key Predistribution in Wireless Sensor Networks. Security Comm. Networks (2011)
6. Eschenauer, L., Gligor, V.D.: A Key-management Scheme for Distributed Sensor Networks. In: ACM CCS, pp. 41–47. ACM (2002)
7. Lee, J., Stinson, D.R.: Deterministic Key Predistribution Schemes for Distributed Sensor Networks. In: Handschuh, H., Hasan, M.A. (eds.) SAC 2004. LNCS, vol. 3357, pp. 294–307. Springer, Heidelberg (2004)
8. Lee, J., Stinson, D.R.: Common Intersection Designs. Journal of Combinatorial Designs 14, 251–269 (2005)
9. Lee, J., Stinson, D.R.: A Combinatorial Approach to Key Predistribution for Distributed Sensor Networks. In: IEEE WCNC, pp. 1200–1205 (2005)
10. Lee, J., Stinson, D.R.: On The Construction of Practical Key Predistribution Schemes for Distributed Sensor Networks Using Combinatorial Designs. ACM Trans. Inf. Syst. Secur. 11(2) (2008)
11. Paterson, M.B., Stinson, D.R.: A Unified Approach to Combinatorial Key Predistribution Schemes for Sensor Networks. IACR Cryptology ePrint Archive (2011)
12. Ruj, S., Roy, B.: Key Predistribution Using Partially Balanced Designs in Wireless Sensor Networks. In: Stojmenovic, I., Thulasiram, R.K., Yang, L.T., Jia, W., Guo, M., de Mello, R.F. (eds.) ISPA 2007. LNCS, vol. 4742, pp. 431–445. Springer, Heidelberg (2007)
13. Ruj, S., Roy, B.: Key Predistribution Schemes Using Codes in Wireless Sensor Networks. In: Yung, M., Liu, P., Lin, D. (eds.) Inscrypt 2008. LNCS, vol. 5487, pp. 275–288. Springer, Heidelberg (2009)
14. Stinson, D.R.: Combinatorial Designs: Constructions and Analysis. Springer, New York (2003)

FLAME: A Flexible and Low-Power Architecture for Wireless Mesh Networks

Seyed Dawood Sajjadi Torshizi[1], Sadra Mohammadalian[2], Fazirulhisyam Hashim[1], and Subramaniam Shamala[2]

[1] Department of Computer and Communication System Engineering, Faculty of Engineering,
Universiti Putra Malaysia (UPM), Serdang, Selangor, Malaysia
dawood.sajjadi@ieee.org, fazirul@eng.upm.edu.my

[2] Department of Communication Technology and Networks, Faculty of Computer Science
and Information Technology, Universiti Putra Malaysia (UPM)
sadra.m.alian@gmail.com, shamala@fsktm.upm.edu.my

Abstract. Nowadays, Wireless Mesh Network (WMN) is known as a promising technology for fast, robust and low-cost deployment of network infrastructures. Establishment of high throughput and reliable links, deduction of the interference effect and saving the energy are some of the serious concerns in construction of WMNs. Although, dominant solutions in organizing these networks are based on omnidirectional antennas, utilizing the directional antennas due to their potential advantages in terms of attaining higher throughput and coverage is a real interest in WMNs. In this paper, a novel framework for building well-organized and low power consumption WMNs by means of flexible directional antennas is presented which to the best of our knowledge can be considered as an innovative solution in deployment of optimized and green WMNs. Acquired preliminary results substantiate not only the efficiency of offered framework in flexible implementation of mesh networks, but also in construction of high-throughput and resilient WMNs.

Keywords: WMN, directional antenna, power conservation, NS3 simulator, QoS, battery lifetime.

1 Introduction

Recently, Wireless Mesh Networks (WMN) is considered as a propitious solution for rapid and cost-effective deployment of various network services and applications. The multi-hop wireless links, which are constructed by means of the familiar 802.11a/b/g/n standards, constitute the primary elements of this paradigm. Particularly, this kind of wireless networks is similar to the wired network in the sense that it is distinguished by a static topology with infrequent changes.

Over recent years, using directional antennas in the deployment of wireless mesh networks has attracted a lot of interests. Directional antennas in comparison to the omnidirectional antennas are able to cover more larger transmission range and also use more spatial reuse to achieve higher throughput. Through this commodity, the

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 171–184, 2013.

wireless nodes can communicate to each other with less interference and get more radio coverage, simultaneously. Hence, by establishment of long distance links, fewer numbers of routing hops will be used which, it leads to the elevation of QoS for determined services.

In this research, we attempted to take advantage of directional antennas by proposing a dynamic sectorial coverage in the deployment of a WMN infrastructure. In fact, we propose a new architecture based on directional antennas, which provides self-healing and on demand connectivity among all utilized mesh routers to improve the overall QoS and reduce the amount of consumed power within each mesh router. Simulation as one of the most dominant methodologies was chosen to direct our experiments over the proposed solution. In our simulation, we regarded only the topology of stationary nodes, which decreases the complexity level of assessment procedure in the early experiments. Demonstrated simulations results under NS3 confirm that applying the offered framework to dynamic directional antenna and flexible transmission power improves the functionality of mesh networks in several aspects and outperforms omnidirectional antennas in terms of energy consumption and coverage areas. The main contribution of this work can be regarded as the provisioning of a green architecture for resilient deployment of wireless mesh networks with respect to attain the maximum achievable throughput between every determined points. Since one of the main global concerns of recent research topics is the practice of selecting energy-efficient networking technologies and minimizing resource usages whenever possible, we believe that the offered framework provides a proper solution for better construction of mesh networks.

The remainder of this paper proceeds as follows. In the next section, well-known related works about the utilization of directional antennas in construction of wireless mesh networks are discussed. Then in section 3 the proposed framework and its major components are described in detail. Towards evaluating the fundamental functionalities of offered solution, multiple simulation scenarios are conducted in the succeeding section. In section 5, acquired results and relevant discussions for performed experiments are deliberated. Finally, concluding statements and future plans are presented in section 6.

2 Related Works

So far, several research works have been conducted on employment of directional antennas to promote regular performance metrics in wireless mesh networks. As an instance, DMesh [1] is one of the first architectures, which has integrated the spatial separation of directional antennas with frequency segregation of non-overlapping channels to enhance the throughput of WMNs. The authors (through directing different scenarios over simulation and test-bed environments) have represented the promotion of attained throughput by means of the offered solution to more than 200% compared to the mesh networks, which are based on omnidirectional antennas. Also, S. Muthaiah et al. in [2] presented a case study on using smart (directional) antennas to improve spatial reuse and coverage range in addition to the achieving higher throughput at lower powers.

Another related research is directed in [3] through simulation, which in a multi-radio multi-channel scenario, authors utilized directional antennas in each mesh router to degrade the amount of interference and provide simultaneous transmission capability to promote physical data rate and consequently aggregate throughput in the mesh network. In the performed simulation, it was supposed that the transmission power (TX power) in each node is tuned appropriately to decrease the interference effect just to the adjacent nodes. J. Zhang directed related researches in [4] and [5] to analyze the capacity of mesh networks with Omni and directional antennas. Authors proposed a mathematical model in [4] to show the capacity enhancement of WMN by means of directional antennas.

At the other work [5], an algorithm was presented to increase the traffic delivery ratio of mesh nodes by adjusting the orientations of directional antennas in the mesh network. It is assumed that each node has multiple antennas, which by interconnecting various nodes to each other, the delivery ratio can be improved up to 280%. It should be noted that after running each round of the offered algorithm, the antenna orientation and mesh topology will be rearranged just to ameliorate the end-to-end throughput within the simulation, but the authors didn't present any practical mechanism for adjustment procedure in the real scenarios. Furthermore, in [6], [7] and [8] the scholars through modeling and simulation represented the efficaciousness of MIMO and directional antennas in different aspects of WMN operations.

H. Okada et al. also in [9] and [10] proposed a 3-sectore antenna system for utilization in multi-radio multi-channel WMN. Through extensive analytical modeling and test-bed implementations, they proved the enhancement of several metrics especially throughput by means of sector antennas, but they didn't investigate the efficiency of their offered solution in term of power consumption. J. Ben-Othman et al. directed analogous works in [11] and [12] by means of NS3 simulator for using sector antennas to enhance overall QoS. Customized versions of OLSR [13] and IEEE 802.11s [14] routing protocols were employed in several scenarios while utilization of directional antennas were taken into account.

In addition to the aforementioned researches, multiple works performed over channel assignment and topology control in wireless mesh networks which directional antennas were regarded as the key design component [15-18]. In [17], a 3-step topology control mechanism for adjusting the orientation of directional antennas and channel allocation is proposed. The offered solution facilitates the construction of mesh networks and maximizes the delivery ratios of traffic demands. It begins by organizing a set of routing trees to balance the traffic among the tree links. Then, the interface allocation for each node of tree with regards to the load distribution among served links by that node will be done. At last, it performs the antennas orientation and channel assignment to mitigate the interference while covering all specified destination nodes. Several relevant researches were inspired from this scheme, which tried to decrease its complexity and improves its topology formation in terms of various performance metrics such as end-to-end delay and Packet Delivery Ratio (PDR) among mesh nodes [18].

Almost in all mentioned works on using directional antennas in WMNs, the major concern of scholars was promotion of throughput and PDR among the mesh routers and energy conservation through dynamic assignment of transmission power and antenna orientation has not been fully investigated yet. Hence, in the next section, the main focus of the proposed architecture is achieving to the higher level of power conservation and system efficiency through dynamic allocation of TX power and antenna orientation in each radio. Different components of offered solution and their respective functions are deliberated in details in section 3.

3 Proposed Architecture

The key challenge in deployment of wireless mesh networks is how to construct a network that connects all nodes to the mesh gateways, such that the end-to-end throughput of the network is maximized. The antennas orientation alignment and channels assignment procedures can be regarded as the major steps of this process, which are dependent to each other. Another serious concern in implementation of mesh networks, especially those are deployed in none-urban and low population areas, is power conservation and utilizing renewable energies to feed different network elements.

Proposed mechanism in this paper is applicable for mesh networks with stationary nodes, which are using single or multiple radios to establish wireless links in different channels. The essence of the offered solution in power saving is based on using the receiver sensitivity values of utilized wireless adapters and finding the optimum required transmission power on each radio. To attain this goal, the Received Signal Strength Indication (RSSI) values of peer nodes will be monitored and compared with predefined thresholds quantities, continuously.

Table 1. Typical QoS parameters for well-known applications [19]

Type	Application	Data Rate	Latency	Jitter	Loss
Audio	Conversation	4,64 kbps	< 150 ms	< 1ms	< 3%
	Speech/Music	5,128 kbps	< 10 s	< 1ms	< 1%
Video	Real-time video	16,384 kbps	< 10 s	< 1ms	< 1%
Data	Interactive game	< 8 kbps	< 250 ms	n/a	0 %
	Web browsing	< 80 kbps	< 250 ms	n/a	0 %
	Email	< 80 kbps	< 4 s	n/a	0 %

In fact, based on the offered algorithm, there is an initial network topology, which all nodes are placed and configured to setup a basic mesh topology at the beginning of the experiment. Then, all mesh routers start tuning procedure of TX power values in their radios while they should maintain the quality of established wireless links with regard to the predetermined QoS thresholds. Considering the type and sensitivity of utilized applications over mesh networks, appropriate margins can be defined to verify the acceptable level of QoS for wireless links. TABLE I demonstrates some of the most familiar network applications and their corresponding required QoS parameters to present a reliable service to the end-users.

Offered solution is divided into interconnected steps among which construction of a multi-radio multi-channel mesh network topology is the first step. In the next step, reduction of the transmission power to conserve much more energy will be begun. This process will lead to the mitigation of the interference effect on the link quality and attaining to the maximum link throughput in each node. Fig. 1 represents the essential components of offered architecture, with the Antenna Alignment Engine (AAE) as the key section, which integrates the rest of the system's components.

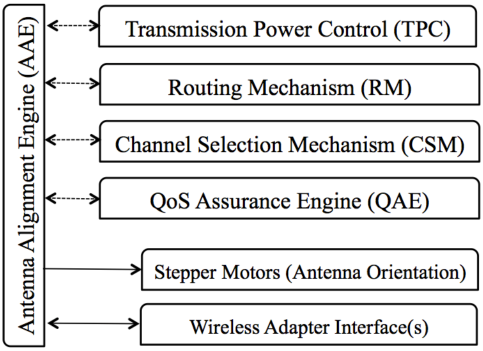

Antenna Alignment Engine (AAE)
Transmission Power Control (TPC)
Routing Mechanism (RM)
Channel Selection Mechanism (CSM)
QoS Assurance Engine (QAE)
Stepper Motors (Antenna Orientation)
Wireless Adapter Interface(s)

Fig. 1. Essential components of proposed architecture

Since antenna alignment procedure has a direct impact on the functionality of other system components, they can be controlled and optimized through this entity. RM and CSM respectively as the routing and channel selection modules in the aforementioned framework accommodate basic addressing and connectivity among mesh routers which regarding the number of nodes and network topology can be chosen as static or dynamic.

It should be noted that although multiple energy efficient and battery aware routing schemes have been presented for wireless mesh networks so far [20-23], almost all of them applied only over Omni directional antennas and none of them proposed a practical approach for deployment of their offered mechanisms. According to the predetermined QoS margins, QAE module is in charge of quality guarantee for established links in each mesh routers. Mentioned values in TABLE I can be considered as a reliable reference for operation of this component. By quality degradation of any wireless links, the transmission power of corresponding radio should be increased which this task will be done through TPC component after notification of QAE module. Then, the determined level of TX power based on proposed algorithm will be committed over each radio adapter by means of AAE module The last major component of the proposed framework, which distinguishes it from other similar works, is the Antenna Alignment Engine (AAE) module. For better explanation of this element, a multi-radio multi-channel mesh topology is displayed in Fig. 2. As it is shown in Fig. 2 (a), 8 nodes, which each one is equipped with 2 radios, are interconnected to each other through wireless links that operate in different 2.4 GHz frequency channels. Suppose that as it is presented in Fig 2 (b), the established link between nodes F and G which functions in channel 6 is broken due to any unexpected reason, such as the existence of high interference in the utilized channel or elimination of line of sight.

Since in the illustrated topology the main objective of using dual radio in each mesh router is attaining higher link throughput, the broken radio in each node attempts to find an alternate path to exploit the available link capacity as much as possible. By means of the latest topology information in each node which is acquired through mesh nodes' broadcast messages, the new redundant paths will be constructed through node pairs (F,C) and (G,H) as it is displayed in Fig 2 (c).

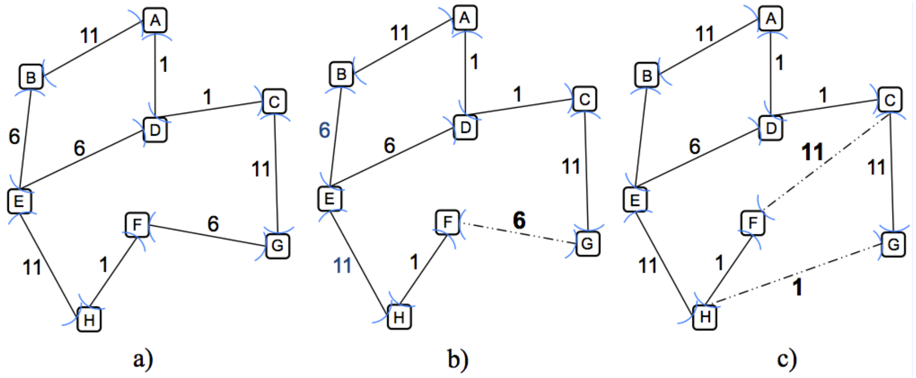


Fig. 2. Functionality of flexible antennas in a sample WMN topology

After determination of the best neighbor, each mesh router requires to change its antenna alignment properly in vertical and horizontal directions to establish a resilient wireless link. Fig. 3 illustrates a practical structure for implementation of the proposed solution within the wireless mesh nodes. Toward deploying a multi-radio multi-channel mesh network, it is recommended each node be equipped with 2 wireless interfaces that each adapter could be connected to a separate directional antenna to radiate in a specific direction. To justify the antenna orientations in appropriate directions, 2 stepper motors are utilized which are being controlled by the implemented algorithm within the mesh router.

In Fig. 3, side and top views of a single mesh node with 2 directional antennas connected to it, are shown. As it is demonstrated, by means of stepper motors, each antenna is able to rotate in clockwise or counterclockwise directions up to 180 degree to find the best succeeding node for passing the network traffic to it. This capability, in addition to the formation of high throughput links and optimizing the amount of consumed energy, will results in the promotion of mesh network stability in the occurrence of any link failure among the mesh nodes. In fact, in this case, the rotating antennas are able to discover another proper neighbor node to create a reliable connection with it based on the predetermined QoS margins by means of QAE module.

One of the most challenging steps of the offered solution is tuning the orientation of directional antennas to cover enough nodes, such that the network topology is preserved; on the other hand, it is necessary to set the antenna to the orientation that the network interference is minimized. This issue is further complicated by the channel assignment, because only the radios in each other's interference range and using the same channel interfere with each other. Through the proposed algorithm, it is attempted to address the whole aforementioned concerns properly to achieve the main goals of the presented architecture.

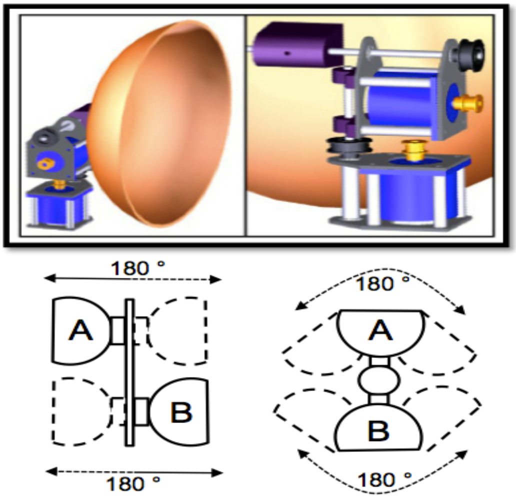


Fig. 3. Flexible antennas with dynamic orientations

Fig. 4 represents the offered algorithm for the operation of each node to attain higher level of energy conservation and throughput. As it is shown, the process continuously monitors the functionality and quality of established links in specific intervals. When initialized timer reaches zero, the link quality will be checked and if

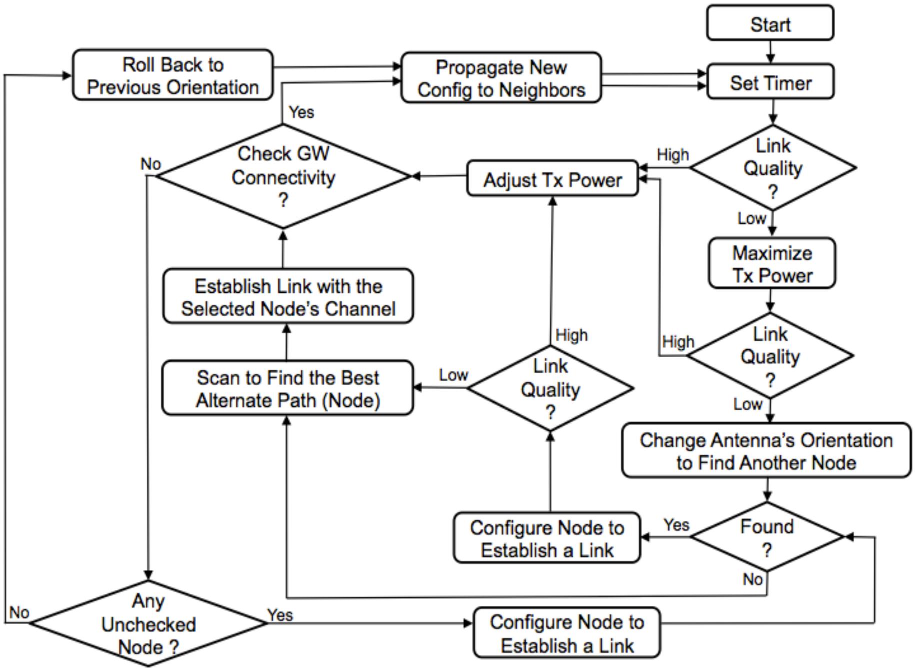

Start
Set Timer
Link Quality ?
High
Low
Maximize Tx Power
Link Quality ?
High
Low
Change Antenna's Orientation to Find Another Node
Found ?
Yes
No
Configure Node to Establish a Link
Link Quality ?
High
Low
Adjust Tx Power
Check GW Connectivity ?
Yes
No
Propagate New Config to Neighbors
Roll Back to Previous Orientation
Scan to Find the Best Alternate Path (Node)
Establish Link with the Selected Node's Channel
Any Unchecked Node ?
No
Yes
Configure Node to Establish a Link

Fig. 4. Proposed algorithm to adjust antenna orientation and TX power

it is less than predetermined thresholds, TX power of the corresponding radio will be maximized. Then, link quality will be rechecked again and if it is elevated, the TX power based on RSSI and receiver sensitivity values of respective adapter will be adjusted to the optimum quantity.

Through checking the gateway connectivity in the next stage will be identified that the applied changes haven't had any undesired effect on the functionality of mesh network. According to this fact that each node broadcasts its latest links' status within the mesh network domain, all nodes have a map of current network topology. If the relevant mesh node couldn't find any peer node to establish a proper wireless connection with it to achieve higher level of QoS, it resumes its previous orientation and waits to make another try in the next time period. In the case that link quality after amplification of TX power to the highest value still is lower than the specified margins, scanning procedure to find another choice to establish more resilient link will be taken place. In the last step of each round in the presented flowchart, the latest node's information will be propagated through the network to inform the others about the recent situation. In the succeeding sections, 2 separate scenarios for evaluating the primary functionalities of offered architecture are expressed and acquired outcomes are delineated in detail.

4 Simulation Setup

To investigate and validate the fundamental operation of offered framework, two elementary scenarios were conducted on multi-radio multi-channel mesh topologies. The main purpose of running these experiments is the correlation of battery lifetime and achievable throughput for various TX powers in each mesh router. In both scenarios, fixed UDP offered load from source node A, is transmitted to node B through directional antennas of intermediate mesh routers that radiate in 5 GHz frequency band.

Table 2. Simulation setup parameters

Simulation Time	300 seconds
Simulation Iteration	10 times
Packet Size	1024 Bytes
CBR for UDP offered load	5.5 Mbps
Utilized Frequency Channels	36, 44, 48, 56
Initial Charge of Battery	1900 J, 6100 J
Node distance	900 m
Channel Bandwidth	20 MHz
Max Transmission Power	25 dBm
Transmission Rate	24 Mbps
Wireless Standard (PHY/MAC)	802.11a
Routing Mechanism	Open11s (802.11s)
Node Distribution	Static
Antenna gain/beamwidth	11 dBi / 120°
Path Loss Model	Friis Free Space

The second experiment is performed in 2 phases; in the first one, such as the first scenario the achievable throughput and battery lifetime are measured for nodes A and B while they are connected to each other by 3 hops. In the next one, it is assumed that the default wireless link that interconnects the node B to the mesh backbone is broken and this node finds an alternate path through its other neighbor node to reach node A.

Respective utilized configurations for the conducted simulation scenarios are mentioned in TABLE II in detail.

5 Results and Discussion

To ground our discussion, we present the obtained results for the first scenario, which its relevant topology is illustrated in Fig. 5. As it is shown in demonstrated graphs in Fig. 7, the remaining battery energy and throughput metrics are quantified in duration of 300 seconds for two different transmission power values. In the second diagram, measured throughput at node B and remained battery energy for both mesh nodes were displayed while the TX power is equal to 25 dBm (~316 mW). By beginning of the simulation, the battery energy of each node starts reducing from its initial value in a linear tension. In the middle of simulation period i.e. 150s, node's B battery is finished and consequently the link throughput value degraded to zero. On the contrary, when

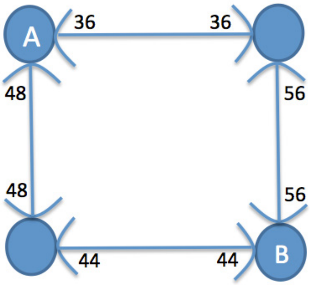


Fig. 5. Multi-radio Multi-channel mesh topology for the first scenario

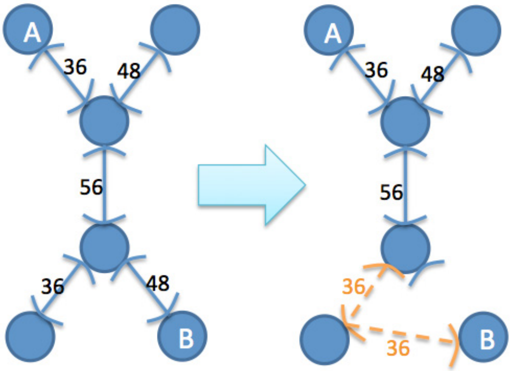


Fig. 6. Topology of second scenario in 2 phases

the TX power is deducted to 20 dBm (~100 mW) as it is presented first graph in Fig. 7, not only at the end of simulation, both nodes still are alive, but also high throughput values between 5.46 and 5.5 Mbps regarding the UDP offered load are attained. In spite of long determined distance (900 m) between paired links in simulation setup, the achievable throughput is almost as equal as the UDP offered load. It is important to note that since the selected transmission rate is assumed 24 Mbps and by using the non-overlapping channels to alleviate the interference impact, the attained throughput in the conducted experiment should be close to 5.5 Mbps offered load value. Acquired throughput quantities, which are presented in Fig. 7, are consistent with this fact.

For the first scenario, in addition to the aforementioned TX powers, similar experiment was directed when the TX power was set to 15 dBm, but the average RSSI in the neighbor nodes respecting the long predetermined distance in the simulation

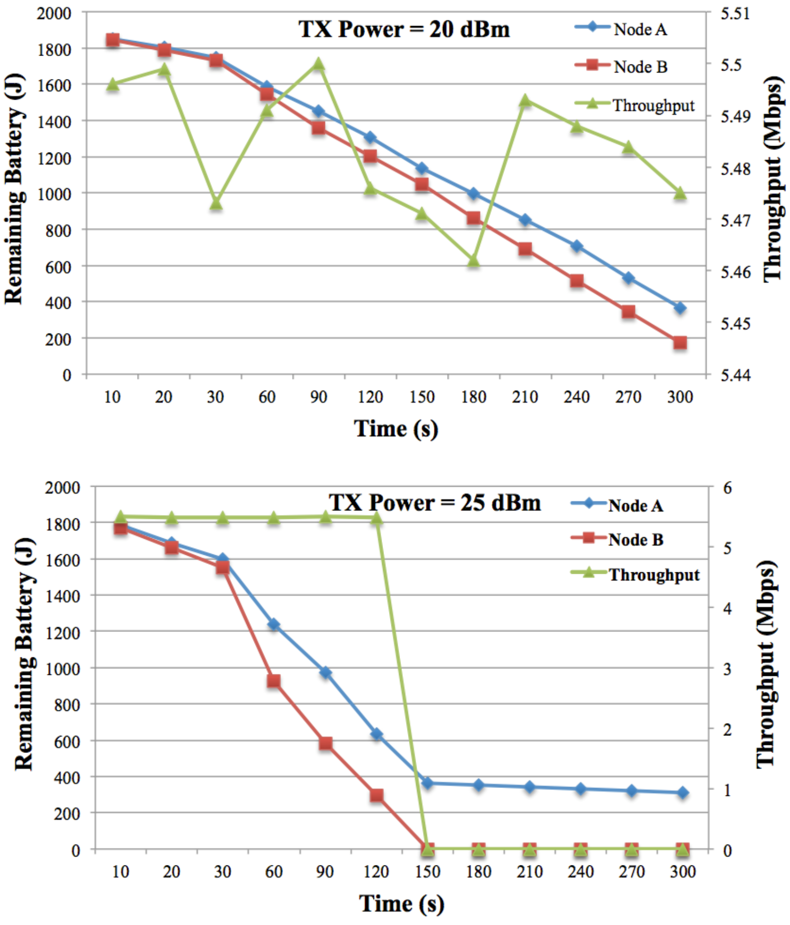


Fig. 7. Acquired results of the first scenario

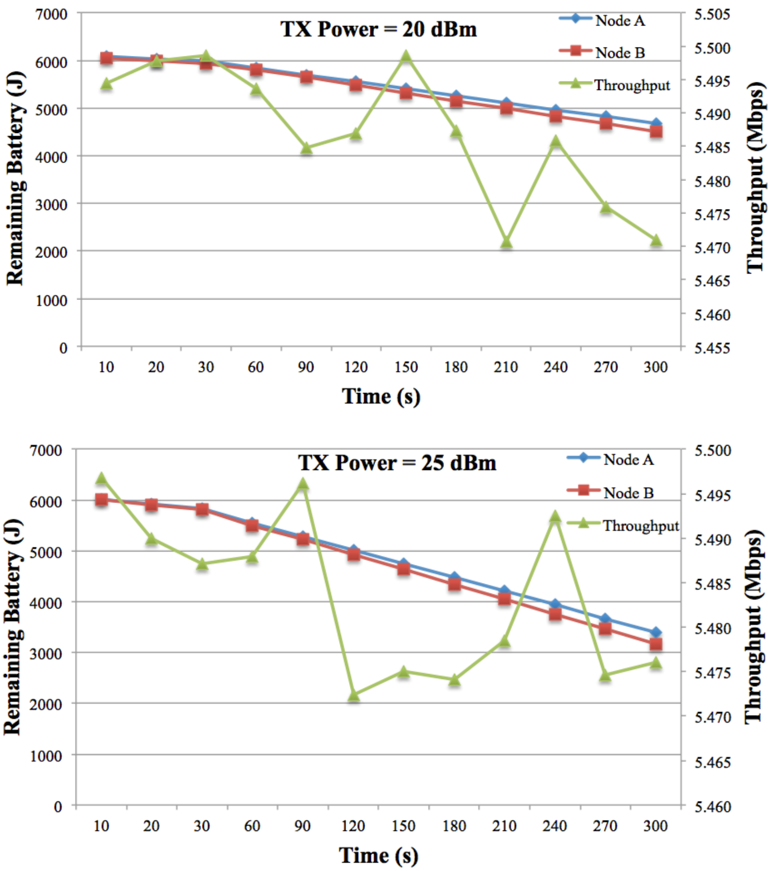


Fig. 8. Acquired results of the second scenario (first phase)

setup, is not strong enough to establish a reliable wireless link. Average quantified RSSI values at node B for TX powers equal to 15, 20 and 25 dBm are respectively -73.15, -63.11 and -60.51 dBm while the defined threshold value to create a reliable wireless link is regarded as -70 dBm. Analogous simulation setting is applied for the second scenario, which is conducted in 2 phases and its topology is illustrated in Fig. 6.

In the first stage, generated UDP traffic at node A, after traversing 3 hops will be delivered to node B. Same performance metrics were quantified at node B when transmission powers are 20 and 25 dBm. Regarding the represented graphs in Fig. 8, by tuning the amount of TX power in each radio of every mesh router, tangible amount of battery charge will be preserved while the link throughput through the whole experiment persists in the proximity of expected value.

To emphasize the importance of TX power optimization process in the offered framework, it is suffice to regard that in Fig. 8, the remaining battery of nodes A and B at the end of the experiment for TX power 20 dBm are 4679 and 4508 J respectively, while for the TX power 25 dBm, remained charge in the nodes' batteries at least 1300 J is less than the former case. In the next phase of second scenario, it is supposed that the wireless link, which was functioning in channel 48, is broken and according to the proposed flowchart, disconnected radio of node B modifies its antenna orientation and channel frequency to establish a reliable link with its nearest neighbor, which is operating in channel 36. Although by this topology change the number of available hops between nodes A and B will be increased, there is not any evident difference between acquired results in conducted phases of this scenario. Also, in addition to the reduction of remaining battery charge in node B in comparison to the previous phase, specific amount of energy should be considered for feeding the stepper motors to change the orientation of respective antenna at this node. With regard to the physical dimension of selected antenna and motor specifications, particular value should be included in the consumed energy at node B which respecting the battery lifetime and stationary position of nodes can be neglected. Fig. 9 demonstrates the obtained results for the second phase of second scenario.

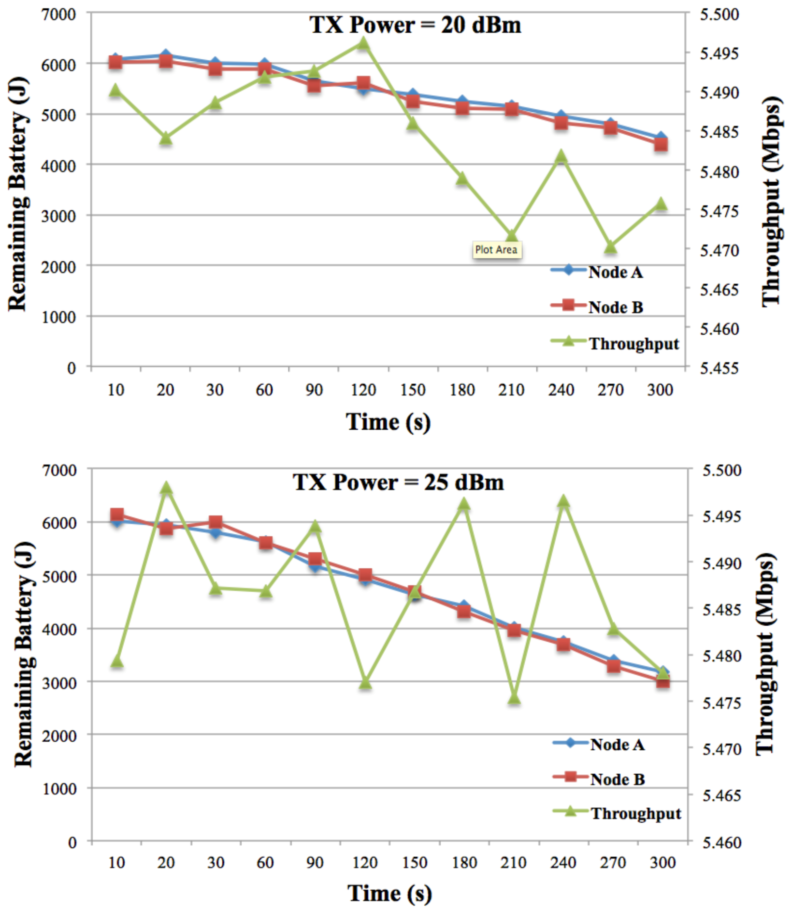


Fig. 9. Acquired results of the second scenario (second phase)

6 Conclusion

As a matter of fact, this study is ended to the findings, which corroborate the outcome of a great deal of the previous works in this field. Actually, the proposed framework for utilization of flexible antennas in the mesh networks not only preserve tangible amount of energy in long term operation of wireless topology, but also alleviate the impact of inter-interference among the utilized nodes through the mesh network. Thus, conducted research equips the scholars with an effective solution to deploy green wireless mesh networks through different geographical locations regarding the promotion of achieved throughput and reduction of power consumption for each mesh router. However, more research on this topic needs to be carried out before the appropriate association of system components is more clearly perceived.

In the future works, we are planning to improve the operation of each component within the proposed architecture and evaluate the system functionality by means of other familiar performance metrics in much more complicated network scenarios and topologies.

References

1. Das, S.M., Pucha, H., Koutsonikolas, D., Hu, Y.C., Peroulis, D.: DMesh: Incorporating practical directional antennas in multichannel wireless mesh networks. IEEE Journal on Selected Areas in Communications 24, 2028–2039 (2006)
2. Muthaiah, S., Iyer, A., Karnik, A.: Design of high throughput scheduled mesh networks: A case for directional antennas. In: GLOBECOM, pp. 5080–5085. IEEE (2007)
3. Zhou, W., Chen, X., Qiao, D.: Practical Routing and Channel Assignment Scheme for Mesh Networks with Directional Antennas. In: 2008 IEEE International Conference on Communications, pp. 3181–3187 (2008)
4. Zhang, J.: Capacity analysis of wireless mesh networks with omni or directional antennas. In: INFOCOM 2009, pp. 2881–2885. IEEE (2009)
5. Zhang, J., Zheng, Z., Jia, X.: Improved Topology Control Method for Maximizing Traffic Delivery Ratio in Wireless Mesh Networks with Directional Antennas. In: 2009 IEEE International Conference on Communications, pp. 1–5 (2009)
6. Luo, L., Raychaudhuri, D., Liu, H.: Channel Assignment, Stream Control, Scheduling and Routing in Multi-Radio MIMO Wireless Mesh Networks. In: 2009 IEEE International Conference on Communications Workshops, pp. 1–5 (2009)
7. Huang, B., He, Y., Perkins, D.: Investigating Deployment Strategies for Multi-radio Multi-channel Residential Wireless Mesh Networks. In: 2009 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, pp. 147–153 (2009)
8. Kandasamy, S., Campos, R., Morla, R.: Using Directional Antennas on Stub Wireless Mesh Networks: Impact on Throughput, Delay, and Fairness. In: ICCCN IEEE, pp. 1–6 (2010)
9. Okada, H.: Kenichi Mase: Performance analysis of wireless mesh networks with three sector antennas. In: Proceedings of the 6th International Wireless, p. 1232. ACM Press, New York (2010)

10. Liu, X., Okada, H.: Kenichi Mase: Performance of Wireless Mesh Networks with Three Sector Antenna. In: Sixth International Conference on Mobile Ad-hoc and Sensor Networks, pp. 146–153 (2010)
11. Ben-Othman, J., Mokdad, L.: Sectorial Coverage in a Deployment of a WMN Backbone Based on Directional Antennas. In: 2011 IEEE Global Telecommunications Conference, GLOBECOM 2011, pp. 1–5 (2011)
12. Ben-Othman, J., Mokdad, L., Cheikh, M.: A new architecture of wireless mesh networks based IEEE 802.11s directional antennas. In: IEEE International Conference on Communications, ICC (2011)
13. Optimized Link State Routing Protocol (OLSR), `http://www.ietf.org/rfc/rfc3626.txt`
14. IEEE 802.11s Tutorial, `http://www.ieee802.org/802_tutorials/06.../802.11s_Tutorial_r5.pdf`
15. Siris, V.A., Delakis, M.: Interference-aware channel assignment in a metropolitan multi-radio wireless mesh network with directional antennas. Computer Communications 34, 1518–1528 (2011)
16. Ding, Y., Xiao, L.: Channel allocation in multi-channel wireless mesh networks. Computer Communications 34, 803–815 (2011)
17. Liu, Q., Jia, X., Zhou, Y.: Topology control for multi-channel multi-radio wireless mesh networks using directional antennas. Wireless Networks 17, 41–51 (2010)
18. Sadeghianpour, N., Chuah, T.C., Tan, S.W.: Improved topology control for multiradio multichannel wireless mesh networks with directional antennas. In: 2011 7th International Wireless Communications and Mobile Computing Conference, pp. 1708–1712 (2011)
19. Shillingford, N., Poellabauer, C.: A framework for route configurability in power-constrained wireless mesh networks. Ad Hoc Networks 8, 857–871 (2010)
20. Avallone, S.: An energy efficient channel assignment and routing algorithm for multi-radio wireless mesh networks. Ad Hoc Networks 10, 1043–1057 (2012)
21. Saputro, N., Akkaya, K., Uludag, S.: A survey of routing protocols for smart grid communications. Computer Networks 56, 2742–2771 (2012)
22. De la Oliva, A., Banchs, A., Serrano, P.: Throughput and energy-aware routing for 802.11 based mesh networks. Computer Communications 35, 1433–1446 (2012)
23. Ma, C., Yang, Y.: A Battery Aware Scheme for Energy Efficient Coverage and Routing. In: IEEE GLOBECOM, pp. 1113–1117 (2007)

Scheme for Assigning Security Automatically for Real-Time Wireless Nodes via ARSA

Rajesh Duvvuru[1], Sunil Kumar Singh[1], Gudikhandula Narasimha Rao[2], Ashok Kote[3], Bangaru Bala Krishna[4], and Moturu Vijaya Raju[5]

[1] Department of Computer Science Engineering, National Inst. of Technology, Jamsehdpur, Jharkhand, India
[2] Department of Computer Science Engineering, K.I.T.S, Guntur, A.P., India
[3] Department of Computer Science Engineering, L.I.M.A.T, Vijayawada, A.P., India
[4] Department of Computer Science Engineering, T.I.T.S, Hyderabad, A.P., India
[5] Department of Information Technology, U.R.C.E.T, Vijayawada, A.P., India
{rajeshduvvuru.cse,sunilkrsingh.cse}@nitjsr.ac.in,
{gudikhandula,kote.ashok,bbk.tits,vijayaraju.m}@gmail.com

Abstract. Security and ease of use are two fundamental requirements of wireless network users. But they conflict with each other. The strongly secure network will put a lot of load on the server for security related work which may hamper the packet delivery ratio. But strong security is indispensable for maintaining the confidentiality of information in current real-time wireless communication networks. This research work combines both, the concepts of network security and the packet scheduling issues of the wireless data packets. Most of the users using wireless network are unaware about what level of security is needed for them. We present a new Automated Security-Aware Packet Scheduling Strategy or ASPS for real-time wireless network. This ASPS algorithm assigns the desirable level of security automatically to the respective data packets with guarantee of deadlines for the packets. Our simulation result proves that our proposal is performing better than existing algorithms in terms of the quality of security, guarantee ratio and reducing the load on the network switch.

Keywords: Load on network switch, security identification adapter, advanced radius authentication server, Automated Security-Aware Packet Scheduling Strategy, Wireless LAN Security.

1 Introduction

Wireless technology has created one of the greatest revolutions in the usage of mobile gadgets. In most of the wireless applications like accessing internet, video calling, live TV, and many other useful applications, wireless technology plays a vital role.

According to the recent survey report of 2008, approximately 86% of total network across the surveyed cities appeared to be vulnerable, since 37% of these networks appeared to have no encryption and 49% of networks to be using WEP encryption [1].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 185–196, 2013.

Most of innovative applications made the wireless technology even more interesting a research area for the scientists and research scholars. Exchanging the information from the source to destination in a confidential and reliable manner, from one mobile node to another mobile node is always a challenging task. For example, data in wireless networks is broadcast in the form of radio waves. In the computer science literature there are lot of data-security related algorithms, yet information exchange faces a lots of security allied attacks like identity theft(MAC spoofing) [2], Man-in-middle [3], Denial of service [4], Network injection [5], and the like. All these security threats have worsened the secure data transmission in wireless networks which leads to bad service being provided by the wireless communication companies. It has also affected them badly in the commercial sense. This is the one of the main reason why security became a foremost part in the field of wireless networks. Plenty of security protocols have been introduced in the arena of wireless network for secure data transmission. The IEEE 802.11 family has mostly concentrated on the security issues like authentication and confidentiality [6]. Most of the data transmission was in static mode level of security. This reflects that, mobile nodes are using same level of security for the data transmission. Different levels of security should be maintained for different type of users. If the user is a highly important, for that particular user, security should be high, compared to the other mobile and wireless nodes [8]. These security levels should be maintained for each and every specific user. For instance, in the Banking system, the current data transaction record will need more level of security than data transaction record twelve years before. Modelling the security mechanism for wireless networks in flexible way is always a challenging task. In this paper, we present a new Automated Security-Aware Packet Scheduling Strategy or ASPS for real-time wireless network. This ASPS algorithm assigns the desirable level of security automatically to the desired data packets with guarantee of deadlines for the packets. Assigning the security level represents encrypting the data by applying the different cryptographic algorithms at each level. Maintaining same level of security is a disadvantage. This is one of the important fragile points of the wireless communication. To avoid this sort of security issues, we are assigning different levels of security for each and every wireless node according to the priority.

This paper comprises five major contributions (1) Simulator for which ASPS algorithm was implemented and tested. (2) Automatic assigning of security levels to the specific nodes (3) It combines the both security and guarantee ratio of packets (4) A new model of wireless data-packet have been designed (5)Novel performance evaluation by combining security, guarantee ratio and load on network switch. This paper is organized in the fallowing way: section 2 describes how the security and packet scheduling were achieved in the wireless communication. Section 3 explains about the ASPS algorithm. Section 4 discusses the simulation results. Lastly we will conclude and assert the future scope for this work in section 5.

2 Related Work

Our research work provides an automatic schema for assigning security levels to real-time wireless nodes. We are going to overcome the high level complexity of assigning security to the deserved N number of users sending or receiving the data through their wireless devices. These are some real-time wireless packet scheduling algorithms [9][10]. These packets are going to be scheduled to reduce the latency time and increase the performance of data transmission. Normally most of the people who are using wireless networks for data transmission are not much aware of technical aspects of the security issues, but every user wants to send the data in a confidential and secure way. In 802.11 wirelesses LAN, we can provide common level of security to all users where there is no differentiation between the high-end user who requires high-level of security, middle-level user who requires average-level of security and low-end user who require very less amount of security. By classifying the users in different levels we can provide an efficient data transmission and which will reduces the burden and delay in the network. For instance, in WLAN we have applied AES algorithm for encryption and decryption of data. Assume the network is capable of transmitting 2KB/sec of data, where the network devices are bounded at maximum 10 meter range. For the similar WLAN with same bandwidth and same network setup if DES algorithm is applied for secure data transmission it will transmit more than 2KB/sec for maximum 10 meter. Because the key length of AES ranges up to 256 bits and for DES is 56 bits. No doubt AES algorithm is much secure than DES algorithm in terms of security [11] whereas data transmission is low due to key size. The wireless device users who are not much considered about applying AES in WLAN are giving additional burden to the network; for these sorts of users, DES will be the best for data transmission and data-packets exchange is also fast. Our research combines the better secure transmission of the wireless data-packet, packet scheduling algorithm and reducing the network load for the fast data transmission.

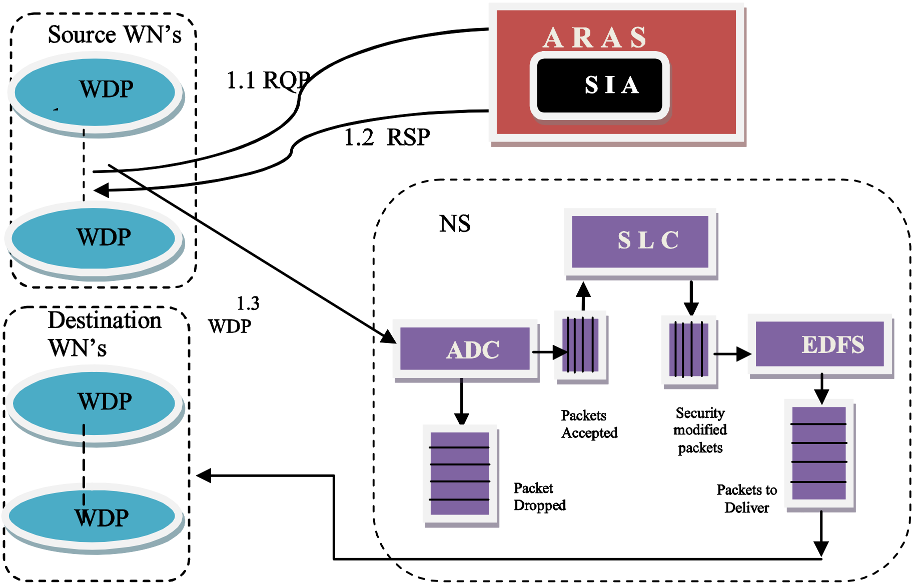


Fig. 1. Schematic Diagram of Network System

3 Advanced Security-Aware Packet Scheduling

In our research work we have modelled a novel Authentication server, Advanced Radius Server Authentication (ARSA), which will perform authentication of WN(Wireless Node) of particular network connected in Wireless Local Area Network (WLAN). Each and every WN works as both sender and receiver, which we call as the transceiver. Normal Radius Server Authentication (RSA)[12] will do perform authentication, authorization and accounting (AAA) of one particular WN by using its IP address, whether it is belongs to that network or not, if the WN is valid, then it will allow that particular WN to access all other resources in that specific network. In addition, we are making RAS to assign the security level for that particular WN by including the SIA (Security Identification Adapter). This SIA will recognize the IP address and assign the deserved security level to that particular WN. The IP addresses are classified into different classes [14], where each class got its own importance and according to its importance, SIA will assign the security level and acknowledge with AAA. For instance class E IP address of WN will be assigned higher level of security when we compared to the class D or C. Depending upon future requirement, further these IP address classes can be classified into sub classes. In addition to the Network Switch (NS), it contains the Admission Controller (ADC), Security Level Controller (SLC) and Earliest Dead line First Scheduling (EDFS), for further detail information please refers this research paper [13]. ADC will accept or reject the Wireless Data Packets (WDP) which are sent from the Wireless Node (WN), depending on their deadlines. SLC will increase the security level if it is having enough dead line. EDFS follows the policy of Earliest deadline first and processes WDPs from Security modified packets queue to Packets to Deliver queue.

3.1 Assumptions and Notations

In this Network we have designed three different types of packets they are request-packet (RQP), Response-Packet (RSP) and WDP. Two for communication between WNs and ARSA and other is communication between the WNs and NS. Before accessing the WLAN, WN should authenticated by means of ARSA, if the WN is valid then ARSA will grant the permission to access the network otherwise authentication will be rejected. Firstly, WN will send a RQP to the ARSA. RQP contains only IP Address (IPAi), and then the ARSA will check the RQP. If it is valid RQP, ARSA will sent back the RSP to WN over NS. RSPi represented by a tuple (ACi,SLi), where ACi denotes Permit/Reject of Access and SLi represents Security Level.

Once the source WN granted permission to access the network, then it will start communicating with NS. Then the WDP are directed from source WN to the destination WN through NS. Basically WDP is represented with a set of fields (ATi,PTi,SLi,Di)[8]. Here SLi and Di is represented with the security level and deadline of the packet i. ATi and PTi denotes arrival and processing time of packet i.

Equation (Eq)-1 specifies the formula for the calculation of deadline.

$$DLi >= CT_i - ST_i \tag{1}$$

Where STi starting time of the transmission of the ith packet, CTi is the completion time of the transmission, DLi is the packet's deadline.

To calculate the security operating cost without loss of simplification, we make use of formula Eq-2 to mold the security operating cost as the extra processing time experienced by packet i.

$$SOC_i = TT_i * (LS_i/MS) \tag{2}$$

Where SOCi represents security operating cost of the ith packet, TTi is the transmission time of the packet i, LSi is the level of security of the packet i, and MS is the maximum security level ranges from 1 to 10 according to the classes in the IP address.

Thus, the total processing time TPTi of packet i can be articulated as:

$$TPT_i = TT_i + SOC_i \tag{3}$$

For computing the load on network (LNS) switch we have used following equation 4.

$$LNS_i = (LS_i/MS) \tag{4}$$

3.2 The ASPS Algorithm

The ultimate aim of this research is to make the wireless users, confusion free, from assigning the security level for their wireless devices and reduce the load on NS, which results in improving the guarantee ratio and security level.

To accomplish this objective, we have designed the ASPS scheduling algorithm. Goal of ASPS is to maintain high guarantee ratio with automated security level. We can achieve high performance in terms of security level and guarantee ratio by instigating ARSA to our ASPS algorithm.

Fig.2 describes flow of the ASPS algorithm for the wireless transmission. The following steps will demarcate the procedure of the ASPS scheduling algorithm .Step1: WN will send RQP to the ARSA. RQP contains IPAi. Step2: After arrival of RQP from source WN, ARSA will check the RQP. If it is valid RQP, ARSA will send back the RSP to WN over NS. RSPi represented by a tuple (ACi,SLi). Step3: Once the RSP is received by the WN, depending on the RSP wireless node will be permitted to send and receive WDP on the network.Step4:If WN is permitted by ARSA it will access the Network through NS.Step5:WN sends WDPi to the NS , where WDPi denotes a set of attributes (ATi,PTi,SLi,Di).Step6: initializing the ASPS scheduler; the security standard of incoming WDPi and the number of rejected WDPi is set to zero. Hold until any incoming WDPi.Step7: If a packet i arrive at NS and it is the only packet available then process the packet instantly using its highest security level. The starting time (STi) and the completion time (CTi) of the WDPi are calculated.Step8: All the packets arriving in the NS during the time period [STi, CTi] are temporarily stored into a waiting queue in the ascending order of their deadlines. The starting time of the next packet STi + 1 is set to CTi.

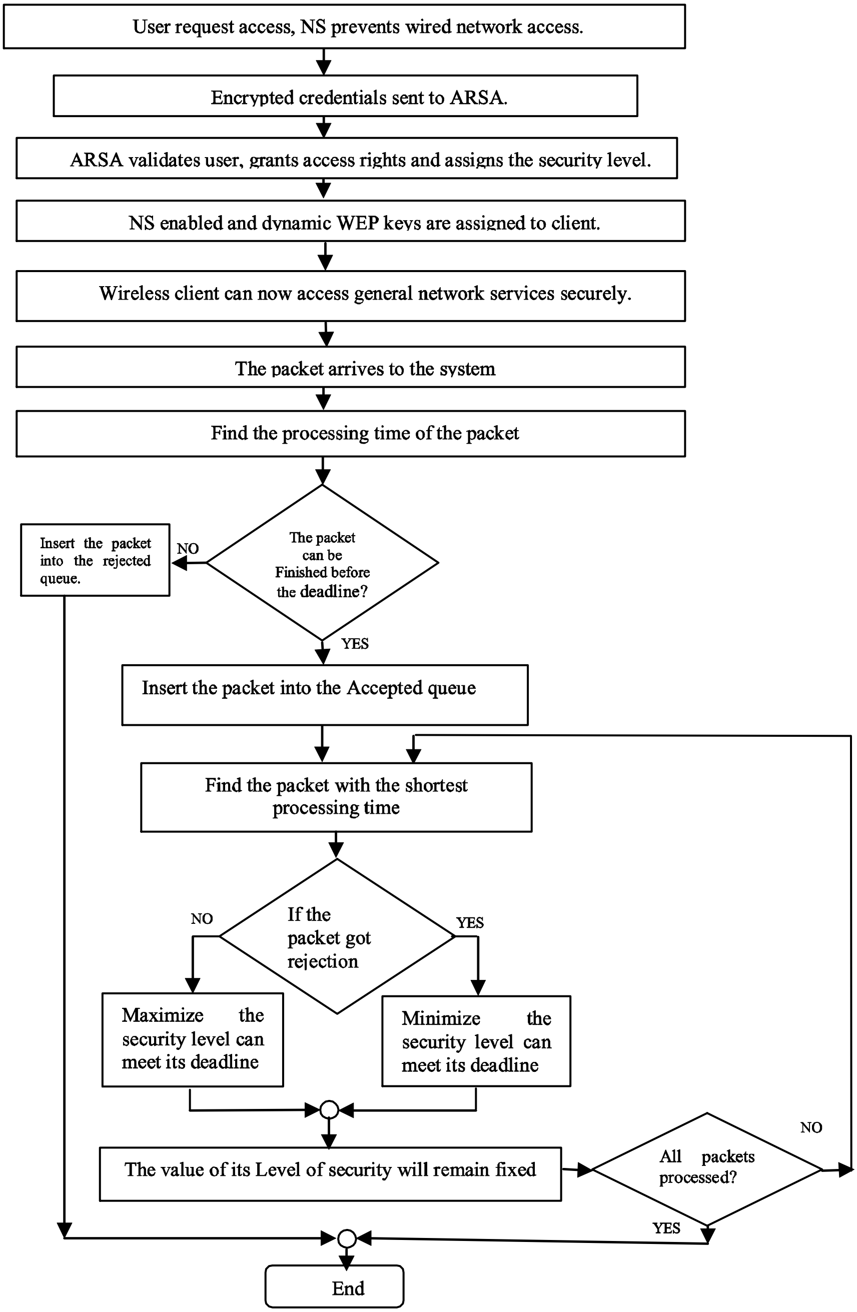

User request access, NS prevents wired network access.
Encrypted credentials sent to ARSA.
ARSA validates user, grants access rights and assigns the security level.
NS enabled and dynamic WEP keys are assigned to client.
Wireless client can now access general network services securely.
The packet arrives to the system
Find the processing time of the packet
The packet can be Finished before the deadline?
NO
Insert the packet into the rejected queue.
YES
Insert the packet into the Accepted queue
Find the packet with the shortest processing time
If the packet got rejection
NO
YES
Maximize the security level can meet its deadline
Minimize the security level can meet its deadline
The value of its Level of security will remain fixed
All packets processed?
NO
YES
End

Fig. 2. The flow chart of the proposed system

Step9: The ADC will decide, if a WDPi in the waiting queue will be accepted or rejected with the respect to the deadline of WDPi. The WDPi which will meet its deadline will be sent to the accepted queue or else guided towards the rejected queue. Step10: The WDPi that are there in accepted queue will be forwarded to the SLC for modification(increment or decrement) of security level dynamically depending upon the various aspects like availability of bandwidth, congestion and data traffic in the network etc. Step11: The security modified WDPi will be sent to the EDF scheduler through security modified queue. Step12: The WDPi arrived in EDF scheduler will be delivered to the destination WN, according to their processing time. The WDPi having less processing time will be delivered first. Step13: Until all the WDPi are processed, the step 7 to step12 will be repeated.

3.3 Implementation Issues

We have implemented and tested the performance of ASPS, SPSS, MIN and MAX algorithms in NS-2 . We have chosen IEEE 802.11 WIFI, wireless LAN protocol for data transmission in our simulation environment.

Simulation of MIN, MAX and SPSS algorithm: We have considered four basic parameters i.e. data size = 0.3 KB, Bandwidth = 0.5 MBPS, dead line = 0.2 No/ Sec. Load on network switch and Level of Security (it will vary depending upon the algorithm) for all four algorithms. MIN, MAX and SPSS algorithm is having common simulation environment. Here, we have taken total six WN, one RAS and one NS. We have assumed WN {0-2} as source nodes, WN {3-5} as destination nodes, RSA represented with number 6 and lastly NS as 7. Initially all the three WN {0-2} will communicate with the RAS for authentication to access the network, once the WN is permitted by the RAS. WN starts transmission of WDP to the destination WN {3-5} through NS. Here RAS will not assign any sort of LS to any WN. In MIN algorithm by default, WN-{0-2} data packets is encrypted with a very low encryption algorithm WEP by the WN itself before transmitting. Here we have considered both LS and LNS as 1. Whereas in MAX algorithm WN-{0-2} data packets are encrypted with WPA2 .Here LS and LNS is 10. Finally SPSS algorithm encrypts with WEP for WN-1(unimportant user),WPA 2 for WN-2 (unimportant user) and WPA2 for WN-3 (important user) and in this algorithms the LS and LNS varied from 1to 3. WN-2 is using WPA2 for encryption; even that user is not deserved for it, which leads to the load on network switch. These simulation results were plotted and discussed in section 4.

Simulation of ASPS algorithm: Lastly, we have conducted simulations for proposed algorithm. Here the simulation environment will differ a bit from rest of the three base line algorithm. Also we have considered similar environment, but RAS is replaced with ARAS. WN-{0-2} has started to access the network, and it is validated and assigned a specific LS depending up on IPA by ARAS.Here we have intentionally assumed WN-1 as unimportant user (requires low LS1, encrypted with WEP),WN-2 as important user (requires average LS-5, encrypted with WPA) and lastly WN-3 as extreme important user (requires average LS-10, encrypted with WPA2) [6]. After encrypting WDP with specified algorithms, the communication will take place between source and destination WN through NS. Figure 3 represents snapshot of simulation environment of ASPS algorithm.

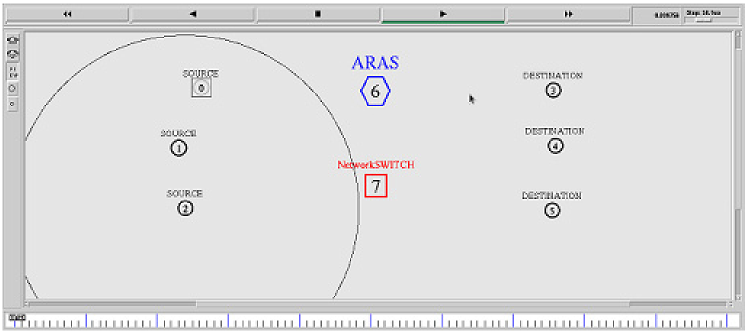


Fig. 3. Simulation of ASPS algorithm

4 Simulation Results

4.1 Comparison of ASPS with MIN, MAX and SPSS Algorithms

At this instant we are going momentarily summarize the baseline algorithms-SPSS, MIN and MAX. Further these algorithms are going to compare with the proposed ASPS algorithm. 1. MIN: All the WN's sends WDP's to NS with lower security level, here the specific WN user, who need high level security, can't do it and cost of guarantee ratio is enhanced at the cost of dropping SOC. 2.MAX: All the WN's sends WDP's to NS with higher security level, here most of the WN user, who don't require maximum security level, it leads to the load on the NS, where LS is improved, but guarantee ratio will be reduced. 3. SPSS: This algorithm will provide different levels of security to the different WN's user, which will results medium performance in levels of security and guarantee ratio of WDP. Limitation of SPSS is WN user have given individuality to assign LS to WDP's, which it leads to the unwanted user also will opt for high level of security to WDP's and sent to NS, which will increase burden to NS. We proposed a novel scheduling algorithm, which will give equal level of performance as that of SPSS in terms of LS and GR, in addition to this ASPS will assign automatic security levels to the specific user depending up on their necessity, which will reduce load drastically on NS.

4.2 Performance Evaluation

The performance of our approach will be evaluated by comparing the ASPS algorithm with SPSS, MAX and MIN. The performance measurement is based on mainly four parameters they are Guarantee ratio (GR), level of security (LS), overall performance (OP) and Load-on-Switch (LOS). Overall performance can be calculated by following Eq-5:

$$OP= (GR * LS) + LOS \tag{5}$$

4.3 Impact of Bandwidth

In the following experiment, we have varied bandwidth from 0.1 to 0.3 MBPS and compared our approach result with remaining three baseline algorithms namely SPSS, MIN, and MAX .It is a universal fact, whenever the bandwidth is high we can send more number of packets, high network bandwidth leads to short transfer times, which in turn result in short processing times of packets. This helps the WN to transmit and receive the WDP's efficiently. It leads to the high level of security that can be incremented, so that overall performance can be improved. Fig.5 Defines cross product of both Guarantee Ratio and Level of Security in addition of Load-on-Switch with respect to the bandwidth. The overall performance of ASPS is in peak state whereas SPSS, MIN and MAX follow next.

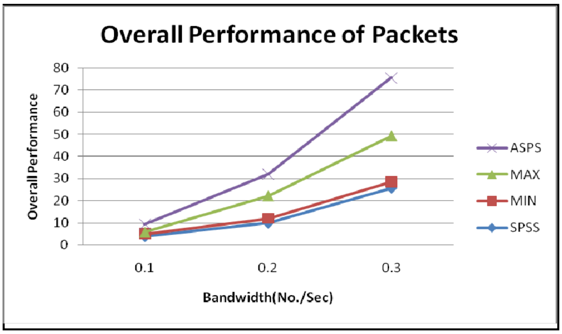


Fig. 4. Impact of bandwidth on Overall performance when data size = 0.3 KB, arrival rate = 0.5 No/Sec, and deadline = 0.2 No/Sec.

4.4 Impact of Arrival Rate

This research work is intended to compare ASPS stratagem with three baseline method. The first baseline method is called SSPS, which will assign different level of security for the incoming data packets .Secondly, MIN method which will assign smallest amount level of security to the incoming packets given to the network switch (NS) and finally, MAX method assigns the maximum amount of security to the data packets arriving at the network switch. To achieve this target, the arrival rate was increased of the incoming packets from 0.1 to 0.3 No./Sec. and the data size was set to 0.5 KB, the Bandwidth to 0.5mbps, and the deadline to 0.2 No./Sec. Fig. 5 Delivers ASPS will execute peak performance when we compared to the other baseline algorithms MIN, MAX, and SPSS in respect to the Arrival rate over the overall performance of the network system.

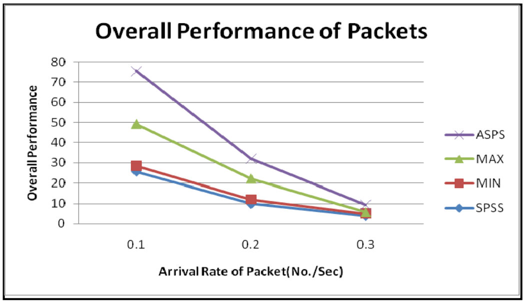


Fig. 5. Impact of arrival rate on Overall performance when data size = 0.3 KB, Bandwidth = 0.5MBPS, and deadline = 0.2 No/Sec

4.5 Impacts of Data Size

In this experimentation, we evaluate the efficiency of ASPS aligned with SPSS, MIN and MAX schemes changing the data size from 0.5 to 1.5KB. Further, we observe Fig. 6 On every occasion the data size is increased the overall performance of ASPS is incremented comparatively.

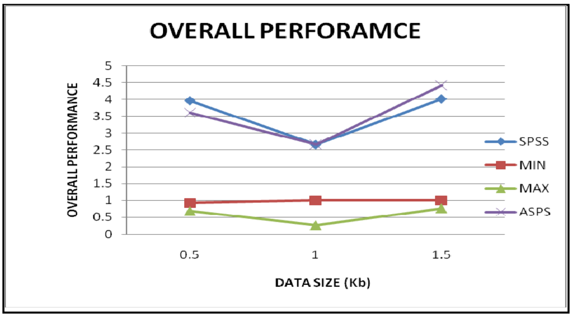


Fig. 6. Impact of Data size on Overall performance when Bandwidth = 0.5MBPS, arrival rate = 0.5 No/Sec and deadline = 0.2 No/Sec

4.6 Impacts of Deadlines

The deadline is another important concern, which has a great impression on our wireless network system. Initially, we vary our deadline from 0.1 to 0.3 no/sec. and we observe the consequence on the network system in detail. The figure below clearly reveals that when we vary the deadline from 0.1 to 0.3 no/sec, then the security level of the WDP's is automatically incremented in the network switch (NS) and eventually the performance is enhanced. The concept which unfolds this mechanism is whenever the deadline has some flexibility, then obviously there will be an advantage of time for that particular packet to get delivered and hence the security level will be increased for that particular packet because load on the network switch will be reduced. This verifies that the ASPS system has a high impact over SPSS, MIN or MAX in terms of security enhancement. Fig.19 exposes clearly that the overall performance is increasing exponentially along with growth in the deadline.

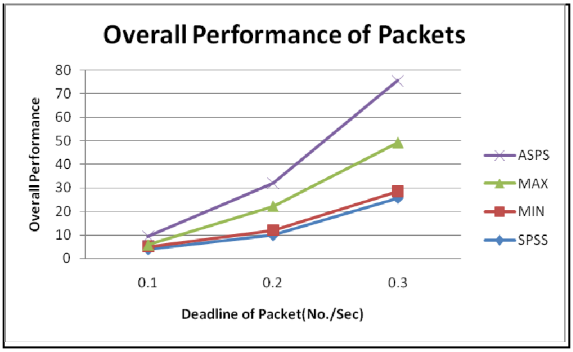


Fig. 7. Impact of Deadline on Overall performance of packets when data size = 0.3 KB, Bandwidth = 0.5MBPS, and arrival rate = 0.5 No/Sec

5 Conclusion and Future Scope

In the current real-time scenario wireless network has reached its summit. In spite of innumerable demands there still exists a lot of challenging task like quality of service, high-speed data transmission and secure communication etc. To overcome such a challenging task, we have designed and simulated an innovative approach (ASPS), which will result better in terms of guarantee ratio, level of security and load on network switch (for the fast data transmission).In our algorithm we are making Advanced Radius Authentication Server (ARAS) to think intelligently while assigning the security level for the incoming request packets, which will increase the data transmission and reduce the load on network switch. Comparing ASPS with other baseline algorithms, we have observed better performance regarding Load on network switch and Overall performance. Assigning dynamic security level to the wireless mobile node became interesting research area for the researchers. This research work provides future scope for Hybrid mechanism for assigning the security level to the specified user.

References

1. http://www.deloitte.com/assets/DcomIndia/Local%20Assets/Documents/All%20India%20Wifi%20Survey.pdf
2. Nagarajan, V., Arsaan, V., Huang, D.: Using Power Hopping to Counter MAC Spoof Attacks in WLAN. In: Conference on Consumer Communications and Networking Conference, pp. 1–5 (2010)
3. Eberz, S., Strohmeier, M., Wilhelm, M., Martinovic, I.: A Practical Man-In-The-Middle Attack on Signal-Based Key Generation Protocols. In: Foresti, S., Yung, M., Martinelli, F. (eds.) ESORICS 2012. LNCS, vol. 7459, pp. 235–252. Springer, Heidelberg (2012)
4. Huang, H., Ahmed, N., Karthik, P.: On a New Type of Denial of Service Attack in Wireless Networks: The Distributed Jammer Network. IEEE Transactions On Wireless Communications, 2316–2324 (2011)

5. Park, J.C., Kasera, S.K.: Securing Ad Hoc Wireless Networks Against Data Injection Attacks Using Firewalls. In: IEEE Wireless Communications and Networking Conference, Salt Lake City, pp. 1525–3511 (2007)
6. Lashkari, H., Mohammad, M., Danesh, S.: A Survey on Wireless Security protocols (WEP, WPA and WPA2/802.11i). In: Second IEEE International Conference on Computer Science and Information Technolog 2009, Kuala Lumpur, Malaysia, pp. 48–52 (2009)
7. m [7] Xiao Qin, Mohamed Alghamdi, Mais Nijim, Ziliang Zong, Kiranmai Bellam, Xiaojun Ruan,and Adam Manzanares. : Improving Security of Real-Time Wireless Networks through Packet Scheduling. In: IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS 2008, IEEE Communications Society, pp. 3273- 3279, (2011).
8. Moh'd, A., Jararweh, Y., Tawalbeh, L.: AES-512: 512-Bit Advanced Encryption StandardAlgorithm Design and Evaluation. In: 7th International Conference on Information Assurance and Security, pp. 292–297. Halifax, NS (2011)
9. Han, S.-J., Oh, H.-S., Park, J.: The improved Data Encryption Standard (DES) Algorithm. In: IEEE 4th International Conference on Spread Spectrum Techniques and Applications Proceedings, Jongan Park, pp. 1310–1314 (1996)
10. Dua, A., Bambos, N.: On The Fairness Delay Trade-off in Wireless Packet Scheduling. In: IEEE Proceedings of the Global Telecommunications Conference, Palo Alto, pp. 25–48 (2005)
11. Alanazi, H.O., Zaidan, B.B., Zaidan, A.A., Jalab, H.A., Shabbir, M., Al-Nabhani, Y.: New Comparative Study Between DES, 3DES and AESwithin Nine Factors. Journal of Computing, 152–157 (2010)
12. http://www.rsa.com/rsasecured/guides/imp_pdfs/Cisco_VPN3K_47_AuthMan7.1.pdf
13. Mattihalli, C.: Designing and Implementing of Earliest Deadline First scheduling algorithm on Standard Linux. In: 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing, Green Computing and Communications (GreenCom), Hangzhou, China, pp. 901–906 (2010)
14. Bruno, R., Conti, M., Gregori, E.: Throughput Analysis and Measurements in IEEE 802.11 WLANs with TCP and UDP Traffic Flows. IEEE Transactions On Mobile Computing, 171–186 (2008)

The Outage Analysis of Inter-relay Coded Cooperation over Nakagami-m Fading Channels

Prabhat Kumar Sharma[1] and Parul Garg[2]

[1] Division of Electronics and Communication Engineering
Netaji Subhas Institute of Technology
New Delhi 110078, India
prabhatmadhavec1@gmail.com

[2] Division of Electronics and Communication Engineering
Netaji Subhas Institute of Technology
New Delhi 110078, India
parul_saini@yahoo.co.in

Abstract. In this paper we examine the outage behavior of a cooperative system, which utilizes the inter-relay coded cooperation. The outage analysis is done for independent, flat Nakagami-m fading channels. A closed form solution for outage probability is derived constrained on instantaneous received power which follows the Gamma distribution.

Keywords: Cooperative communication, outage probability, Nakagami-m fading

1 Introduction

Transmit diversity can be achieved in a wireless system by the cooperation among single antenna users to form a virtual array of antennas (VAA) [1,2]. Various cooperative protocols and their performance analysis were presented in [3] - [5]. In [5], an alternative mechanism, known as *coded cooperation*, was proposed in which concept of channel coding was associated with cooperative signalling. In coded cooperation each code word is essentially partitioned into two blocks, each of which is transmitted by one of the cooperative partners. Therefore the partners transmit some incremental redundancy instead of just repeating the received bits as in amplify and forward or decode and forward schemes.

An inter-relay cooperation scheme was presented in [6] for decode and forward relaying. This scheme allows the exchange of information among the relays (forming the VAA) to mitigate the propagation errors. Our contributions through this paper are (i) we present an inter-relay coded cooperation (IRCC) scheme, which improves the resource utilization within the VAA, and (ii) we have reached to a closed form expression of outage probability for the presented IRCC scheme. We consider Nakagami-m fading which is more generalized form of fading, as it can be used to model different fading channels through distinct values of parameter m [7]. The outage probability was defined in [8] in terms of instantaneous received power as the probability that the received power falls

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 197–206, 2013.

below a certain specified threshold. The present outage analysis is constrained on the instantaneous received power [9], however, the system model of present analysis is entirely different to that of [9].

The rest of this paper is outlined as follows. The system model is described in section 2. The closed form expression for the outage probability is derived in section 3. Numerical results are given in section 4. Section 5 concludes the paper with a brief note on the future scope of this work.

2 System Model

We have taken a simple model which consists of a source (s), destination (d) and two relays (r_1 and r_2) as shown in Fig. 1. As a practical assumption for low cost systems, we consider half duplex transmission, where a node cannot transmit and receive simultaneously. The inter-relay link is assumed to be reciprocal. To avoid multiple access interference links are considered to be orthogonal. Convolutional coding is been employed to convert N_0 bits of each source block into a codeword of length N.The coding mechanism is known to the source s, both the relays r_1 and r_2, and destination d.

The transmission is taking place in three phases. In first frame, source (s) broadcasts N_1 of N bits, which are received by each relay and destination. We define here the reliable relay as the relay which has correctly received the first phase broadcast. Further, If a relay is not reliable, it is unreliable. In second phase source (s) and reliable relay (or relays) will broadcast N_2 bits and destination along with unreliable relay, if any, will listen to it. N_1 and N_2 bits are obtained from length N codeword by partitioning it with rate R_1 and R_2 using a rate compatible punctured convolutional (RCPC) code [10]. Hence $N_1 + N_2 = N$. We define the cooperation ratio (α) as

$$\alpha = \frac{N_1}{N}. \tag{1}$$

The correctness of reception at relays will be validated on the basis of their cyclic redundancy checks (CRCs). The relay at which CRCs are not validated, will remain silent in the second frame. The quality of reception at each relay is obviously a function of the statistics of its channel with source. In the third phase the unreliable relay, will transmit the N_2 bits, if it received these bits correctly from the source and reliable node's broadcast in second phase. It may be noted that, although the link from source to unreliable relay was not good enough to receive correctly in the first phase as the power was not sufficient enough to label the threshold, this insufficient power will add up to the power received by the unreliable relay over relay to relay (inter-relay) link in second phase. In third phase, while making the decision at unreliable relay, we will consider the sum of two powers for comparison with threshold. As the reception is very energy economic than transmission, IRCC requires no additional energy while enhancing the resource utilization. However, the gains of IRCC don't come for free, it requires an extra time slot.

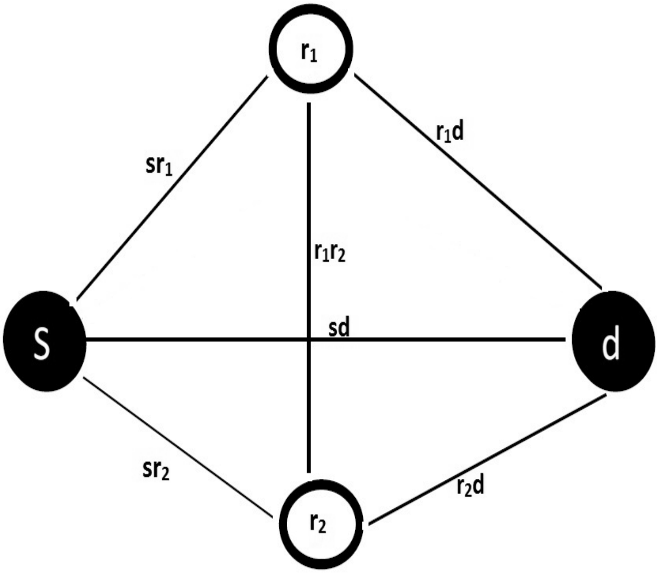


Fig. 1. System Model

We define P_n to be the threshold power so that if the instantaneous received power over any link falls below this threshold, the link is said to be in outage and corresponding outage event is $Pr\{P < P_n\}$. Thus for a link, corresponding outage event can be defined as,

$$P_{out} = Pr\{P < P_n\} = \int_0^{P_n} f(p)dp, \tag{2}$$

where p is the instantaneous received power and $f(p)$ is the probability density function (p.d.f) of p.

3 Outage Analysis

For Nakagami-m distribution, instantaneous received power (p) has a Gamma distributed p.d.f., hence outage probability as given in (2) for Nakagami-m fading can be written as,

$$P_{out} = \int_0^{P_n} \frac{1}{\Gamma(m)} \left(\frac{m}{\tilde{P}}\right)^m P^{m-1} \exp\left(\frac{-mP}{\tilde{P}}\right) dP = 1 - \frac{\Gamma\left(m, \frac{mP_n}{\tilde{P}}\right)}{\Gamma(m)}, \tag{3}$$

where $\tilde{P}$ denotes the average value of received power over the fading and shadowing effects. The function $\Gamma(m)$ is the Gamma function defined as $\Gamma(m) = \int_0^{\infty} x^{m-1} \exp(-x)dx$, and $\Gamma(m, \tau)$ is the upper incomplete Gamma function defined as $\Gamma(m,\tau) = \int_{\tau}^{\infty} x^{m-1} \exp(-x) dx$. We now define the outage events corresponding to each of four cooperative cases possible for the given system model described in section 2, as

- Case 1: When r_1, and r_2 both decode correctly i.e. if $P_{sr_1} > P_n/\alpha$, and $P_{sr_2} > P_n/\alpha$, the outage event is,
  $[\{\alpha P_{sd} + (1-\alpha)\left[P_{r_1 d} + P_{r_2 d}\right]\} < P_n]$
- Case 2: When r_1, and r_2 both don't decode correctly i.e. if $P_{sr_1} < P_n/\alpha$, and $P_{sr_2} < P_n/\alpha$
  the outage event is, $[\alpha P_{sd} < P_n]$
- Case 3: When only r_1 decodes correctly, and r_2 doesn't i.e. if $P_{sr_1} > P_n/\alpha$, and $P_{sr_2} < P_n/\alpha$
  then there are two further possibilities,
- Case 3(a): If $\{P_{sr_2} + P_{r_1 r_2}\} > P_n/\alpha$,
  the outage event is,
  $[\{\alpha P_{sd} + (1-\alpha)\left[P_{r_1 d} + P_{r_2 d}\right]\} < P_n]$, or
- Case 3(b): If $\{P_{sr_2} + P_{r_1 r_2}\} < P_n/\alpha$,
  the outage event is,
  $[\{\alpha P_{sd} + (1-\alpha) P_{r_1 d}\} < P_n]$
- Case 4: When only r_2 decodes correctly, and r_1 doesn't i.e. if $P_{sr_1} < P_n/\alpha$, and $P_{sr_2} > P_n/\alpha$
  then there are two further possibilities,
- Case 4(a): If $\{P_{sr_1} + P_{r_1 r_2}\} > P_n/\alpha$,
  the outage event is,
  $[\{\alpha P_{sd} + (1-\alpha)\left[P_{r_1 d} + P_{r_2 d}\right]\} < P_n]$, or
- Case 4(b): If $\{P_{sr_1} + P_{r_1 r_2}\} < P_n/\alpha$,
  the outage event is,
  $[\{\alpha P_{sd} + (1-\alpha) P_{r_2 d}\} < P_n]$

Assuming all the links to be mutually independent, the overall outage probability for these four disjoint cases, can be written as given in (4).

$$\begin{aligned}
P_{out} = {} & Pr\{P_{sr_1} > P_n/\alpha\} Pr\{P_{sr_2} > P_n/\alpha\} Pr\{\alpha P_{sd} + (1-\alpha)\left[P_{r_1 d} + P_{r_2 d}\right] < P_n\} \\
& + Pr\{P_{sr_1} < P_n/\alpha\} Pr\{P_{sr_2} < P_n/\alpha\} Pr\{\alpha P_{sd} < P_n\} \\
& + Pr\{P_{sr_1} > P_n/\alpha\} Pr\{P_{sr_2} < P_n/\alpha\} Pr\{\{P_{sr_2} + P_{r_1 r_2}\} > P_n/\alpha\} \\
& \times Pr\{\alpha P_{sd} + (1-\alpha)[P_{r_1 d} + P_{r_2 d}] < P_n\} \\
& + Pr\{P_{sr_1} > P_n/\alpha\} Pr\{P_{sr_2} < P_n/\alpha\} Pr\{\{P_{sr_2} + P_{r_1 r_2}\} < P_n/\alpha\} \\
& \times Pr\{\alpha P_{sd} + (1-\alpha) P_{r_1 d} < P_n\} \\
& + Pr\{P_{sr_1} < P_n/\alpha\} Pr\{P_{sr_2} > P_n/\alpha\} Pr\{\{P_{sr_1} + P_{r_1 r_2}\} > P_n/\alpha\} \\
& \times Pr\{\alpha P_{sd} + (1-\alpha)[P_{r_1 d} + P_{r_2 d}] < P_n\} \\
& + Pr\{P_{sr_1} < P_n/\alpha\} Pr\{P_{sr_2} > P_n/\alpha\} \times Pr\{\{P_{sr_1} + P_{r_1 r_2}\} < P_n/\alpha\} \\
& \times Pr\{\alpha P_{sd} + (1-\alpha) P_{r_2 d} < P_n\} \qquad (4)
\end{aligned}$$

We have further assumed the average link powers of all the relay to destination links to be same, i.e.

$$\tilde{P}_{r_1 d} = \tilde{P}_{r_2 d} = \tilde{P}_{av}$$

We also consider the average power of inter-relay link to be equal to that of sorce to destination link. Now, using (3), the overall outage probability can be written as given in (5), (details of (5) are given in appendix).

$$
\begin{aligned}
P_{out} = {} & \left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right]\left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right]\left[A_1 - A_2\sum_{k=0}^{2m-1} A_3 C\beta(k+1, m)\,{}_1F_1\left[m; k+m+1; \frac{BP_n}{\alpha}\right]\right] \\
& + \left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right]\left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right]\left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right] + \left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right] \\
& \times \left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right]\left[\frac{\Gamma\left(2m, \frac{mP_n/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right]\left[A_1 - A_2\sum_{k=0}^{2m-1} A_3 C\beta(k+1, m)\,{}_1F_1\left[m; k+m+1; \frac{BP_n}{\alpha}\right]\right] \\
& + \left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right]\left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right]\left[1 - \frac{\Gamma\left(2m, \frac{mP_n/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right] \\
& \times \left[A_1 - A_2\sum_{k=0}^{m-1} A_3 C\beta(k+1, m)\,{}_1F_1\left[m; k+m+1; \frac{BP_n}{\alpha}\right]\right] + \left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right]\left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right] \\
& \times \left[\frac{\Gamma\left(2m, \frac{mP_n/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right]\left[A_1 - A_2\sum_{k=0}^{2m-1} A_3 C\beta(k+1, m)\,{}_1F_1\left[m; k+m+1; \frac{BP_n}{\alpha}\right]\right] \\
& + \left[1 - \frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_1}}\right)}{\Gamma(m)}\right]\left[\frac{\Gamma\left(m, \frac{mP_n/\alpha}{\tilde{P}_{sr_2}}\right)}{\Gamma(m)}\right]\left[1 - \frac{\Gamma\left(2m, \frac{mP_n/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right] \\
& \times \left[A_1 - A_2\sum_{k=0}^{m-1} A_3 C\beta(k+1, m)\,{}_1F_1\left[m; k+m+1; \frac{BP_n}{\alpha}\right]\right]
\end{aligned} \tag{5}
$$

where $\beta(u, v)$ is beta function defined as $\beta(u, v) = \int_0^1 x^{u-1}(1-x)^{v-1}dx$, and ${}_1F_1[a; b; z]$ is the confluent hypergeometric function [11], and

$$
\begin{aligned}
A_1 &= [1 - \varphi], \\
A_2 &= \left[\frac{1}{\Gamma(m)}\left(\frac{m}{\tilde{P}_{sd}}\right)^m \exp\left(-\frac{mP_n}{(1-\alpha)\tilde{P}_{av}}\right)\right], \\
A_3 &= \left[\frac{1}{k!}\left(\frac{m\alpha}{(1-\alpha)\tilde{P}_{av}}\right)^k\right], \\
B &= \left[\frac{m\alpha}{(1-\alpha)\tilde{P}_{av}} - \frac{m}{\tilde{P}_{sd}}\right], \\
C &= \left[\frac{P_n}{\alpha}\right]^{k+m}
\end{aligned} \tag{6}
$$

We, in (5), have found analytically the closed form expression of the outage probability P_{out} for generalized Nakagami-m fading.

4 Numerical Results

For numerical analysis we assume that the average power of all the links is equal. Fig. 2 shows the effect of cooperation ratio over the outage behavior for different

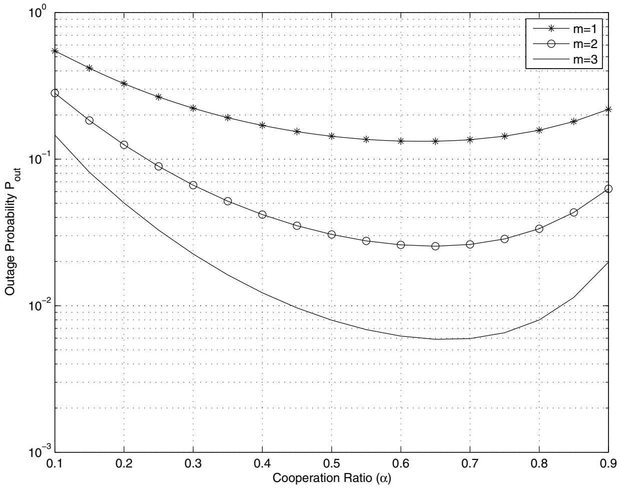


Fig. 2. Outage Probability vs Cooperation Ratio(α) for different values of m

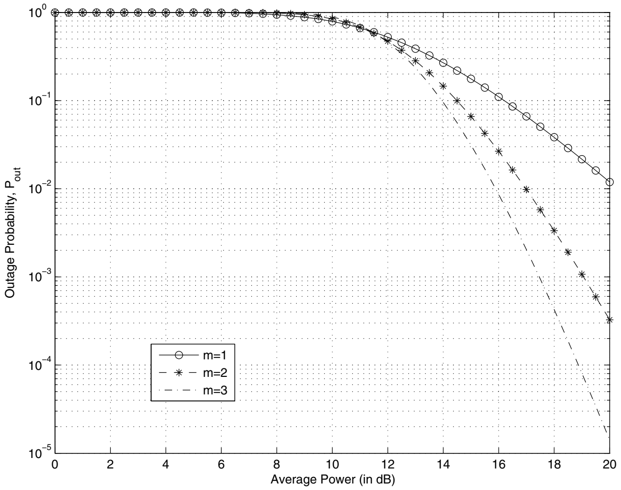


Fig. 3. Outage Probability vs Average Power(P_{av}) for different values of m

values of parameter m, assuming the threshold power and the average power both to be equal to 10 dB. Fig. 3 shows the plots of the outage probability given in (5), w. r. t. average power for different values of Nakagami parameter m. Here we consider threshold power $P_n = 5$ dB, and cooperation ratio $\alpha = 0.5$. The observations from this figure are very intuitive in sense that the performance of the system is increasing (i.e. outage probability is decreasing) on improving the channel conditions (i.e. increasing the value of parameter m).

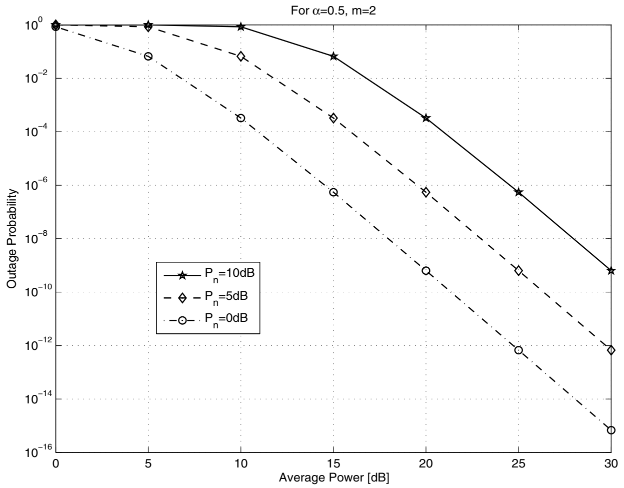


Fig. 4. Outage Probability vs Average Power(P_{av}) for different values of P_n

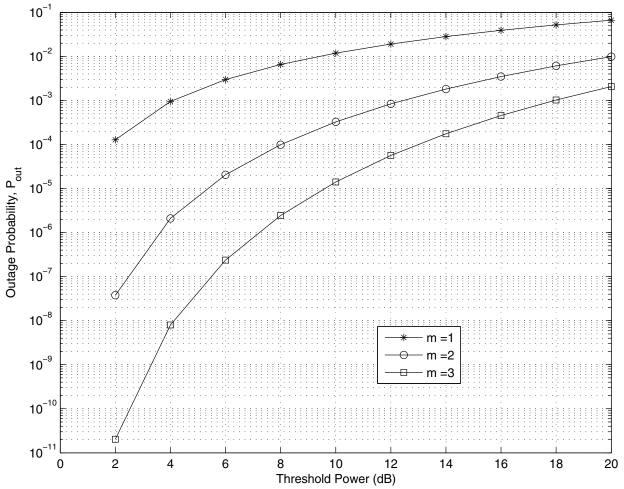


Fig. 5. Outage Probability vs Threshold Power (P_n) for different values of m

Fig. 4 plots the outage probability with respect to average power for different values of threshold power with $m = 2$ and $\alpha = 0.5$. We, from this figure, can observe that the system remains in the outage as long as the average power is below the threshold value. The variation of outage probability with threshold power for different values of parameter m, assuming average power $\tilde{P} = 20$ dB and cooperation ratio $\alpha = 0.5$, can be observed in Fig. 5.

5 Conclusion

This paper, using instantaneous power based approach, investigates the outage behavior of the given cooperative communication system with inter relay coded cooperation. The exact closed form expression for outage probability is obtained under Nakagami-m fading environment for the same. This work can be extended for multiple relays based system.

References

1. Sendonaris, A., Erkip, E., Aazhang, B.: User cooperative diversity Part-I: System description. IEEE Trans. commun. 51(11), 1927–1938 (2003)
2. Sendonaris, A., Erkip, E., Aazhang, B.: User cooperative diversity Part-II: Implementation aspects and performance analysis. IEEE Trans. commun. 51(11), 1939–1948 (2003)
3. Laneman, J.N., Tse, D.N.C., Wornell, G.W.: Cooperative diversity in wireless networks: Efficient protocols and outage behaviour. IEEE Trans. Inform. Theory 50(12), 389–400 (2004)
4. Laneman, J.N., Wornell, G.W., Tse, D.N.C.: An efficient protocol for realizing cooperative diversity in wireless networks. In: Proc. IEEE Int. Symp. Inform. Theory, Washington, DC, p. 294 (June 2001)
5. Hunter, T.E., Nostratinia, A.: Cooperation diversity through coding. In: Proc. IEEE Int. Symp. Inform. Theory, Laussane, Switzerland, p. 220 (June 2002)
6. Wu, M., Xue, W., Wubben, D., Dekorsy, A., Paul, S.: Energy-aware design of inter-relay cooperation for distributed relaying networks. In: 8th International Symposium on Wireless Communication Systems (ISWCS), pp. 864–868 (November 2011)
7. Nakagami, M.: The m-distribution -A general formula of intensity distribution of rapid fading. In: Hoffman, W.G. (ed.) Statistical Methods in Radio Wave Propagation. Pergamon, Oxford (1960)
8. Goldsmith, A.: Wireless Communications, 1st edn. Cambridge University Press, Cambridge (2005)
9. Bansal, A., Garg, P.: An Analytical Approach to Outage Analysis and Critical Cooperation Ratio in Coded Cooperation. In: Wireless Personal Communications. Springer (April 2012), doi:10.1007/s11277-012-0608-4
10. Hagenauer, J.: Rate-compatible punctured convolutional codes(RCPC codes) and their applications. IEEE Trans. Wireless commun. 36(4), 389–400 (1988)
11. Bell, W.W.: Special functions for scientists and engineers. Van Nostrand, London (1968)
12. Gradshetyn, I.S., Ryzhic, I.M.: Table of Integrals, Series and Products, 6th edn. Academic Press, San Diego (2000)

Appendix

The terms in (4), containing the probability $Pr\{\alpha P_{sd}+(1-\alpha)\sum_{i\in\{\Theta_1\}}P_{xy} < \mathcal{P}\}$ where $\{\ l \in 1,2\ \}$ are number of relays cooperating, and P_{xy} is the power over xy link, is given as

$$Pr\{\alpha P_{sd} + (1-\alpha) \sum_{\substack{i=1 \\ \forall l = \{1,2\}}}^{l} P_{xy} < \mathcal{P}\} = \int_0^{\frac{\mathcal{P}}{\alpha}} \int_0^{a} f(P_{sd}) f(z) dP_{sd} dz \tag{7}$$

where $z = \sum_{i=1}^{l} P_{xy} \sim \text{Gamma}\left[lm, \frac{\tilde{P}_{av_1}}{m}\right]$, and $a = \frac{\mathcal{P}-\alpha P_{sd}}{1-\alpha}$. Using (2) and (3), the integral $\int_0^a f(z)dz$ can be written as

$$\int_0^{a} f(z)dz = 1 - \frac{\Gamma\left(lm, \frac{ma}{\tilde{P}_{av_1}}\right)}{\Gamma(lm)}\,, \tag{8}$$

therefore

$$Pr(\alpha P_{sd} + (1-\alpha)\sum_{i=1}^{n} P_{xy} < \mathcal{P}) = \int_0^{\frac{\mathcal{P}}{\alpha}} f(P_{sd}) \left[1 - \frac{\Gamma\left(lm, \frac{ma}{\tilde{P}_{av_1}}\right)}{\Gamma(lm)}\right] dP_{sd} \tag{9}$$

Further, using [12, Eq. (8.352.2)], $\Gamma\left(lm, \frac{ma}{\tilde{P}_{av_1}}\right)$ can be written as

$$\Gamma\left(lm, \frac{ma}{\tilde{P}_{av_1}}\right) = (lm-1)! \exp(-\frac{ma}{\tilde{P}_{av_1}}) \sum_{k=0}^{lm-1} \frac{1}{k!} \left(\frac{ma}{\tilde{P}_{av_1}}\right)^k, \tag{10}$$

we get,

$$\begin{aligned} &Pr(\alpha P_{sd} + (1-\alpha)\sum_{i=1}^{n} P_{xy} < \mathcal{P}) = \\ &A_1 - A_2 \sum_{k=0}^{lm-1} A_3 \int_0^{\frac{\mathcal{P}}{\alpha}} \left(\frac{\mathcal{P}}{\alpha} - P_{sd}\right)^k {P_{sd}}^{m-1} \exp(BP_{sd}) dP_{sd} \\ &= A_1 - A_2 \sum_{k=0}^{lm-1} A_3 C \beta(k+1, m)\, {}_1F_1\left[m; k+m+1; \frac{B\mathcal{P}}{\alpha}\right]. \end{aligned} \tag{11}$$

This simplification is achieved using [12, Eq. (3.383)]. Here, terms A_1, A_2, A_3, B and C can be written as

$$\begin{aligned}
A_1 &= \left[1 - \frac{\Gamma\left(m, \frac{m\mathcal{P}/\alpha}{\tilde{P}_{sd}}\right)}{\Gamma(m)}\right], \\
A_2 &= \left[\frac{1}{\Gamma(m)}\left(\frac{m}{\tilde{P}_{sd}}\right)^m \exp\left(-\frac{m\mathcal{P}}{(1-\alpha)\tilde{P}_{av_1}}\right)\right], \\
A_3 &= \left[\frac{1}{k!}\left(\frac{m\alpha}{(1-\alpha)\tilde{P}_{av_1}}\right)^k\right], \\
B &= \left[\frac{m\alpha}{(1-\alpha)\tilde{P}_{av_1}} - \frac{m}{\tilde{P}_{sd}}\right], \\
C &= \left[\frac{\mathcal{P}}{\alpha}\right]^{k+m}, \qquad (12)
\end{aligned}$$

where $\beta(u, v)$ is beta function defined as $\beta(u, v) = \int_0^1 x^{u-1} 1 - x^{v-1} dx$, and ${}_1F_1[a; b; z]$ is the confluent hypergeometric function [11].

Power Efficient MAC Protocol for Mobile Ad Hoc Networks

Sohan Kumar Yadav and D.K. Lobiyal

School of Computer and Systems Sciences, Jawaharlal Nehru University,
New Delhi-110067, India
{jsmsohan,lobiyal}@gmail.com

Abstract. Maximizing throughput and battery power are the key factors to design an efficient power control MAC protocol. Power management in MANET is a critical issue, since these are powered by batteries. It is also due to mobility of MANETs, the size of batteries is of great concerns. This paper presents a novel power control protocol, namely Power Efficient MAC protocol for mobile ad hoc networks. The PEMAC protocol uses minimum power level to transmit RTS instead of maximum power level. Here the minimum power level is predefined value, and it is noted that it is sufficient to reach the receiver. The receiver transmits CTS by using maximum power level. The data and acknowledgement are transmitted using power level respectively which is calculated according the power level of RTS transmission. This protocol conserves energy as it uses less energy in transmitting RTS packet, and it also increases the spatial reuse in the network. It has been shown through simulation that the proposed protocol is energy efficient without degrading throughput.

Keywords : IEEE 802.11, MAC, MANETs, Power control.

1 Introduction

Mobile Ad Hoc Networks (MANETs) is a multi-hop in nature where, mobile nodes are operated in a distributed manner without any fixed infrastructure. MANETs are gaining popularity due to their ability to provide temporary and instant networking solution in situations i.e. military, hazardous, flood, natural calamity. In this network nodes are operated only by battery power, therefore energy saving is most important to maximize the lifetime of network. This can be implemented into two ways are: (i) allow nodes to go in sleep mode and (ii) use of power control schemes for transmission of packets. Using different transmission power level for DATA packet transmission causes the decrease of the networks capacity due to exposed terminal problem.

The Ad Hoc mode of the IEEE 802.11 standard is the most popular MAC protocol for Ad Hoc networks. This protocol generally follows the CSMA/CA (Collision Avoidance) and the exchange of RTS/CTS between the transmitter and the receiver. In this method a transmission floor is reserved for the data packet transmission. This protocol uses maximum power level for transmitting data transmission to prevent other nodes transmission present in its carrier sensing range. Nodes that are hearing

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 207–217, 2013.

RTS/CTS message will defer their transmission until ongoing transmission. This scheme is useful in solving the problem of hidden and exposed terminal problem.

This scheme does not allow concurrent transmission of nodes present in the carrier sensing range of the ongoing transmitting node. We can say that in this scheme simultaneous transmission is not possible i.e. it degrade the networks spatial reuse capacity. For example, we consider the situation as in figure 1, here nodes A and B are trying to established communication between them. Node A transmitting RTS packet with maximum power level to node B. Node C hear node A's RTS message (it is present in the carrier sensing range of node A) and therefore, it postpone its transmission to the node D [1, 3,5, 8].

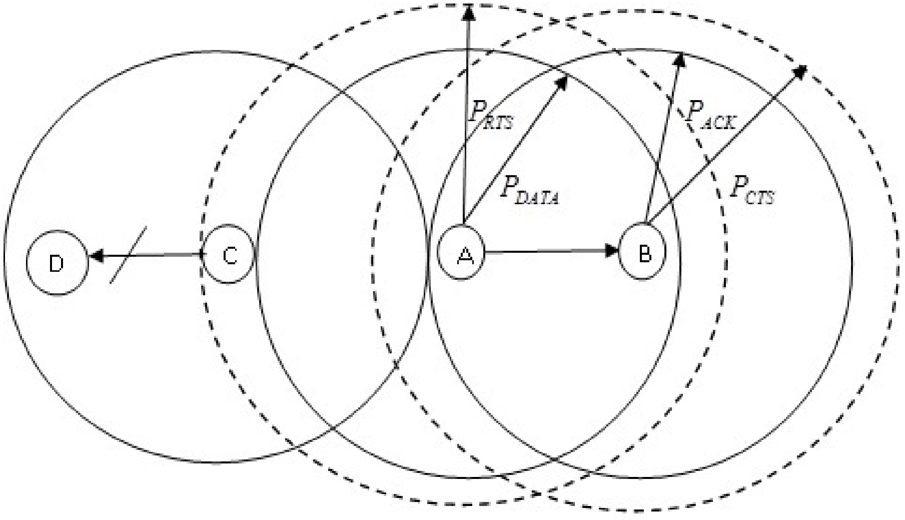


Fig. 1. Simultaneous Transmission of nodes A and C are not possible if it uses $P_{\max}$ for transmitting RTS, CTS packets

To improve the efficiency of the IEEE 802.11 protocol, we propose a power efficient MAC protocol for mobile ad hoc networks that uses minimum power level for RTS packet and maximum power level for CTS packets transmission. The DATA/ACK packets transmission power level will be calculated after the completion of RTS/CTS handshake. And it will be according to the power level for transmitting RTS which will complete the RTS/CTS handshake. The rest of the paper is organized as follows. In section 2 we reviewed related work. Section 3 will present the proposed protocol. Simulation results are given in Section 4. Finally, conclusions are given in Section 5.

2 Related Work

Power Control for MANETs has been studied in the literature. Therefore, the goal for MAC protocol for MANETs is to coordinate the channel access among the number of nodes to achieve high channel use. The IEEE 802.11 deals with the both physical and medium access control layers of network. Most widely used MAC protocol is IEEE 802.11, and MAC layer is the host for set of protocols. It is responsible for the regulating the use of shared medium. In this paper the study will focus on the IEEE 802.11 MAC protocol and its power control schemes. We will give detailed discussions about the scheme in the following text.

2.1 IEEE 802.11 MAC Protocol

It is a widely used protocol, in which RTS/CTS handshake will be done using maximum transmission power level [2, 10]. In this protocol DATA/ACK will also transmitted, using maximum transmission power level. Headers of RTS, CTS, and DATA include the time duration to inform the nodes which are in carrier sensing range of the ongoing transmitting when ACK will be sending. In this scheme, nodes present in carrier sensing range of the sender and receiver are capable of decoding RTS, CTS, and DATA, knowing about that the ongoing transmission. This scheme avoids the very first stage collision. It prevents the hidden terminal problem. For example, in figure 2, if node A wants to transmit to node B, it first send RTS packet. After receiving RTS node B will reply with CTS packet which include the total time period for ongoing transmission. Since node C within the carrier sensing range of node B, clearly it decodes and extracts information about the neighbor ongoing transmission.

Therefore, even if node C wants to transmit to node D, it will keep silent till ongoing transmission of nodes A and B. It is clear that hidden terminal problem is solved using RTS/CTS mechanism.

Clearly, IEEE 802.11 MAC protocol solved the most common problem such as hidden terminal problem in MANETs. However, there is no consideration of power control in IEEE 802.11 MAC protocol. It uses maximum and same transmitting power level for all type of packet transmissions, which leads to more battery consumption.

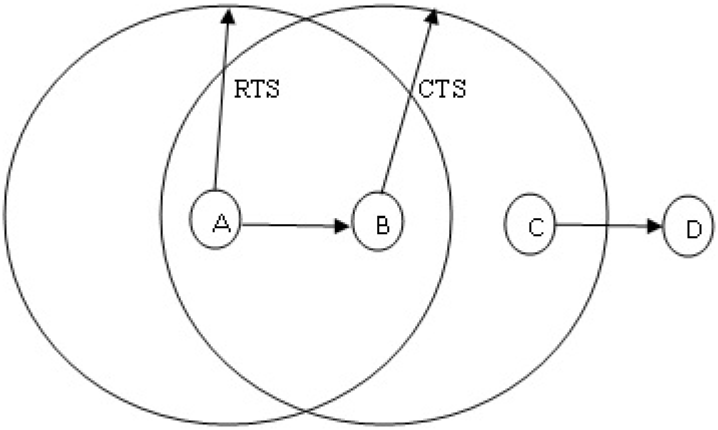


Fig. 2. RTS/CTS handshaking (prevent hidden terminal problem in IEEE 802.11) [5]

2.2 BASIC Power Control MAC Protocol

The basic power control protocol (BPCMP) uses different power level for handshake of RTS/CTS packets and for DAT/ACK packets [4]. In this scheme RTS/CTS will be sends with maximum power level. Therefore, all the nodes presents in the neighbor of the transmitting node, know about ongoing transmission. And DATA/ACK packets sent by using minimum power level. Let $P_{DESIRED}$be the power level for transmitting DATA/ACK, and is given by

$$P_{DESIRED} = \frac{P_{MAX}}{P_r} \times R_{XTHRESH} \times \beta \tag{1}$$

Where $R_{XTHRESH}$ is the minimum signal strength necessary to receive the signal, which is determined by the physical characteristics of the node, β is a constant, P_r the amount of power level received at the receiver side, when it will be sends by maximum power level by the sender. From the figure 3, it is clear that the basic power control MAC protocol sends RTS/CTS using maximum power level (P_{MAX}), and DATA/ACK using lowest possible power level ($P_{DESIRED}$).

This protocol may introduce more collisions. Therefore, using this protocol increases the number of retransmission to achieve better throughput as compare to IEEE 802.11. This scheme gives better network performance at the cost of more energy consumption. Therefore, this scheme is not suitable for us because we are interested in finding higher throughput with minimum energy consumption. Due to mobility for nodes the low power level transmission also causes more retransmission. This leads to higher energy consumption

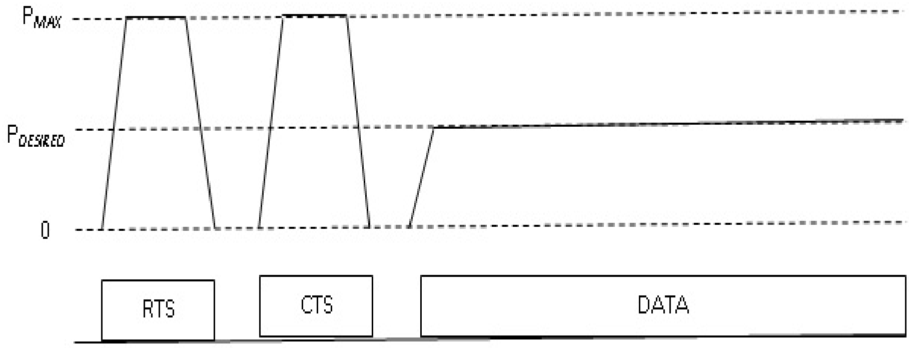


Fig. 3. BPCMP uses different power level for RTS/CTS and DATA

2.3 The PCM Protocol

As we discussed earlier, BASIC scheme consumes more energy and also degrade the networks throughput. Authors [1], proposes an improved power control MAC protocol for MANETs. It is similar to the BASIC power control protocol. This PCM protocol can avoid the collision by using maximum power level for transmitting DATA periodically. The time for periodically transmission should be less than the EIFS (extended inter-frame space) duration to let other nodes to know ongoing transmission of its neighboring node.

This scheme maintains the carrier sensing area periodically and also using less power level as in BASIC protocol. However, if nodes are mobile, the low power level causes more retransmission. Thus it consumes higher energy consumption [1, 6,9,11]. Therefore, this scheme is not acceptable when nodes are mobile, and network is dense.

3 Proposed Protocol

In this section, we proposed the power efficient MAC protocol for mobile ad hoc networks. This protocol can be considered as an improved version of BASIC protocol. In the proposed protocol different power level are used to transmit RTS and CTS packets. Let us take situation as in figure 4. From figure 4 it is clear that if we use minimum power level for transmitting RTS packet instead of maximum power level, simultaneous transmission of node group A-B and C-D can be possible. Thus, this increases the spatial reuse, and the network capacity. Therefore to increase the spatial use of network and also to increase the life of battery we propose a power control MAC protocol for mobile ad hoc networks. This protocol will able to solve above discussed problem.

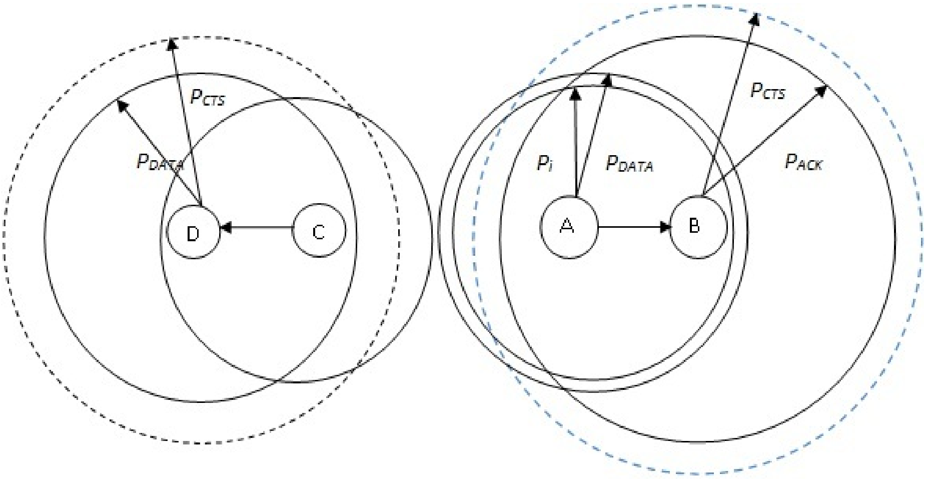


Fig. 4. Spatial reuse in PEMAC (Here nodes A and C can communicate simultaneously)

The proposed protocol uses the following steps to transmit the data to the intended receiver.

I. Initially, sender node sends RTS frame with minimum power level i.e. P^{i}_{RTS}. It is predefined value and is sufficient enough to reach to the receiver node. If receiver node receives RTS, it replies with CTS to the sender node. Otherwise the value of P^{i}_{RTS} will be increased, since it not enough to reach the receiver. The value of this will be increased as follows

$$P^{i+1}_{RTS} = P^{i}_{RTS} + \frac{P_{MAX} - P^{i}_{RTS}}{k \times i} \tag{2}$$

Where k is system parameter and is set to be 3 for better network performance.

II. CTS frame is sent by using maximum power level i.e. P_{MAX}. Nodes in carrier sensing range of receiver are sensing it and defer their transmission for period of ongoing transmission.

III. After RTS/CTS handshake is over, the DATA/ACK will be sent by using power level defined as follows

$$P_{DATA/ACK} = P^{i}_{RTS} \times \varepsilon \tag{3}$$

Where P^{i}_{RTS} is the power level of RTS transmission when receiver responds with CTS packet, and ε is a constant and set to be greater than 1.

IV. When the retransmission occur, transmission power level is increased as P_{MAX} instead of P^{i}_{RTS}.

In BASIC and PCM protocol, the power level for transmitting data packets is not changed whenever retransmission occurs. This may create a problem of network failure. The possibility of network failure will increase when nodes are mobile. Therefore, it will increase the number of retransmission of the packets. Most of the retransmissions are due to insufficiency of power (nodes are mobile). Therefore,

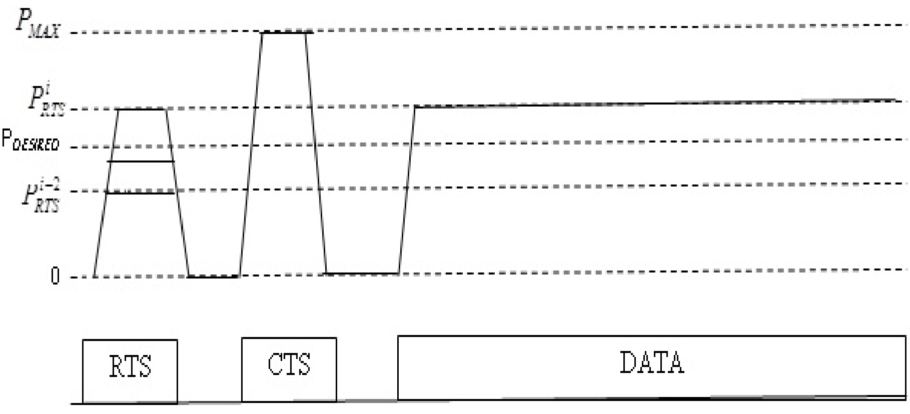


Fig. 5. Different power level will be used in PEMAC protocol for RTS, CTS, and DATA

proposed PEMAC protocol solves this problem by using the maximum power level for retransmission. This will also solve somehow the problem of mobility. The different power level used in sending RTS and CTS packets are shown in the figure 5. This also uses different power level for DATA packets.

4 Performance Evaluation

In this section, we evaluate the performance of the proposed PEMAC protocol using the computer simulator GloMoSim [7]. Here we also compared the performance of the PEMAC protocol with the other power control MAC protocol i.e. PCM, BASIC, and IEEE 802.11.

4.1 Simulation Environment

We have considered varying node density i.e. 3, 6, 9, 12, and 15 with and without mobility. Further the other parameters used in simulation are listed in table 1. The carrier sensing range is twice the transmission range (approx.).

Table 1. Parameters used in simulation

Simulation Time	15 second
Network Area	500×500
Radio Transmission Range	250m
Radio Carrier Sensing Range	500m(approx)
Propagation Model	Two-ray path loss
Bandwidth	2Mbps
Node Placement	Random
Mobility-Wp-pause	0.1 Millisecond
Mobility-Wp-Min-Speed	0 m/s
Mobility-Wp-Max-Speed	10 m/s
Noise Figure	4
Radio-Tx-Power	24.5 dBm
Radio-Antenna-Gain	0.0 dBm
Radio-Rx-Sensitivity	-71.42 dBm
Routing	LAR1
Promiscuous-Mode	No
Mobility Model	Random way point

4.2 Simulation Results

We used following metrics to evaluate the performance of the PEMAC, PCM, BASIC, and IEEE 802.11 protocols under both the mobile and static environment. In mobile condition the maximum speed is considered 15 m/s.

Throughput: It is the total number of data bits transferred from source to destination in per unit time i.e. per second. This shows the performance in terms of network capacity.

Rate of Energy Efficiency: It is the number of data bits delivered per joule energy consumed. Higher the rate of energy efficiency means protocol more efficient. It is measured in bits/joules.

Rate of Energy Efficiency = total number of data bits / total energy consumed in joule

Figure 6 shows, throughput in bits per second when nodes are not mobile. In this we have taken 3, 6, 9, 12, and 15 nodes for evaluation of the scenario. From the figure 6 it is clear that our protocol performs better than the other.

In Figure 7, we compared network throughput when nodes are mobile with maximum moving speed 10 m/s. Here, we are using same parameters as used in the above case. Figure 7, shows that proposed protocol is performing better as compared to PCM, BASIC, IEEE 802.11.

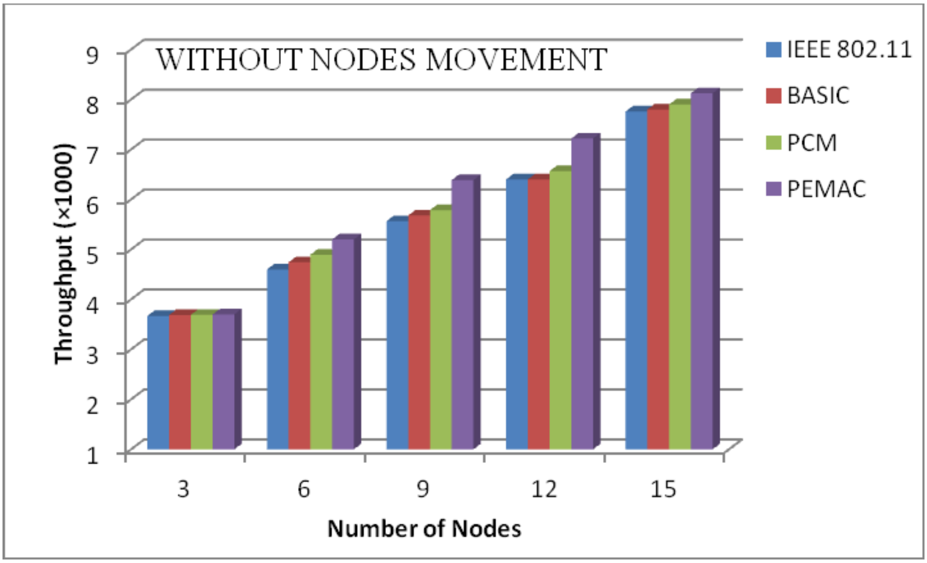


Fig. 6. Comparison of Throughput (moving speed = 0)

Figure 8, shows that ratio of energy efficiency as nodes are static. In this situation we take 3, 6, 9, 12, and 15 nodes. It shows that PEMAC performs better in comparison of PCM, BASIC, and IEEE 802.11.

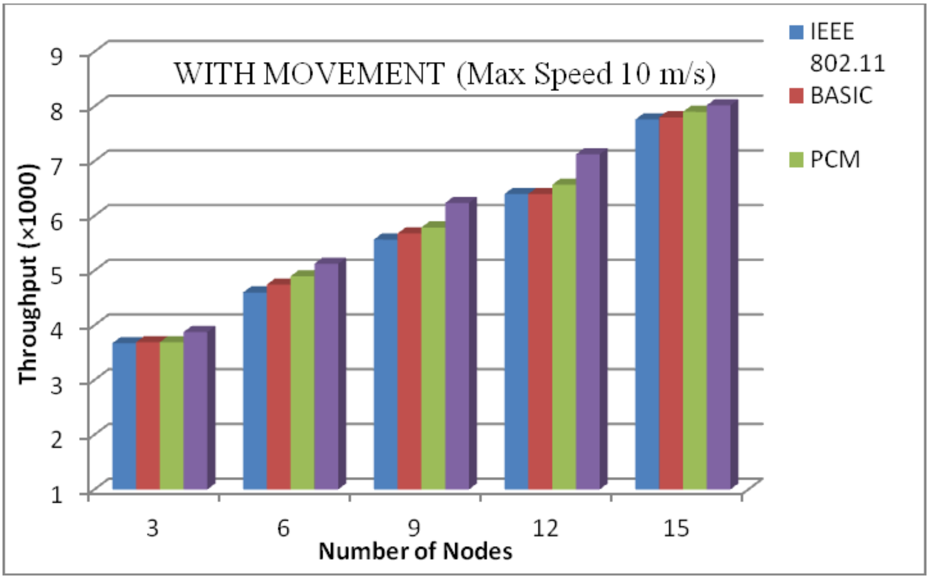


Fig. 7. Comparison of Throughput (moving speed = 10)

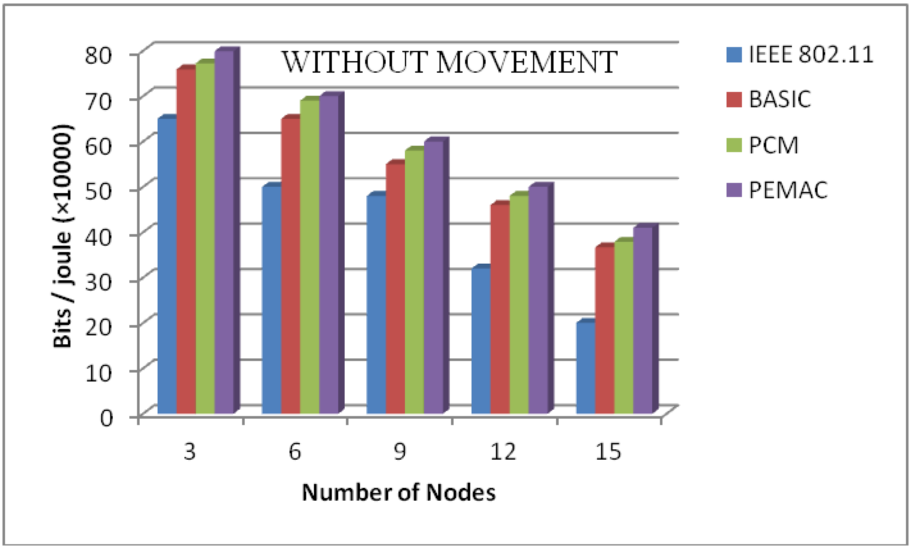


Fig. 8. Compariosn of energy efficiency (moving speed = 0)

Figure 9 shows that the comparison of energy of network in mobile state (moving speed = 10 m/s). And other parameters are same as static situation. Form the figure 9 it is clear that proposed protocol works better with the existing protocol.

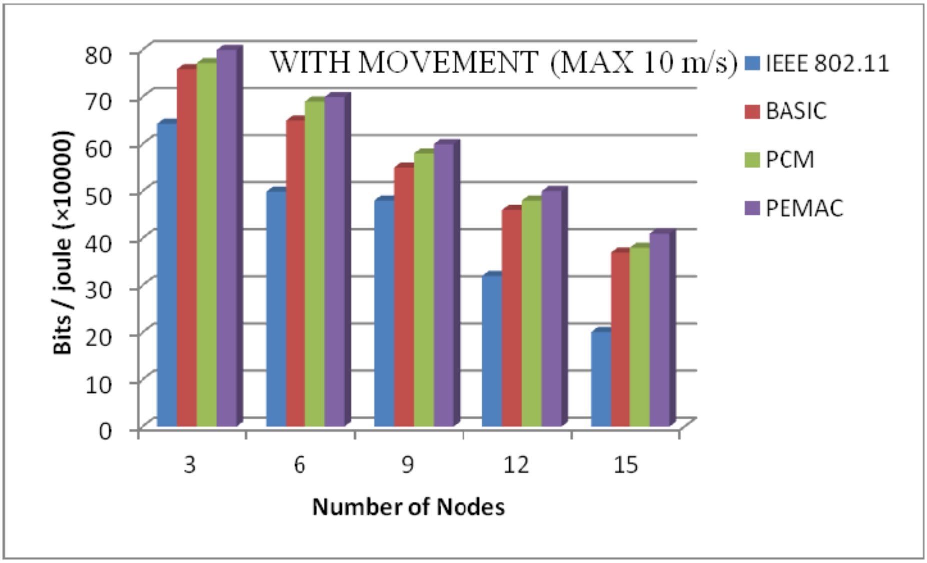


Fig. 9. Compariosn of energy efficiency (moving speed = 10)

5 Conclusion

We presented a new power control MAC protocol (named, PEMAC) for MANETs. Introduction of IEEE 802.11 and BASIC protocol forms the basis to propose this protocol, which gives better performance in terms of battery life and data delivery. The proposed PEMAC protocol uses minimum power level to transmit RTS packet and uses maximum power level for CTS. Hence it calculates the power level for DATA/ACK according to the RTS transmitting power level. It has been observed from the simulation results that PEMAC protocol conserved energy without degrading network throughput as generally other protocol degrading network performance while going for power control.

Acknowledgement. The work presented in this paper is supported by Council of Scientific & Industrial Research (Human Resource Development Group), New Delhi, India.

References

1. Jung, E.-S., Vaidya, N.H.: A Power Control MAC Protocol for Ad Hoc Networks. In: Proceeding of the International Conference on Mobile Computing and Networking (MOBICOM 2002), pp. 36–47 (September 2002)
2. Muqattash, A., Krunz, M.M.: A Distributed Transmission Power Control Protocol for Mobile Ad Hoc Networks. IEEE Transactions on Mobile Computing 3(2), 113–128 (2004)

3. Murthy, C.S.R., Manoj, B.S.: Ad Hoc Networks: Architectures and Protocols, 2nd edn. Pearson Education (2005)
4. Chen, H.H., Fan, Z., Li, J.: Autonomous Power Control for MAC Protocol for Mobile Ad Hoc Networks. Hindwani Publishing Corporation, EURUSIP Journals on Wireless Communication and Networking 36040, 1–10 (2006)
5. Yan, H., Li, J., Sun, G., Chen, H.-H.: An Optimistic Power Control MAC Protocol for Mobile Ad Hoc Networks. In: IEE ICC Proceedings, pp. 3615–3620 (2006)
6. Varvariyos, E.M., Vasileios, G., Nikolaos, K.: The Slow Start Power Control MAC Protocol for Mobile Ad Hoc Networks and its Performance Analysis. Ad Hoc Networks 7, 1136–1149 (2009)
7. Nuevo, J.: Comprehensible GloMoSim Tutorial. INRS-Universite du Quebec, Nuevo@inrs-telecom.uquebec.ca, March 26, pp. 1–34 (2004)
8. Liu, C.-Y., Lin, C.-H.: Distributed MAC Protocol to Improve Energy and Channel Efficiency in MANET. In: Langendoerfer, P., Liu, M., Matta, I., Tsaoussidis, V. (eds.) WWIC 2004. LNCS, vol. 2957, pp. 1–12. Springer, Heidelberg (2004)
9. Chang, C.Y., Chang, H.R.: Power Control and fairness MAC mechanisms for 802.11 WLWNs. Computer Communications 30, 1527–1537 (2007)
10. Krunz, M., Muqattash, A.: Transmission Power Control in Wireless Ad Hoc Networks: Challenges, Solutions, and Open Issues. IEEE Network 18(5), 8–14 (2004)
11. Gupta, P., Kumar, P.R.: The Capacity of Wireless Networks. IEEE Transaction on Information Theory 46(2), 388–404 (2000)

Fault Tolerant Range Grouping Routing in Dynamic Networks

Akanksha Bhardwaj, Prachi Badera, and K. Rajalakshmi

Department of Computer Science,
Jaypee Institute of Information Technology, Noida
{busy.akanksha,prachbadera}@gmail.com,
k.rajlakshmi@jiit.ac.in

Abstract. A characteristic feature of dynamic networks is the notion of failure. A failure can be a partial failure [as in distributed systems] or total failure. A partial failure may happen when one component in a system fails. This failure may affect the proper operation of other components, while at the same time leaving yet other components totally unaffected. In contrast, a failure in any system is often total in the sense that it affects all components, and may easily bring down the entire system. Hence, it becomes important to design a system which can work even if (partial) failures occur. This paper proposes various approaches which help in making the designed system fault tolerant. Mainly the routing mechanism is focused upon with the help of the concept of acknowledgment and negative acknowledgment.

Keywords: Routing, Crash failure, Omission failure, Fault detection, Fault recovery.

1 Introduction

Dynamic networks have received significant attention in recent years due to the widely increasing application of Dynamic Networks in various fields. The nodes in such a network collaborate with each other to perform tasks such as data communication, data processing, data analyzes, etc. Since the network formed is dynamic in nature, its components have high probability to fail under given scenario [1].

Fault tolerance is the ability of a system to deliver a desired level of functionality in the presence of faults. Actually, extensive work has been done on fault tolerance and it has been one of the most important topics in Dynamic Networks. The objective of this research paper is to investigate the current research work on fault tolerance in Dynamic Networks.

Failures determined by nature are the identity of the faulty processes and the details of their faulty behavior. These depend on the particular failure model being assumed. In this application, we consider two closely related failure models, called crash and omission. Here crash implies departure of the node from the topology. In the omission model, a faulty process may omit to send/receive messages [2].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 218–227, 2013.

Another problem faced in dynamic network is arrival of a new node. In a node arrival /departure [crash model], it is necessary to provide the user with a new topology.

A brief overview of the network designed is given in section 2. In section 3 how fault happens in different levels of Dynamic network is discussed. Fault detection and recovery algorithms are introduced in section 3.4.1. Results of evaluation performed for this approach are shown in section 4.

2 Network Designed

In order to create large networks, involvement of large nodes is required. However, maintenance of large number of nodes is not easy. Therefore, for proper maintenance of large nodes, we first cluster the nodes and then perform routing among the nodes. A network with 9 nodes has been created in our previous work is described in detailed manner in [3]. A short description of the network is as follows:

2.1 Clustering

Let node A be the master and suppose it first runs the application to detect its nearby devices. Node A finds out nodes B, C, D in its cluster and store the address and the name in their data structures. Hence it forms an INTRA cluster with the nodes B, C, D. Similarly, nodes B, C, D, E, F, G, H, I will find the nodes lying in its range and store this information in their lists respectively. Next, these lists are exchanged with the nearby nodes. At each node, a comparison between the nodes of its list and the list received from its neighbors takes place. A node that is found to be different is appended in its list. These newly added nodes form the INTER cluster where the nodes of inter cluster becomes gateway to reach the nodes of inter cluster. This creates an initial topology of the network. The topology is prone to changes [since it is a dynamic Network], therefore addition and deletion of node takes place according to their Bluetooth ranges.

2.2 Routing

After the formation of the clusters, the main task that remains is forwarding of the data. In order to accomplish this, the nodes have the list available with itself which it can refer to know if the node exists in the network or not. After mentioning the destination node, the node looks at the list. The data are sent to the neighboring node via *hoping*. When the data reach the ***gateway*** node (i.e. the node which is a part of more than one cluster), it looks for the route with the help of the lists again and then again routes the data. If the gateway crashes it looks for another path and sends the packets via another gateway.

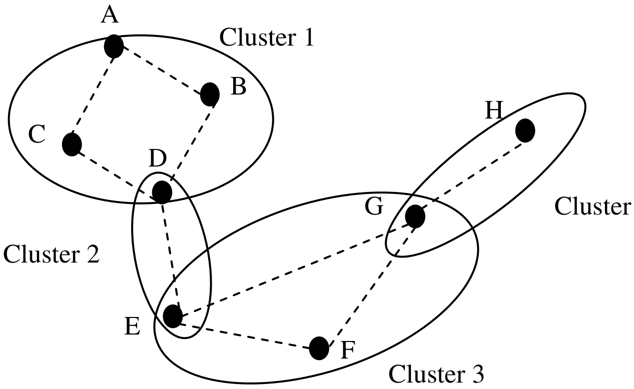


Fig. 1. Four clusters are formed covering all the nodes in the network

3 Fault Tolerance

Five levels of fault tolerance were discussed in [5]. They are physical layer, hardware layer, system software layer, middleware layer, and application layer. On the basis of study, we classify fault tolerance in WSNs into four levels from the system point of view.

To tackle faults the system should follow two main steps. The first step is fault detection. It is to detect that a specific functionality is faulty, and to predict it will continue to function properly in the near future. After the system detects a fault, fault recovery is the second step to enable the system to recover from the faults.

We consider two closely related failure models, called crash and omission. Here crash implies departure of the node from the topology. In the omission model, a faulty process may omit to send/receive messages.

3.1 Assumptions

- Synchronous system
- Communication delay is bounded.
- Message delivery is ordered.
- Uni-casting.

3.2 Hardware Layer

Faults at hardware layer can be caused by malfunctioning of any hardware components. As in the network designed, the hardware component used is Bluetooth device. Due to power limitations of Bluetooth devices, excessive use of a Bluetooth device can result in its malfunctioning. Secondly, due to the environment which may contain other radio wave radiations which might result in interfere with Bluetooth's working and hence create problem for Bluetooth users.

Solution. Try to use the Bluetooth in the environment where there are no other radiations. And for power consumption, always put the device on charging while using it.

3.3 Software Layer

The network has two software components:

a. Operating system
b. Middleware such as routing.

Solution. In order to reduce the probability of operating system failure, the designed application should be tested with different versions so that even if one results in a failure, the application can still run in other version.

For fault tolerant routing, the concept of acknowledgement and negative acknowledgement is introduced. The detailed algorithm is explained in the following sections.

3.4 Network Communication Layer

The routing in Dynamic Networks is the most prone area to faults. There may be a possibility that the receiver doesn't receive the sent packets due to link failure. Or there may be a possibility that the sender is not able to send the packets to the desired node.

As described in the related research, the fault tolerance can be implemented either with the help of check-pointing or with the help of message logging.

3.3.1 Proposed Approach

In our approach, we combine the two concepts for fault tolerant routing. When the data to be forwarded is packetized, it is piggy-backed with a randomly generated number called the transaction ID and the total number of packets to be sent. All the generated packets are assigned a serial number so that the receiver can keep a track of lost messages.

Thus, the packet is shown in fig(2)

Packet Type	File format	Data	Transaction Id	Seq no.	Total no. of packets	Receiver Name

Fig. 2. Packet formed before forwarding the data

A sender log is maintained which records the filename and the receiver's address.

Filename	Receiver's [Bluetooth] address

Fig. 3. Sender Log

On sending a message

1) A file/data is broken into small chunks of data.
2) Each chunk is then packetized as shown in fig (2).
3) An entry is made into sender log as shown in fig (3).
4) The packet is sent to the receiver.

3.3.1.1 Omission Failure

Omission failure[4] occurs when a message inserted in sender hosts message buffer never arrives at the receiver hosts incoming message buffer.

3.3.1.1.1 Fault Detection

For detecting omission failure a receiver log is maintained which stores the transaction ID, the number of packets received, the total number of packets it should receive and the sender's address and the filename.

Transaction ID	Total no. of packets	No.of packets received	Filename	Sender's address	Time at which packet was received

Fig. 4. Reciever Log

Negative Ackowledgement	Transaction Id	Sender Name	Receiver Name

Fig. 5. Negative Acknowledgement

Acknowledgement	Transaction Id	Sender Name

Fig. 6. Acknowledgement

Transaction ID	File Name

Fig. 7. Sender Log

On receiving a packet:

1) If the receiver received a packet with sequence number as 1, A new entry is made in the receiver log as shown in fig (4).

2) On receiving each packet with sequence number greater than 1, based on transaction ID its sequence number is compared with the sequence number of the received packet.

3) If the sequence number is consistent the packet is processed and number of received packets is updated.

4) If the sequence number is not consistent the packet is rejected and a negative acknowledgement (NACK) is created as shown in fig (5).

The receiver log is monitored periodically for incomplete transactions. A transaction is said to be incomplete if its last packet was received 30 sec before the current time. On detecting an incomplete transition a NACK is again generated and sent.

An acknowledgement is sent to the sender, as shown in fig (6), after receiving the last packet. On receiving the acknowledgement entry in sender log is deleted. After sending the acknowledgement the receiver deletes the entry of file from the receiver log.

3.3.1.2 Fault Recovery

On receiving a NACK:

1) Checks the entry in sender log based on transaction Id to get the filename.
2) The file is broken into chunks.
3) Chunks following the sequence number received are packetized.
4) An entry is made into sender log as shown in fig (7).
5) The packet is sent to the receiver.

3.3.2 Crash Fault/Entry of New Nodes

Crash fault occurs when a node due to mobility gets disconnected from the topology. Similarly a new node can also enter the existing topology. All such updations should be diagnosed and communicated.

3.3.2.1 Fault Detection

The receiver checks his receiver log periodically. A node arrival and departure is detected.

3.3.2.2 Fault Recovery

The notification of the arrival and departure of nodes is broadcasted in the topology.

4 Evaluation

Logs are maintained at both sender and receiver end which helps in fault detection and recovery. But these logs can be a load on memory. Hence after complete transfer of packets the entry from the log is deleted. In this way all transactions that are incomplete are stored and rest are deleted to decrease the load on memory.

The most important factor to be dealt with in any routing algorithm is Reliability. The problem faced in most of the applications based on dynamic networks is the problem of failure. Especially for a network type application, failure can occur easily due to change in available bandwidth, failure of links, nodes. Hence the goal here is to make the system work properly under any circumstances either by fixing the problem or to make the system work by neglecting the faults without its performance being affected.

Before the integration of fault tolerance module to the routing algorithm when there is a low link failure rate, as shown in case (a), the packets are dropped in between and when the link sets up again, the dropped packets are not transferred. However, after application of fault tolerance module in the application, the dropped packets are asked to be sent first and then the latter packets are accepted. Hence, after the integration of fault tolerance, the channels become more reliable.

In case (b), when there is a high link failure rate, more packets are dropped i.e. complete information is not transferred from sender to receiver and large number of packets are dropped. These simulations have been done in ns2.The simulation results for case (a) and case (b) are shown in fig (8) and fig (9) respectively.

a. Low Link Failure Rate

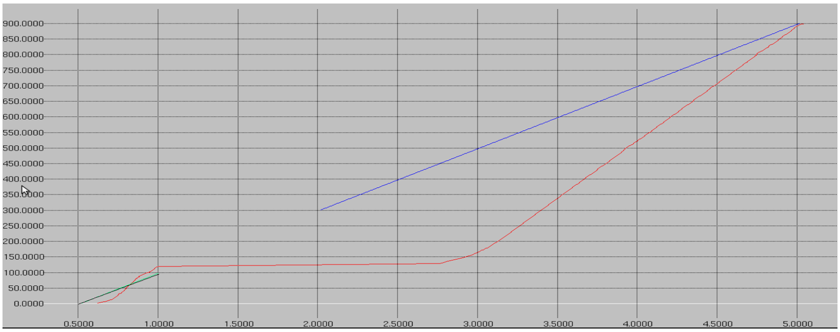


Fig. 8. Case with Low Link failure Rate showing before and after the implementation of the module

b. High Link Failure Rate

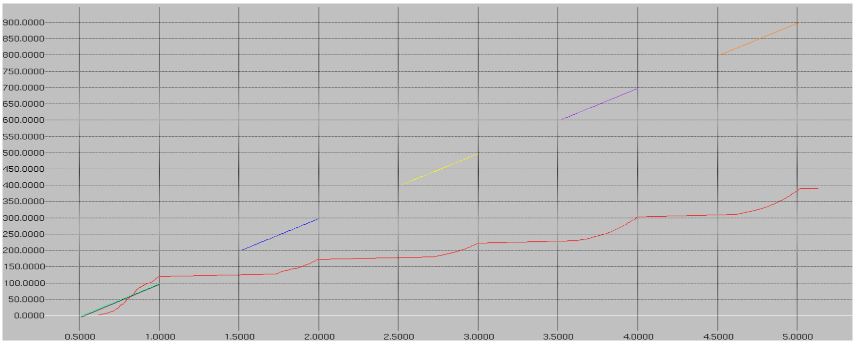


Fig. 9. Case Showing High link failure rate before and after implementation of the module

For any network based application, the transmission time and response time are the most important characteristics.

In this application, for transmission of the data, the data is first packetized in the application and then forwarded. Computation time with respect to this application is defined as the time it takes to form the packet initially. Delay is the time introduced due to the hoping of the data which includes comparing the address with its own address and correspondingly either forward or keep it with itself or delay introduced due to the link failure between nodes. The response time per packet is defined as the sum of delay introduced and transmission time. And the total response time i.e. the total time taken to transfer the file via hoping or directly is derived by multiplying the number of packets with the response time per packet.

The following table depicts different scenarios with different data input files with respect to total time taken in the network.

Table 1. Response time with respect to Link Failure Rate in different scenarios. n–Number of hops done to reach the receiver.

Link Failure Rate	**No. of hops**	**Computation Time at sender (per packet) (sec:milisec)**	**Transmission Time (per packet) (sec:milisec)**	**Delay (per packet) (sec:milisec)**	**Response Time (Delay + Transmission time + Computation time) (per packet) (sec:milisec)**
0	0	01:22	00:03	N/A	01:22
0	1	01:22	00:03	00:32	01:22 + 00:35 =01:57
0	N	01:22	00:03	00:32*n	01:22+00:32*n +00:03 = 01:25+00:32*n
1	0	01:22	00:03	00:07 + 00:03[#]+ 01:30 = 01:40	01:22+01:40 +00:03 = 03:05
1	1	01:22	00:03	00:07+ 00:03+ 01:30+ 00:03 = 01:43	01:22+01:43 +00:03 = 03:08
1	N	01:22	00:03	00:07 + (01:30*n-1)+ 00:03*n+ 1 =03:33*n –00:93	01:22+00:03 +03:33*n-00:93 = 2:12+3:33*n

[#] Assuming the link breaks after transferring one packet only.

5 Related Research

Faults in any system can be classified in three categories: Transient, Intermittent and Permanent faults. Transient faults are the ones which occur once and then disappear. An intermittent fault occurs, vanishes of its own accord, then reappears, and so on. Intermittent faults cause a great deal of aggravation because they are difficult to diagnose. A permanent fault is one that continues to exist until the faulty component is replaced [2].

Since, in a large network, the machines [servers/clients] are highly dependent on each other, failure in one system can lead to failure in others and so on. Therefore, it is a necessity to define failure models. Gottfried Fuchs has described various failure models in his paper [10].

MNLIR scheme proposed in [6] uses interval routing. Interval routing is a space-efficient routing method for networks, but the method is static and determinative, and it cannot realize fault-tolerance.

A drawback of the topology construction in method of [7] is that it is not particularly efficient for very dynamic environments. A node that joins or leaves the network could trigger a complete restructuring of the topology. Although algorithm proposed in [8] generates topologies with low total weight, the question of their sub-optimality with respect to the unique minimal k-connected and k-regular overlay graph arises. M.K, Das, et al. [11] has given a comparative study of various routing algorithms along with their analysis. However, they have not focused on fault tolerant routing. A new routing approach named DRS has been described in [12]. This protocol basically looks for faults from time to time in the system.

6 Conclusion

The goal of this paper is to propose a fault tolerant routing algorithm which deals with two types of failure models, crash fault and omission fault. The focus mainly has been on network communication layer. Our approach periodically monitors the network to form the most updated topology with the nodes. It also checks for alternate routes if the gateway is disconnected from the sender node. The application uses minimum memory for the application of fault tolerance. This fault tolerance approach brings redundancy but it is a reasonable trade off for providing reliability.

References

1. Cao, G.: Designing Efficient Fault-Tolerant Systems on Wireless Networks. In: The Proceedings of 8th Annual International Conference on Computer Networks, pp. 263–270 (2004)
2. Tanenbaum, A.S., Van Steen, M.: Distributed Systems, 2nd edn. Pearson (2006)
3. Badera, P., Bhardwaj, A., Rajalakshmi, K.: Range grouping for routing in dynamic networks. In: Parashar, M., Kaushik, D., Rana, O.F., Samtaney, R., Yang, Y., Zomaya, A. (eds.) IC3 2012. CCIS, vol. 306, pp. 95–105. Springer, Heidelberg (2012)

4. Dolev, D., Friedmant, R., Keidar, I., Malkhi, D.: Failure detectors in omission failure environments. In: PODC 1997 Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing, p. 286 (1997)
5. Koushanfar, F., Potkonjak, M., Sangiovanni-Vincentelli, A.: Fault Tolerance in Wireless Sensor Networks. In: Mahgoub, I., Ilyas, M. (eds.) Handbook of Sensor Networks, Section VIII, vol. 36. CRC press (2004)
6. Feng, X., Han, C.: A Fault-Tolerant Routing Scheme in Dynamic Networks. Journal of Computer Science and Technology 16(4), 371–380 (2001)
7. Thallner, B.: Fault tolerant communication topologies for wireless ad hoc networks. In: 1st Workshop on Dependability Issues in Wireless Ad Hoc Networks and Sensor Networks (DIWANS 2004), Florence, Italy (June 2004)
8. Thallner, B., Moser, H.: Topology Control for Fault-Tolerant Communication in Highly Dynamic Wireless Networks. In: Third International Workshop on Intelligent Solutions in Embedded Systems, pp. 89–100 (2005)
9. Thallner, B., Moser, H., Schmid, U.: Topology control for fault-tolerant communication in wireless ad hoc networks. Published in: Journal Wireless Networks 16(2), 387–404 (2010)
10. Fuchs, G.: Implications of VLSI Fault Models and Distributed Systems Failure Models – a Hardware Designer’s view. In: The Proceedings of the Conference on Design, Automation and Test, vol. 2, pp. 300–310 (March 2009)
11. Marina, M.K., Das, S.R.: On-demand Multipath Distance Vector Routing in Ad Hoc Networks. In: 9th International Conference on Network Protocols, vol. 4(6), pp. 14–23 (2001)
12. Chowdhury, A., Friederi, O., Burger, E., Grossman, D., Makki, K.: Dynamic Routing System (DRS): Fault Tolerance in Network routing. In: The Proceedings of 10th International Conference on Computer Networks, vol. 3(8), pp. 23–25 (2004)

Deployment of Sensors in Regular Terrain in Form of Interconnected WSN Units

Arup Kr. Chattopadhyay[1] and Chandan Kr. Bhattacharyya[2]

[1] Dept. of Computer Sc. & Engineering
Academy of Technology, West Bengal, India
ardent.arup@gmail.com
[2] Dept. of Computer Sc. & Engineering
Techno India, Saltlake, West Bengal, India
ckbtechno@gmail.com

Abstract. The cost of algorithm for finding the deployment positions of static sensors in a given terrain is one of the most important issues for sensor deployment. In this paper a new scheme of deployment is proposed considering a regular rectangular terrain. In the proposed algorithm the entire coverage of the terrain is provided by interconnecting a number of predefined WSNs. Each WSN in a normal configuration consists of five sensors as deployable unit, unlike other sensor deployment algorithms, where each sensor is considered as a deployable unit. The efficiency of the algorithm is guaranteed by making deployment decision for a group of sensors together, rather than making decision for each sensor node. A group of sensor nodes already form a WSN, where the sensor at the centre acts as a server-node and the remaining four as client-nodes. The client-nodes are responsible for sensing coverage and any object sensed within the WSN will be reported to the server-node; whereas the server-node is mainly responsible for communicating with the neighboring WSNs. Hence the challenge of the algorithm is to organize the WSNs in a particular order so that interconnection between the WNSs can be established for effective sensing coverage in the given terrain.

Keywords: wireless sensor network, sensor deployment, DIW, WSN deployment, network of WSNs, regular terrain, Interconnected WSN units.

1 Introduction

The entire network of sensors providing sensing coverage in a given terrain can be viewed as an inter-network between a number of small WSNs. Deployment of Interconnected WSNs algorithm or DIW algorithm is proposed for deployment of predefined WSNs in a rectangular terrain. The performance analysis of the algorithm is in terms of number of deployments needed to be done for a given terrain. The predefined WSNs are formed by a sensor at centre, which is the server-node, adjacent with four other sensors, which are client-nodes. Client nodes are homogeneous sensors (with same sensing and communication coverage) mainly used for sensing

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 228–237, 2013.

and communication limited to server-node (belong to same WSN) only, hence less energy is consumed and memory requirement is extremely low. Server-nodes are mainly responsible for data communication. The arrangements of sensors within a WSN are done in a way to provide sensing coverage in rectangular region, which can be called as sensing rectangle. The sensing rectangles can be joined one with another to provide a complete coverage in any given rectangular terrain.

The rest of the paper is organized as follows. Section 2 covers the basic topology of the WSNs with the calculations required for its effective sensing coverage and details the assumptions about the environment considered for the algorithm. In Section 3 the objectives of the algorithm are presented and in section 4 few already proposed solutions are discussed. Section 5 introduces the algorithm and in section 6 the complexity of the algorithm is presented. A comparison between the proposed algorithm and LDM algorithm [9] is sited in section 7 and simulated results are shown in section 8. The limitations and future scopes of enhancements of the algorithm are discussed in section 9.

2 Some Preliminaries and Assumptions

In predefined WSN, the five sensors are arranged in a star topology as show in Fig. 1. In the topology, the sensor at the centre of the WSN acts as server-node and is a heterogeneous (sensing range and communication range are different) sensor. The other four are the client-nodes, homogeneous of type (sensing range and communication range are same).

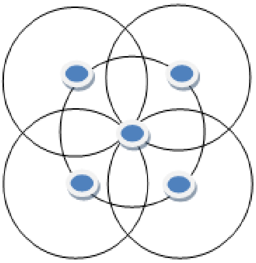


Fig. 1. Topology of WSN

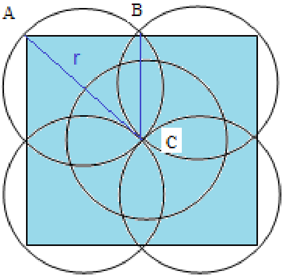


Fig. 2. Calculation of effective coverage area of WSN

Considering the sensing coverage is uniform in all the directions, the sensing area is circular; let it be of radius *r*. Then, the effective coverage of WSN is *8r*^{2} as shown in Fig. 2. The effective coverage of a WSN can be calculated as follows-

Let radius of sensor coverage area is r (in Fig. 2)

So if triangle ABC (in Fig. 2) is considered, $where, AB = BC = w$

$$then, w^2 + w^2 = (2r)^2 \quad or, w = \sqrt{2}r$$

Let, s is each side of total coverage rectangle of WSN, then $s = 2w = 2\sqrt{2}r$.

So, the total coverage area, $A = \left(2\sqrt{2}r\right)^2 = 8r^2$.

Distance between two adjacent WSN canters can be calculated as shown in Fig. 3a and 3b. As per Fig. 3a,

Let $BC{=}w$ then, $w = \sqrt{2}r$.

So, the distance between two adjacent WSNs (as shown in Fig. 3b. i.e. distance between *P1* and *P2*) is $2w = 2\sqrt{2}r$. Then the sensing range of server sensor is $t_s = r$, the communication range $t_c \geq 2\sqrt{2}r.$ For the client sensors, it is $t_s = t_c = r.$

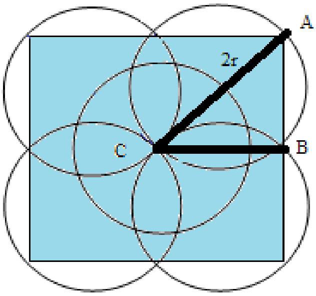


Fig. 3a. Calculation of distance between two deployable positions

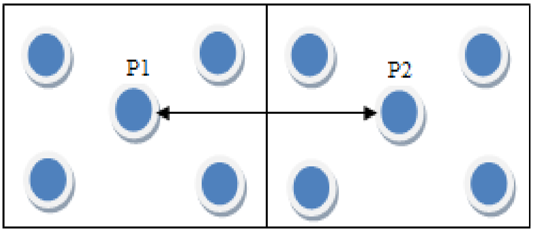


Fig. 3b. Distance between two adjacent deployable positions P1 and P2 calculated as $2\sqrt{2}r$

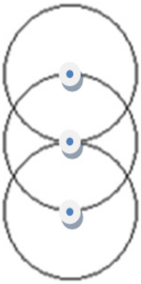


Fig. 4a. Topology of partial WSN-1

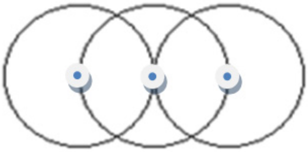


Fig. 4b. Topology of partial WSN-2

Two special kind of WSNs, each consisting of 3 sensors are considered for partial deployments to cover small areas or fragments left after deployments of WSN of 5 sensors or full WSNs. The topologies of partial WSNs as shown in Fig. 4a and 4b are mentioned as partial WSN-1 and WSN-2. The sensor at the center is server-node ($t_s = r, t_c \geq 2\sqrt{2}r.$), and the other two are client-nodes ($t_s = t_c = r$).

The total coverage area by the partial WSN can be calculated as shown in Fig. 5.

$AB = r$ (where r is radius of sensing circle)

$$BC = \frac{r}{2}$$

$$AB = r$$

$$then, r^2 = \left(\frac{r}{2}\right)^2 + (AC)^2 or, AC = \frac{\sqrt{3}r}{2}$$

So, height of coverage area is $\sqrt{3}r$ and length = $6\times(BC) = 6\times\frac{r}{2} = 3r$. Total Coverage = $3r\times\sqrt{3}r = 3\sqrt{3}r^2$.

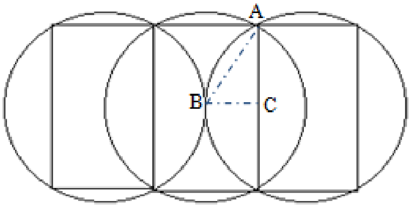


Fig. 5. Calculation of coverage area for partial WSN-2

To keep same alignment with full deployment, the distance between adjacent partial WSNs are kept as $2\sqrt{2}r$. That obviously makes use of $2\sqrt{2}r$ out of $3r$ length. So the actual coverage area = $2\sqrt{2}r\times\sqrt{3}r = 2\sqrt{6}r^2$.

3 Problem Definition and Objectives

The key objectives of this proposed deployment algorithm are as follows-

- Deployment of interconnected WSNs, with a low cost algorithm.
- Complete coverage of the given terrain or AoI [9].
- No effective coverage outside the given terrain or AoI [9].
- Developing algorithm which will deploy the special partial WSNs (as shown in Fig. 4) so that the wastage of sensing coverage can be minimum, i.e. no deployment beyond AoI.

4 Related Works

Different algorithms have already been developed for deployment of sensors in regular terrain, irregular terrain and irregular terrain with obstacles. In [1], [2], [3], [4], [5], [6] and [7], different processes of deployment within a regular rectangular terrain have been proposed. In [1], an irregular shape is partitioned into regular shapes, and then farther sub-partitioned so that the sensors can be deployed. In [2] a fuzzy based key redistribution determining method is used for deployment. In [3], movement assistant principal was introduced for moving sensors from densely deployed area to less densely deployed area. In [4], a unified framework for movement assistant deployment has been proposed. In [5], a "Virtual Rhomb Grid based Movement-assisted Sensor Deployment" algorithm has been proposed, which starts with a rectangular shaped terrain with randomly placed sensors. It partitions the terrain using a virtual rhomb grid (VRG) and moves the sensors on vertices of VRG. In [6], the authors considered a regular shape terrain with holes. They proposed linear-time algorithm to identify the boundary nodes, after a random deployment of sensors in terrain.

An application specific solution is provided in [7], which describes a non-uniform deployment of sensors. In [8], authors consider irregular terrain without any holes, uses robot deployment mechanism for near-minimal number of sensor deployment. In [9] and [10], the LDM algorithm has been proposed for deployment of sensors, routing model and tracking objects and irregular terrain with obstacles has been considered. LDM makes decision for each and every sensor about its deploy-ability, hence reduce performance. A major comparison of DIW and LDM is done in the performance section.

5 Proposed Model

Considering a rectangular terrain the deployment is started from the top-left corner. The number of WSNs to be deployed from left to right is n1 = ⌊ (length of terrain) / (length of each WSN) ⌋. Number of WSNs deployed from top to bottom is n2 = ⌊ (height of terrain) / (height of each WSN) ⌋. So, number of full deployment n1 X n2. The algorithm for full deployment is as follows-

```
algorithm full_deployment(){
    /*let n1 number of WSN to be deployed length-wise;
    let n2 number of WSN to be deployed hight-wise*/
    n1 = floor(length_of_terrain / length_of_WSN);
    n2 = floor(height_of_terrain / height_of_WSN);
    starting_position:=(top_of_terrain+ height_of_WSN/2
    , left_of_terrain+length_of_WSN / 2);
    let, p (point) := starting_position;
    for 1 to n2 do{
          for 1 to n1 do{
          full_deplyment_at(p);
          p.x :=   p.x + 2√2r (with x coordinate)
          }p.y :=   p.y + 2√2r (with y coordinate);
    }
    call partial_deployment();
}
```

The full deployment can leave a small uncovered area at the bottom of AoI and at the right side of AoI. The uncovered areas have to be covered by the partial deployment scheme. The algorithm for partial deployment is as follows-

```
  Algorithm partial_deployment(){
      /*let n1 and n2 number of WSN for partial
      deployment at uncovered area at bottom, length-wise
      and height-wise; let uncovered_bottom_length and
      uncovered_bottom_height are the length and height
      of uncovered area of bottom;*/
      n1 := ceiling(uncovered_bottom_length /2√2r));
      n2 := ceiling(uncovered_bottom_height /√3r));
      partial_deployment of n1 X n2 WSNs at
bottom_uncovered_area.
      /*let n3 and n4 number of WSN for partial
      deployment at uncovered area at right-side, length-
      wise and height-wise; let uncovered_right_length
      and uncovered_right_height are the length and
      height of uncovered area of right side.*/
      n3 := ceiling((uncovered_rihgt_length /√3r));
      n4 := ceiling((uncovered_right_height /2√2r));
      partial_deployment of n3 X n4 WSN at
right_uncovered_area.
  }
```

The approach can plot some of the sensors outside the AoI, which is practically not possible. So, algorithm called *reposition()* will reposition those sensors so that the sensors will be positioned within the terrain and their effective coverage area exactly ends at the border of terrain, so no coverage will be provided outside the terrain.

```
  algorithm reposition(){
      for each WSN whose effective coverage area beyond
      the bottom-border of AoI{
            let h is the height of effective coverage
            area of partial deployment below the AoI;
            deploy the WSN, by h distance upwards;
      }
      for each WSN whose effective coverage area beyond
      the right-border of AoI {
      let w is the width of effective coverage area of
      partial deployment beyond the right border of the
      AoI;
      deploy the WSN, by w distance leftwards;
      }
  }
```

The algorithm guarantees no effective coverage (coverage rectangle) outside the terrain as shown in Section 8, Fig. 10. As a result, the algorithm generates deployment database with X-coordinates and Y-coordinates. The format of the database is as follows-

Table 1. Format of 3 deployment databases for Full-Deployment, for partial-Deployment with WSN-1 and for partial-Deployment with WSN-2

Node ID	**Deployment Position**	
	X-coordinate	**Y-coordinate**

6 Complexity Analysis

Considering a rectangular AoI with no obstacle, with width and height *w* and *h* and the coverage radius of sensors r, DIW construct WSN of side $2\sqrt{2}r$ and area of $8r^2$.

Number of deployments per row is $N = \left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right]$ (as horizontal and vertical distance between adjacent nodes 2√2r). Number of deployments per column is $M = \left[\left(\frac{h}{2\sqrt{2}r}\right) - 1\right]$.

Total number of deployments in the given AoI of is *(w X h)* = $N \times M = \left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right] \times \left[\left(\frac{h}{2\sqrt{2}r}\right) - 1\right]$

In worst case, the maximum number of partial deployments at bottom-side and right-side uncovered areas after full deployment will be for 2 rows and 2 columns, i.e. $2 \times \left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right] + 2 \times \left[\left(\frac{h}{2\sqrt{2}r}\right) - 1\right]$. Considering $h = w$ then total deployment in worst case is $\left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right]^2 + 4\left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right]$.

In the algorithms full_deplyment(), partial_deplyment() and reposition() the deployments are done or adjusted row-wise and column-wise. If total number of row-wise and column-wise deployments are m and n respectively, then the cost is $n \times m$. Assuming $m = n$, the cost is n^2. Hence the complexity is $\theta(n^2)$.

7 Performance

The efficiency of the algorithm depends on number of units (WSN) to be deployed. Compare to LDM in [9] and [10], the number of deployments can be definitely reduced as the deployable units in LDM are sensors, whereas in DIW the deployable are WSNs (consisting of 5 sensors).

Considering a rectangular terrain with no holes and pockets, with width and heights are *w* and *h*,

Considering homogeneous sensors with the radius of sensor coverage *r*, for LDM algorithm,

Number of deployment per row, $N = \left(\frac{w}{r} - 1\right)$.

Number of deployment per column, $M = \left(\frac{h}{r} - 1\right)$.

Total deployment = $N \times M = \left(\frac{w}{r} - 1\right) \times \left(\frac{h}{r} - 1\right)$.

Lets assume *w* = *h*, total number of deployment = $\left(\frac{w}{r} - 1\right)^2$.

For DIW algorithm, total number of deployment (in worst case) considering *w*=*h*, is, $\left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right]^2 + 4\left[\left(\frac{w}{2\sqrt{2}r}\right) - 1\right]$ [as discussed in performance analysis].

For a area 100 X 100, comparison of number deployments of the 2 schemes LDM and DIW, is done below-

Table 2. Number of deployment comparison between LDM and DIW

Radius of sensing area	Number of Deployments		
	LDM	**DIW**	**Gain (less number of deployment)**
10	9801	1317	8484
20	2401	344	2057
30	1045	159	886
40	576	92	484
50	361	61	300

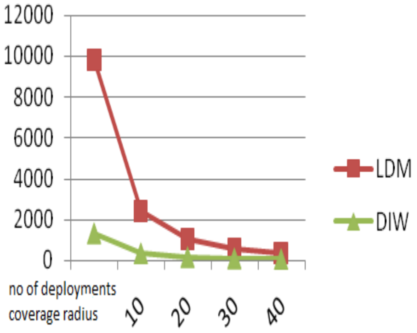


Fig. 6. Comparison of deployment of LDM and DIW

8 Simulated Results

A simple simulator is developed to simulate the DIW algorithm. For example, regular terrain of 500X300 pixels is considered in Fig. 7. The step by step deployment process for the terrain is depicted in Fig. 7, 8, 9 and 10.

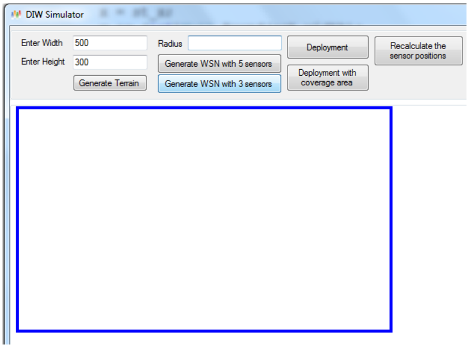

Enter Width 500
Enter Height 300
Radius
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 7. Creation of rectangular terrain (500X300)

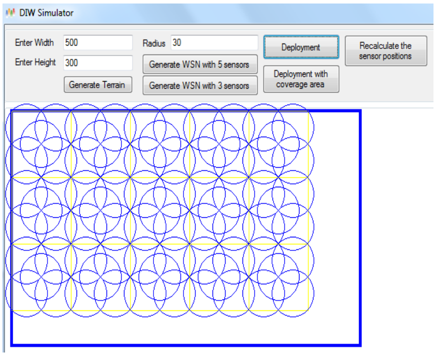

Enter Width 500
Enter Height 300
Radius 30
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 8. Coverage after full deployment

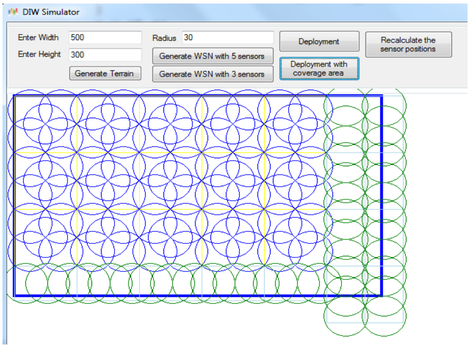

Enter Width 500
Enter Height 300
Radius 30
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 9. Coverage after partial deployment

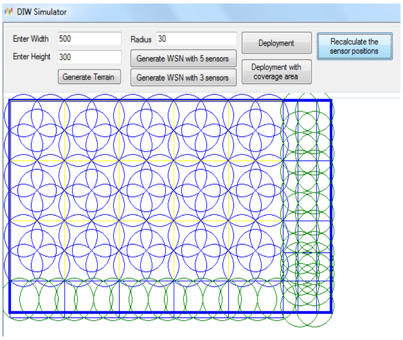

Enter Width 500
Enter Height 300
Radius 30
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 10. Coverage after repositioning WSNs

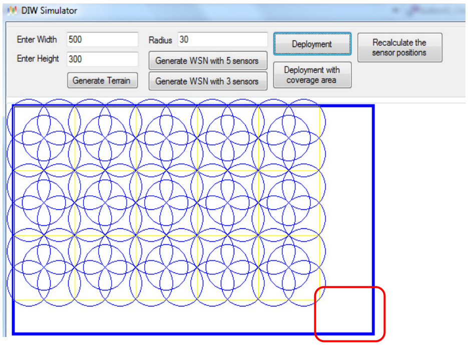

Enter Width 500
Enter Height 300
Radius 30
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 11. Deployment ambiguity marked in red colored box

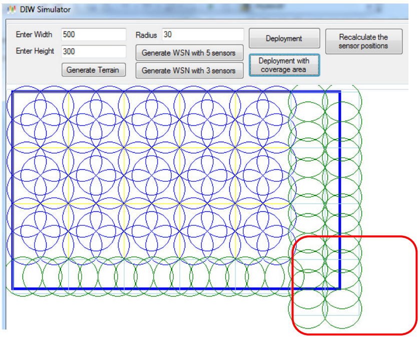

Enter Width 500
Enter Height 300
Radius 30
DIW Simulator
Generate Terrain
Generate WSN with 5 sensors
Generate WSN with 3 sensors
Deployment
Deployment with coverage area
Recalculate the sensor positions

Fig. 12. Resolution of deployment ambiguity in current DIW implementation

9 Conclusions

DIW algorithm is currently restricted to the rectangular terrains. DIW can be further updated to work with irregular terrain with obstacles. Most of the previous algorithms, which start the deployment process from the centre of the terrain, end up with few fragments or uncovered areas at the border of the terrain. One of the key advantages here is that all the fragments will be gathered at a particular side of AoI; so it is easy to address. One of immediate chances of enhancement of this algorithm is in Fig. 11. The uncovered area after full deployment marked in red box can be considered as bottom side uncovered area or right side uncovered area. This can be addressed by right or down partial deployment (deployment ambiguity). Ideally the scheme must be selected so that the number of deployment is minimum. But current implementation always considers it as right-sided deployment as shown in Fig. 12.

References

1. Wang, Y.-C., Hu, C.-C., Tseng, Y.-C.: Efficient deployment algorithms for ensuring coverage and connectivity of wireless sensor networks. In: The Proceedings of First International Conference on Wireless Internet, pp. 114–121 (2005)
2. Zhao, L., Liang, Q.: Fuzzy deployment for wireless sensor networks. In: The Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, pp. 79–83 (2005)
3. Wang, G., Cao, G., Porta, T.L.: Movement-assisted sensor deployment. IEEE Transactions on Mobile Computing 5(6), 640–652 (2006)
4. Fang, C., Low, C.P.: Unified Framework for Movement-Assisted Sensor Deployment. In: The Proceedings of Wireless Communications and Networking Conference, WCNC 2008, March 31-April 3, pp. 2057–2062. IEEE Publication (2008)
5. Wang, X., Yang, Y., Zhang, Z.: A virtual Rhomb Grid-Based Movement-Assisted Sensor Deployment Algorithm in Wireless Sensor Networks. In: The Proceedings of First International Multi-Symposiums on Computer and Computational Sciences, IMSCCS 2006, vol. 1, pp. 491–495 (2006)
6. Wang, Q., Xu, K., Takahara, G., Hassanein, H.: Deployment for Information Oriented Sensing Coverage in Wireless Sensor Networks. In: The Proceedings of Global Telecommunications Conference, GLOBECOM 2006, pp. 1–5. IEEE (2006)
7. Fang, Q., Gao, J., Guibas, L.J.: Locating and by passing routing holes in sensor network. In: The Proceeding of INFOCOM 2004, Twenty-Third Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, pp. 2458–2468 (2004)
8. Chang, C.-Y., Chang, C.-T., Chen, Y.-C., Chang, H.-R.: Obstacle-Resistant Deployment Algorithms for Wireless Sensor Network. IEEE Transactions on Vehicular Technology 58(6), 2925–2941 (2009)
9. Bhattacharyya, C.K., Bhattacharya, S.: LDM (Layered Deployment Model): A Novel Framework to Deploy Sensors in an Irregular Terrain. In: The Proceedings of Wireless Sensor Network, vol. 3, pp. 189–197 (2011)
10. Bhattacharyya, C.K., Bhattacharyya, S.: Tracking a moving object in an irregular area: a guaranteed shortest path routing model. In: The Proceedings of IEEE International Conference INDICON 2009, Ahmedabad, Gujarat, December 18-20, pp. 166–169 (2009)

3-Tier Heterogeneous Network Model for Increasing Lifetime in Three Dimensional WSNs

Samayveer Singh, Satish Chand, and Bijendra Kumar

Department of Computer Engineering,
Netaji Subhas Institute of Technology, Sector-3, Dwarka, New Delhi –110078, India
samayveersingh@gmail.com, schand86@hotmail.com,
bizender@rediffmail.com

Abstract. Homogeneous algorithms assume that the entire sensor node equipped with equal amount of energy. In this paper, a network model has been proposed which incorporate heterogeneity in term of their energy. The term heterogeneity means nodes equipped with dissimilar amount of energy. This model contains three tier node heterogeneity namely tier-1, tier-2, and tier-3 heterogeneity. We assume that nodes are equipped in three dimensions, not mobile, and randomly distributed. It performance is compared with 3D-ALBP, called 3D-hetALBP. Finally, the simulation results demonstrate that our proposed heterogeneous algorithm is more effective in prolonging the network lifetime compared with 3D-ALBP.

Keywords: Heterogeneous, target, sensing range, energy efficiency, sensor network.

1 Introduction

In last couple of years, there have been many studies on wireless sensor networks (WSNs) with respect to their applications in different areas such as security, health, disaster relief, environment, and home applications. Recent advancements in WSNs have enabled the development of low cost sensor networks. There are many issues that affect the design of WSNs such as fault tolerance, scalability, production costs, hardware constraints, network topology, transmission media, power consumption, etc. One of the major restrictions of these sensors is their limited energy. In a WSN, the sensor nodes are scattered in a region of interest and they have self-organizing ability. They collect information from the monitoring region, aggregate it, and then send it to sink or base station through a wireless link for further processing [1]. Since a sensor node has limited energy, it can transmit data to short distance and hence it is not possible in all deployments for a sensor node to directly communicate with sink. One of the most important features of sensor networks is cooperative effort of the sensor nodes. The sensor nodes allow random deployment strategies for different applications. Many distributed and centralized techniques concentrate on minimizing sensing energy by using smart scheduling, adjusting sensing ranges, heterogeneity and maximizing cover sets. In centralized techniques, it is assumed that a single node,

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 238–247, 2013.

usually called base station, has access to the entire network information such as sensing range, location and residual energy of all sensors. Using this information, the base station computes a schedule, which is then provided to each sensor node. In a distributed technique, a sensor can exchange information with its neighbors within a fixed number of hops, which is usually 1- or 2-hop information and then makes scheduling decision, i.e., it decides to move to on or off state. In off state, a sensor saves its energy. One of the solutions for energy problem is to implement mechanisms for efficient energy management. The important methods of efficient energy management are based on scheduling the sensor activity. In scheduling, a sensor node may be in one of the three states at a single time such as active state, deciding state and idle or sleep state. Implementation of energy heterogeneity in network may be another solution for energy problem. There is many type of resource heterogeneity in WSNs such as link heterogeneity, energy heterogeneity and computational heterogeneity, but energy heterogeneity plays a more vital role. All other resource heterogeneities indirectly depend on the energy heterogeneity. In energy heterogeneity, all sensor nodes in a sensor network have different amount of energy. Here, we have use load balancing protocol with adjustable sensing range for incorporating heterogeneity and three dimensions deployment of sensor nodes. In this paper, we propose an energy-efficient heterogeneous network model for maximizing the active duration of a sensor network. For energy-efficient network we use both scheduling and heterogeneity. Our proposed work describes 3-tier of heterogeneity of nodes in a network, called tier-1, tier-2 and tier-3 nodes. The tier-3 nodes have more energy as compared to the tier-1 and tier-2 nodes, and the tier-2 nodes have more energy than that of the tier-1 nodes. The number of sensor nodes in tier-1 is larger than that of the tier-2 and tier-3 nodes each.

The remaining of the paper is organized as follows. In Section 2, we discuss the literature survey related to prolonging the lifetime problems of sensor networks. Section 3 discusses the design of our 3-tier network model and deployment strategy. Section 4 provides a performance evaluation of the proposed work. We conclude the paper in Section 5.

2 Literature Review

In this section, we review the work related to prolonging the lifetime of WSNs. There are different approaches for increasing the network lifetime that include disjoint and non disjoint cover sets, scheduling, adjustable sensing range, and heterogeneity of sensor nodes.

In [2], Slijepcevic & Potkonjak discuss disjoint cover sets by dividing the monitoring area into the fields. Each field is observed by at least one sensor and each sensor covers one or more fields. In this method, the mutually exclusive sets of sensor nodes are selected in such a way that they can cover the entire monitored area. All cover sets remain active for equal amount of time and only a single cover set is active at a time. Berman et al. [3] has extended the work [2] by using scheduling mechanism to schedule the cover sets for enhancing the network lifetime. In [4], Berman et al. discuss another method, called load balancing protocol (LBP), for maximizing the lifetime. The LBP allow sharing the load among sensors and balancing the energy for sensors covering a target. This method uses scheduling mechanism and does not require cover sets to be disjoint. In scheduling, a sensor can be in one of the three states namely active, sleep and vulnerable. Initially, all sensors are assumed to be in vulnerable state. They broadcast their battery tiers along with the target covering information to its neighbouring nodes. By using this information, a neighbouring sensor changes its state and stay in that state until another sensor cannot cover the same target. The drawback of LBP is that it balances the load among the sensors rather the energy for covering the target. This problem has been overcome in a new deterministic energy- efficient protocol (DEEPS) [5]. In this method, the sensors can change their states from idle to active or vice versa while monitoring or communicating. It basically minimizes the energy consumption rate in sensing a target by sending the sensor nodes into sleep mode that have smaller energy with respect to that target. Cardei et al. [6] discuss adjustable range set covers (AR-SC) problem in which a maximum number of (non-disjoint) set covers are determined and the range of each sensor in such a way that each set covers all targets. The energy consumption of the sensor nodes is optimized by using the adjustable sensing range approach. Dhawan et al. [7] discuss two distributed algorithms to maximize the network lifetime for target coverage by providing adjustable sensing and communication ranges capabilities to sensors. Their works may be considered as an enhancement of distributed algorithms for fixed range sensors [4, 5].

The energy consumption can also be optimized if the sensor network supports node heterogeneity. Some of the important recent works based on heterogeneity are discuss in [8,9,10]. In [10], Qing et al. discuss two-level heterogeneity and multi-level heterogeneity model for WSNs. Two-level heterogeneity model considers the two types of sensor nodes. Categorization of sensor nodes is defined in term of their energies. The multi-level heterogeneity considers multiple types of nodes whose energies are defined from a given energy interval. Kumar et al. discuss a three level heterogeneous network model that considers three types of sensor nodes [8,9] namely normal, advanced and super nodes. The energy of a super node is larger than that of a advance node and that of a advance node is larger than that of a normal node. In this model, the number of normal nodes is more than the advance nodes and the number of advance nodes is more than the super nodes. In our proposed work, we also use three types of sensor nodes: tier-1, tier-2 and tier-3 that resemble to normal, advance and super nodes, respectively. Furthermore, we deploy the sensor nodes in 3-D environment. Our proposed model performs better than the existing models [8,9]. In next section, we discuss a proposed heterogeneous network model for wireless sensor networks.

3 Proposed Heterogeneity Network Model

In this section, we propose a heterogeneity energy model for increasing the lifetime of wireless sensor network. In this model network, the nodes are categorized into three types namely tier-1, tier-2 and tier-3. Let *n* be the total number of sensor nodes in a given network. The number of tier-3 nodes is *m/2* times of the total nodes, each having α times more energy than that of a tier-1 node. Thus the number of tier-3 nodes, denoted by N_3, is given by

$$N_3 = ((m/2)*n) \tag{1}$$

The number of tier-2 nodes is *m* times of the total nodes, each having *α/2* times more energy than that of a tier-1 node. Thus the number of tier-2 nodes, denoted by N_2, is given by

$$N_2 = (m*n) \tag{2}$$

The remaining nodes are the tier-1 nodes, each having energy E_0. Thus the number of tier-1 nodes in the network, denoted by N_1, is given by

$$N_1 = n\text{-}N_2\text{-}N_3 = n\text{-}(m/2)*n\text{-}(m*n)$$

$$N_1 = ((1\text{-}(m/2)\text{-}m)*n) \tag{3}$$

The total energy of the network is obtained as the sum of energies of all tier-1 nodes, tier-2 nodes, and tier-3 nodes. Mathematically it is given as follows.

$$E_{total} = N_1*E_0+ N_2*\alpha*E_0+ N_3*(1+\alpha)*E_0 \tag{4}$$

Putting the values of N_1, N_2 and N_3 from (1), (2)and (3) in (4).

$$E_{total} = ((1\text{-}(m/2)\text{-}m)*n)*E_0+ (m*n)*\alpha*E_0+ ((m/2)*n)*(1+\alpha)*E_0$$

$$E_{total} = n*((1\text{-}(m/2)\text{-}m)*E_0+ m*\alpha*E_0+ (m/2)*(1+\alpha)*E_0)$$

After simplifying, we have E_{total} as follows.

$$E_{total} = n*E_0*(1+m*\alpha) \tag{5}$$

From (5), it shows that, the total energy of the network has increased by a factor of *(1+m*α)*. In this network model, the deployment of sensor nodes is made in three dimensions. The points *(x, y, z)* covered by a sensor with range *r* that is centered at (x_1, y_2, z_3) satisfy the following relation.

$$(\mathrm{x} - \mathrm{x}_1)^2 + (\mathrm{y} - \mathrm{y}_1)^2 + (\mathrm{z} - \mathrm{z}_1)^2 = \mathrm{r}^2$$

$$(x\text{-}x_1)^2 + (y\text{-}y_1)^2 + (z\text{-}z_1)^2=r^2 \tag{6}$$

Equation (6) indeed represents a sphere in 3-D whose radius is *r* and centre is (x_1, y_2, z_3). All the points inside the sphere or its boundary will be monitored by the sensor placed at its centre.

4 Simulation Results and Discussions

In this section, we discuss the implementation of ALBP protocol for our proposed network model. This model explains 1-tire, 2-tire, and 3-tire heterogeneity of a WSN, and we call it as 3D-hetALBP. The energy models used in our simulation results are linear and quadratic energy models, commonly used energy models in literature [4,5]. The linear model is given by $e_p = c_1 * r_p$, where c_1, a constant, is given by $c_1 =$

$\frac{E_{Total}}{\sum_{r=1}^{p} r_p}$ and e_p refers to energy to cover a target at distance r_p. The quadratic model is given by $e_p = c_2 * r_p^2$, where c_2, a constant, is defined by $c_2 = \frac{E_{Total}}{\sum_{r=1}^{p} r_p^2}$.

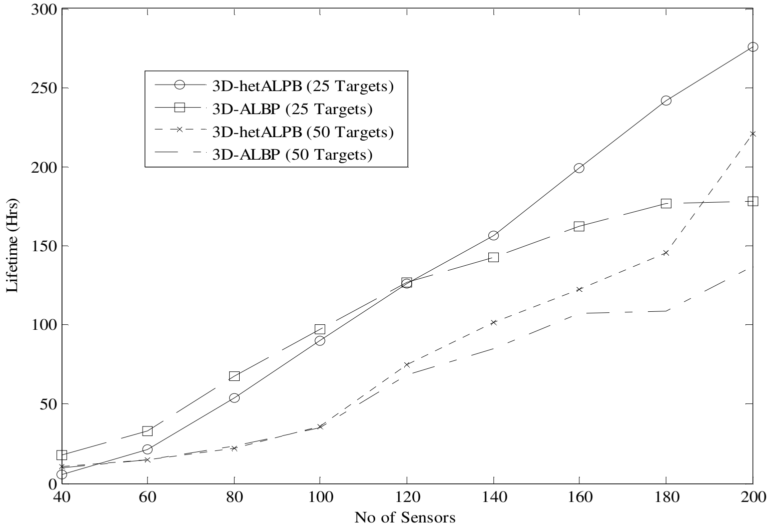


Fig. 1. Network lifetime with respect to number of sensors for linear energy model at 30M sensing range

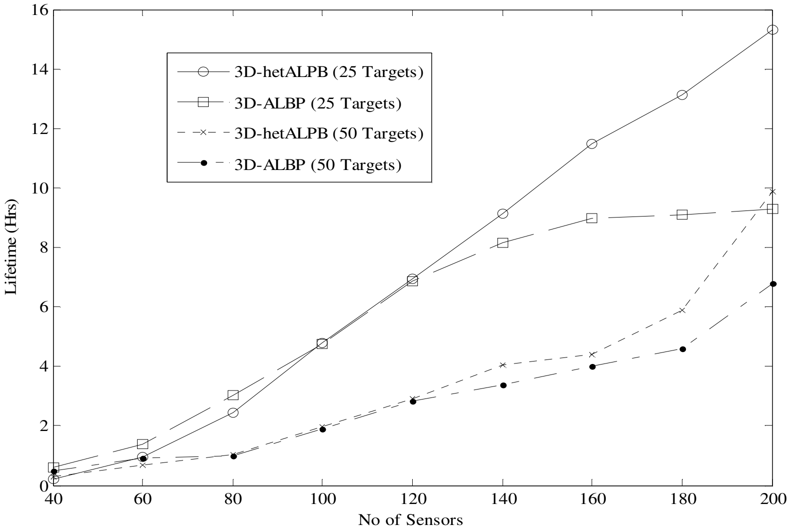


Fig. 2. Network lifetime with respect to number of sensors for quadratic energy model at 30M sensing range

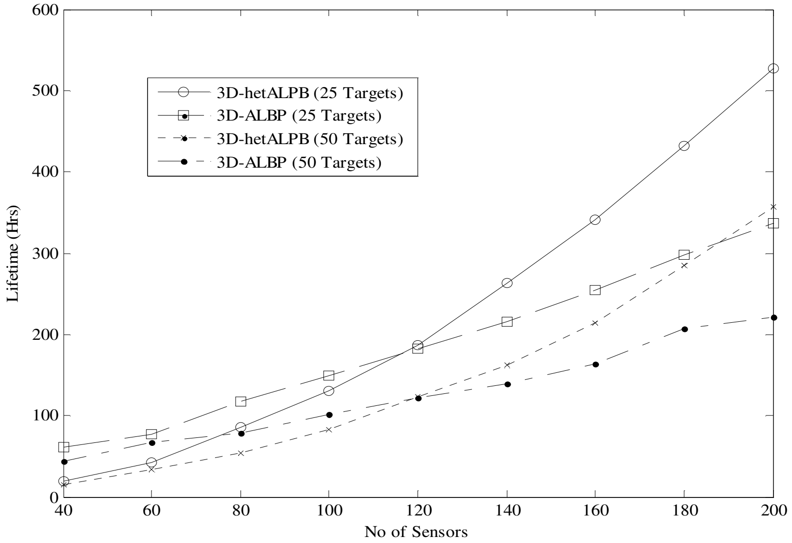


Fig. 3. Network lifetime with respect to number of sensors for linear energy model at 60M sensing range

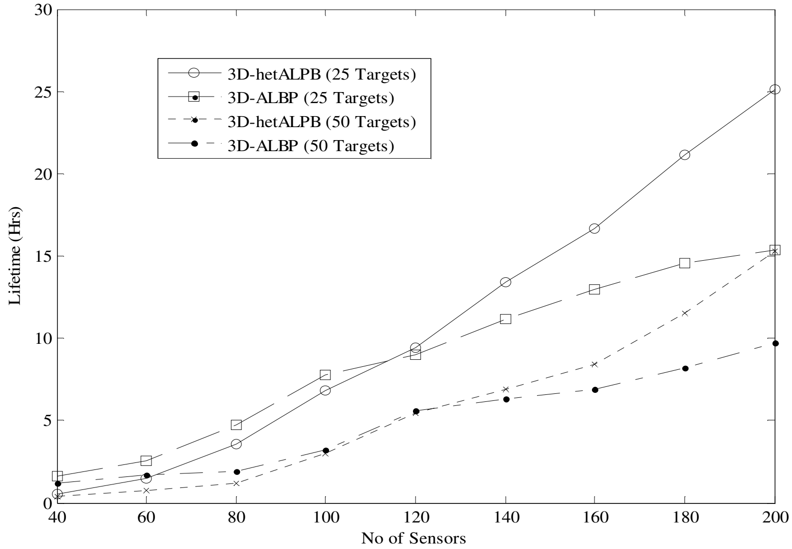


Fig. 4. Network lifetime with respect to number of sensors for quadratic energy model at 60M sensing range

We take monitoring area of size 100Mx100M for hosting two different number of targets i.e., 25 and 50 targets. The number of sensor nodes varies from 40 to 200 by taking two different maximum sensing ranges as 30M and 60M. We have used sensor initial energy as 2J and the values of α and m are taken as 2 and 0.3, respectively. In this scenario, we have carried out simulations for several sets of initial energies of the nodes and parametric values. In all cases, we got similar kinds of results: however, we have shown simulation results for the above mentioned input parameters, which are shown in Figs.1-4.

Figs 1 and 2 show the results for homogeneous 3D-ALBP and 3D-hetALBP using linear and quadratic energy models, respectively. Each of these figures shows network lifetime with respect to the number of sensors for sensing range 30M and, 25 and 50 targets, for homogeneous 3D-ALBP, 3D-hetALBP protocols. It is evident from these figures that the 3D-hetALBP protocol provides longer lifetime than that of the homogeneous 3D-ALBP protocol. We also observe from these figures that increasing the number of sensors increases the network lifetime. In Figs. 3 and 4, the sensing range has been taken as 60M and remaining parameters are kept unchanged. The heterogeneous 3D-hetALBP significantly performs better than the homogeneous 3D-ALBP. Increasing the density of targets decreases the network lifetime, whereas increasing the density of sensor nodes increases the network lifetime.

Figs. 5 and 6 show the network lifetime for our heterogeneous network model and existing heterogeneous network model [8,9] for linear and quadratic energy models, respectively, for the input parameters: 200 number of sensors, 25 & 50 targets, and 30M & 60M sensing range. It is evident from these figures that our proposed heterogeneous network model performs better than the existing model [8,9].

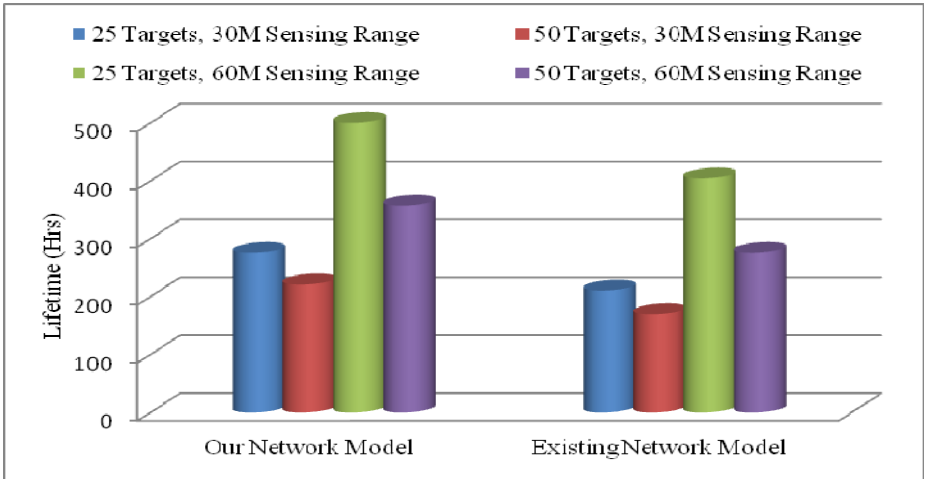


Fig. 5. Network lifetime comparison our and existing network model for linear energy model at different targets and sensing range

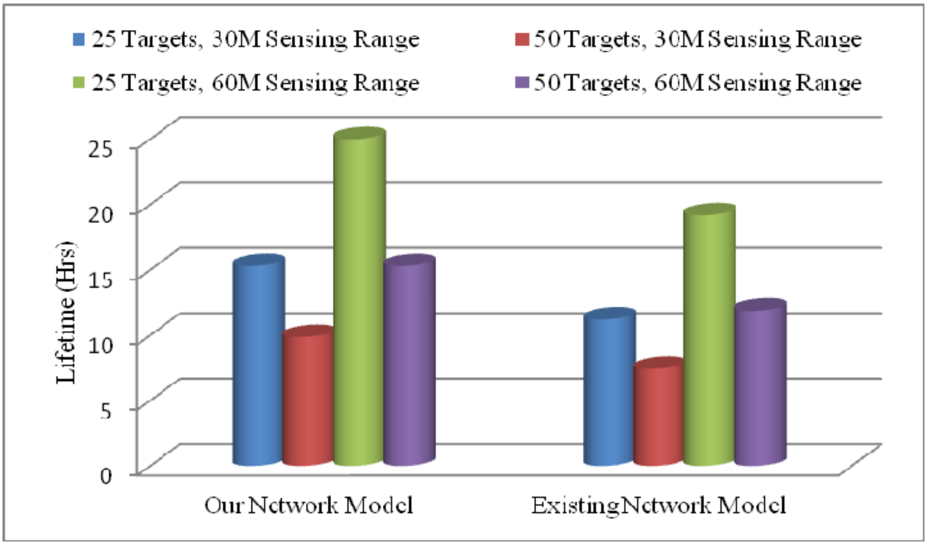


Fig. 6. Network lifetime comparison our and existing network model for quadratic energy model at different targets and sensing range

Another important parameter for network lifetime is the number of rounds when the first and last sensor nodes become dead. Tables I & II show the number of rounds when first and last nodes become dead for linear and quadratic energy models. It is evident from these tables that the first and last nodes become dead in more number of rounds using 3D-hetALBP than that of the homogeneous 3D-ALPB for both the linear as well as quadratic energy models.

Table 1. Round number when first and last nodes dead using linear and quadratic energy models in homogeneous 3D-ALBP and 3D-hetALBP for 30M sensing range, 200 sensors in 100Mx100M

30 M Sensing range and 200 number of sensors				
	Linear energy model		Quadratic energy model	
Cases	First node dead	Last node dead	First node dead	Last node dead
3D-hetALPB (25 Targets)	268	348	19	31
3D-ALBP (25 Targets)	166	186	06	21
3D-hetALPB (50 Targets)	249	257	11	25
3D-ALBP (50 Targets)	138	145	03	23

Table 2. Round number when first and last nodes dead using linear and quadratic energy models in homogeneous 3D-ALBP and 3D-hetALBP for 60M sensing range, and 200 sensors in 100Mx100M

60 M Sensing range and 200 number of sensors				
Cases	Linear energy model		Quadratic energy model	
	First node dead	Last node dead	First node dead	Last node dead
3D-hetALPB (25 Targets)	562	657	18	109
3D-ALBP (25 Targets)	314	409	08	100
3D-hetALPB (50 Targets)	339	418	07	77
3D-ALBP (50 Targets)	206	255	04	69

5 Conclusion

In this paper, we have proposed the 3D-hetALBP, an implementation of ALBP using 3 dimensional deployment and the heterogeneity model of tier-3 for WSNs. The model is capable to describe tier-1, tier-2 and tier-3 heterogeneity. The 3D-hetALBP provides longer network lifetime than the 3D-hetALBP for both linear and quadratic energy models. Furthermore, increasing the sensing range increases the lifetime of the heterogeneous sensor networks.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38, 393–422 (2002)
2. Slijepcevic, S., Potkonjak, M.: Power efficient organization of wireless sensor networks. IEEE International Conference on Communications (ICC) 2, 472–476 (2001)
3. Berman, P., Calinescu, G., Shah, C., Zelikovsky, A.: Efficient Energy Management in Sensor Networks. In: Ad Hoc and Sensor Networks, Wireless Networks and Mobile Computing, vol. 2, Nova Science Publishers (2005)
4. Berman, P., Calinescu, G., Shah, C., Zelikovsky, A.: Power Efficient Monitoring Management in Sensor Networks. In: IEEE Wireless Communication and Networking Conference (WCNC 2004), Atlanta, pp. 2329–2334 (2004)
5. Brinza, D., Zelikovsky, A.: DEEPS: Deterministic Energy-Efficient Protocol for Sensor networks. In: ACIS International Workshop on Self-Assembling Wireless Networks (SAWN 2006), Proc. of SNPD, pp. 261–266 (2006)
6. Cardei, M., Wu, J., Lu, M.: Improving network lifetime using sensors with adjustable sensing ranges. Int. J. Sensor Networks 1(2), 41–49 (2006)

7. Dhawan, A., Aung, A., Prasad, S.K.: Distributed Scheduling of a Network of Adjustable Range Sensors for Coverage Problems. In: Prasad, S.K., Vin, H.M., Sahni, S., Jaiswal, M.P., Thipakorn, B. (eds.) ICISTM 2010. CCIS, vol. 54, pp. 123–132. Springer, Heidelberg (2010)
8. Kumar, D., Aseri, T.S., Patel, R.B.: EEHC: Energy efficient heterogeneous clustered scheme for wireless sensor networks. Int. Journal of Computer Communications 32(4), 662–667 (2009)
9. Mao, Y., Liu, Z., Zhang, L., Li, X.: An Effective Data Gathering Scheme in Heterogeneous Energy Wireless Sensor Networks. In: Int. Conf. on Computational Science and Engineering, vol. 1, pp. 338–343 (2009)
10. Qing, L., Zhu, Q., Wang, M.: Design of a distributed energy-efficient clustering algorithm for heterogeneous WSNs. Int. Journal of Computer Communications 29, 2230–2237 (2006)

A Comparative Study of Reactive Routing Protocols for Industrial Wireless Sensor Networks

Manish Kumar[1], Itika Gupta[2], Sudarshan Tiwari[3], and Rajeev Tripathi[1]

[1] Electronics and Communication Engineering Department,
Motilal Nehru National Institute of Technology, Allahabad
[2] Madan Mohan Malaviya Engineering College, Gorakhpur
[3] National Institute of Technology, Raipur
{rel1101,rt}@mnnit.ac.in, itikagupta28@gmail.com,
sudarshantiwari114@hotmail.com

Abstract. There have been a continuously growing demand for the efficient and infrastructure less Industrial Automation Systems (IAS), to fulfill the rapid changes in real world requirements. During the past decade, wireless sensor networks have been evolved as a powerful tool for the industrial process monitoring and distributed control. In the Industrial wireless control system, sensor nodes are deployed over an area, to monitor physical and or environmental conditions such as temperature, pressure, vibration, motion, sound, humidity etc. For the reliable control function, measured parameters should be delivered to the actuators in real time and reliable manner. Thus, considering the energy constraints of sensor nodes various reactive routing protocols are being used for the real time and reliable message delivery. The aim of this paper is to present a comparative study of existing reactive routing protocols in Industrial Wireless Sensor Networks (IWSNs).

Keywords: AODV, DSR, DYMO, IWSNs, LAR, Reactive routing.

1 Introduction

Today's fast growing competitive industries face growing demand to upgrade their processes for industry automation with improved productivity, efficiency and Quality at low cost [1][15]. Traditionally, IAS are realized through wired communication. However, the wired communication based automation systems require expensive communication cables, to be installed and maintained regularly. Therefore there is a vital need for cost-effective communication systems. Recent developments in the field of Micro Electro Mechanical Systems (MEMS) and WSNs made low cost embedded IAS feasible [2] [3] [15]. The sensor nodes are installed with industrial equipments to monitor various parameters such as temperature, pressure, vibration, power, humidity etc. The sink node (also a sensor node) is installed at remote center for gathering and performing control actions. The nodes work together to analyze and enable control action for maintain quality and productivity. It also warns for repair or replacement of

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 248–260, 2013.

the equipment to meet safety and security norms. A set of sensor node, shares wireless medium, communicate with each other without any predefined infrastructure and central controller. The participating nodes which are not in communication range may communicate in multi hop manner. They follow a well defined hopping sequence in order to comply with the routing policy and to cope well with network dynamics. Hence, the routing protocols play a vital role to select a suitable optimum route, and transmit data on selected path. The reliability of the routing protocol has a direct influence on factors such as network dynamics, coverage, prolonging of the network life time, Packet Reception Rate (PRR), Quality of Service (QoS), fault tolerance and fairness [1][5][15]. Since energy is a main constraint of wireless sensor node, the reactive routing protocols are best suited for reliable data delivery in WSNs. Thus we have chosen and considered reactive routing protocols for the comparative study.

This paper presents the comparative study of performance of well established reactive routing protocols such as AODV, DSR, DYMO and LAR in industrial prospects using QualNet simulator. The organization of the paper is as follows: Section 2 describes routing protocols in IWSNs for real time reliable message delivery. Section 3 describes the simulation scenario, Section 4 presents a discussion over simulation results and finally, Section 5 concludes the paper with future research directions.

2 Routing in Industrial Wireless Sensor Networks

The information routing is most challenging task in the industrial wireless environment due to the inherent characteristics of the wireless sensor networks as dense deployment, nodes mobility and energy constraints. The major issues that need to be addressed are maximizing network lifetime, minimizing latency, resource awareness, topological changes, location awareness, scalability, reliability and real time data delivery [6]. As the sensor networks are a type of Mobile Ad-hoc Networks (MANET), same routing protocols can also be used for IWSNs [4].

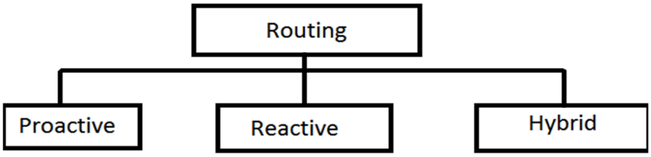


Fig. 1. Routing protocol classification

As shown in figure 1, the routing protocols may be classified as proactive reactive and hybrid. Proactive routing protocol is also known as table driven routing protocol [7]. Each node is required to store routing information in the network. The network status is updated either periodically or when the network topology changes, results in low latency, thereby suitable for real-time traffic, but the bandwidth gets wasted due to periodic updates. Moreover, as these protocols are not energy efficient, its reliability is questionable as for as WSN is concerned. Example protocols are Ballman ford Routing Protocol, Destination Sequenced Distance Vector Routing (DSDV), Optimized Link State Routing Protocol (OLSR) and Source Tree Adaptive Routing (STAR).

Reactive routing protocol discovers the route, when needed, hence called “On Demand routing protocol”. The process of route discovery is done by flooding the Route REQuest (RREQ) packets throughout the network. Reactive routing protocol always uses the current status of network hence the traffic is generated in bursty manner, which may create congestion during high activities. The significant delay may occur as a result of route discovery. But it saves energy and bandwidth during inactivity period. It’s good for low traffic [7]. Examples Protocols are Ad-Hoc On-demand Distance Vector (AODV), Dynamic Source Routing (DSR), Dynamic MANET On-demand (DYMO) and Location Aided Routing (LAR).

The Hybrid routing protocol is the mixed structure of proactive and reactive routing protocol. The best features of proactive routing protocol and reactive routing protocols as low latency and less bandwidth requirement respectively are being incorporated in hybrid routing as attempts to strike balance between the two. The example protocols are, Zone Routing Protocol (ZRP), Core-Extraction Distributed Ad-hoc Routing (CEDAR).

As for as real time reliable delivery is concern the reactive routing protocols may be used because it saves bandwidth and requires fewer resources which may lead it, to be used for industries. The energy saving in inactive period improves reliability and low traffic requirement makes reactive routing protocol suitable for IWSNs.

2.1 Ad-Hoc On Demand Routing Protocol

The Ad hoc On-Demand Distance Vector (AODV) is a reactive routing protocol. On demand basis it establishes a route to the destination hence it is called as AODV. It enables dynamic, self-starting, multihop routing between participating nodes wishing to establish and maintain an ad hoc network. AODV allows obtaining routes quickly for new destinations, and does not require nodes to maintain routes to destinations that are not in active communication [8] [9]. AODV allows mobile nodes to respond to link breakages and changes in network topology in a timely manner. The operation of AODV is loop-free, and by avoiding the Bellman-Ford "counting to infinity" problem offers quick convergence when the ad hoc network topology changes. When a node needs to know a route to a specific destination it creates a Route Request, forward to sink node through intermediate nodes. Intermediate node stores a reverse route. When the request reaches a sink node, creates a reply which contains the number of hops that are requiring reaching the destination [9]. All intermediate node, stores the forward route in their routing table. This route created from each node from source to destination is a hop-by-hop state and not the entire route as in source routing.

2.2 Dynamic Source Routing Protocol

The Dynamic Source Routing protocol (DSR) is routing protocol designed for multi-hop wireless networks where nodes are mobile. DSR allows the network to be completely self-organizing and self-configuring [10]. It uses the route discovery cycle for route discovery and uses the route maintenance cycle to maintain the active routes. These cycles work together, discover and maintain the route in reactive manner, create the routing packet overhead for searching the routes dynamically. It prolongs the network life time, load balance, by providing flexibility to sender to choose and control route among selected routes to destination [11]. When a node S wants to send a packet to node D, but does not know a route to D, node S initiates a route discovery. Source node S floods RREQ packet. The RREQ consist of sender address, destination address and unique request id determined by the sender. Each node appends its own identifier when it forwards the RREQ packet. The Route REPly (RREP) can be send by reversing the RREQ only if links are guaranteed to be bidirectional. If unidirectional links are allowed, then RREP may need a route discovery for S from node D. Route discovery is not needed if D already knows a route to node S. If a route discovery is initiated by D for a route to S then the RREP is piggybacked on the RREQ from D. DSR protocol provides loop- free routing, rapid recovery when routes in the network change. The DSR protocol is designed mainly for mobile ad hoc small networks (Approx. two hundred nodes) with high mobility [10].

2.3 Dynamic MANET On-Demand

The DYnamic MANET On-demand (DYMO) routing protocol is designed for mobile nodes, intended to work in wireless multihop networks [12]. It dynamically changes the network topology and on demand it determines unicast routes between nodes. The route discovery and route management are the two basic operations of the DYMO routing protocol. In route discovery phase the Source node initiates dissemination of a RREQ packet throughout the network to find the destination node. In the dissemination process, each intermediate node records a route to the source node. As the destination node receives the RREQ packet, responds with a RREP packet. RREP packet is a unicast massage intended for source node. The intermediate node that receives the RREP message records the route for destination node. As the source node receives the RREP, the full duplex path between source and destination has been established. At the time of network changes, nodes maintain their routes and monitor their links, if communication between source destinations breaks, the source node is notified by Route ERRor (RERR) packet. Hence the source node again initiates the route discovery if it still has packets to be delivered.

DYMO routing protocol is basically designed for small, medium, and large population mobile ad hoc networks [12], it can handle large mobility ranges, various traffic patterns, but it best suited to lightly loaded networks. It requires small memory to store active destination only instead of storing all destination routing information. Further a Sequence Number (SeqNum) is maintained by each node. This sequence number insures the freshness of related routing information for loop-free routes.

2.4 Location Aided Routing

Location aided routing (LAR) is a demand based routing technique. It utilizes the location information to optimize the routing overhead by flooding route request packets in selected area. When a source node requires a path to destination node, source broadcast RREQ Packet to all its neighbor nodes. The neighbor node that receives RREQ packet compares its identifier with desired destination identifier. If the match found the node knows that request is for a route to itself, forwards the message otherwise node broadcasts the request to its neighbors to avoid redundant transmission. As the RREQ packet being forwarded the path following information is incorporated in the RREQ. As the destination node receives RREQ packet, reply to sender through RREP packet. The RREP packet follows the path mentioned in the RREQ packet received by destination node. To avoid the loss of route discovery the sender set a timeout at the RREQ transmission, if the RREP is not received before timeout, a new RREQ packet is initiated with different sequence number. The sequence number are useful in duplicate packet receptions at same route. During the data transmission, when a packet moving towards destination did not get the path towards the destination sends RERR packet to source node, to discover the new route.

3 Simulation Setup

The performance parameters were observed and analyzed using simulation model based on QualNet [13]. QualNet is discrete event network evaluation simulator software, developed by Scalable Networks and is a commercialized version of GloMoSim. It analyzes the performance of wired, wireless and hybrid network, supports thousands of nodes for simulation and works with 32 and 64 bit operating systems like UNIX, Linux and MacOS.

Reactive routing protocols as AODV, DSR, DYMO and LAR were considered for simulation using QualNet 5.0.2, considering the varying node mobility while maintaining the constant traffic load using 4 CBR, with different node density as 25, 50, 75 and 100 on a 500x500 meter^2 scenario as given bellow.

Table 1. Scenario Parameters

Simulation Parameters	**Value**
Routing Protocols	AODV, DSR, DYMO, LAR
Scenario Dimensions (meters)	500x500
Simulation time (seconds)	100
No of Nodes	25,50,75,100
Mobility Model	None, Random Waypoint
Path Loss Model	Two Ray
Shadowing Mode	Constant
Pause Time (seconds)	0
Minimum speed (mps)	0
Maximum speed (mps)	1,2,3,4,5
No of CBR Application	04
Packet to be send	∞
Inter Packet Interval(second)	1
Start Time (second)	0
Packet size (bytes)	512

The above said routing protocols were configured with the matrix like Packet Reception Rate (PRR), throughput, jitter and end to end delay with respect to mobility. To support walking speed for maintaining network connectivity as handheld device changes its position in industrial environment, 1 meter per second (mps) as minimum and 5 mps as maximum mobility is considered.

4 Results and Discussion

4.1 Packet Reception Rate (PRR)

The Packet Reception Rate (PRR) represents how many percentages of packets are successfully delivered at destination node [14]. For reliability point of view a larger value of PRR is required. Here we have considered mobility, which may results deteriorate the PRR. Figure 2, 3, 4 and 5 represents the performance of PRR-Vs mobility with AODV, DSR, DYMO and LAR routing respectively. The evidence shows that more or less selected routing protocol performs equally well with different node density and mobility but AODV routing protocol performs exceptionally well, remains constant as node density and mobility increases.

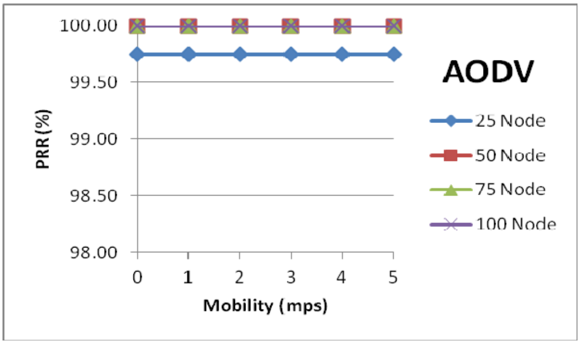


Fig. 2. PRR Vs Mobility for AODV Routing Protocol with varying Node Density

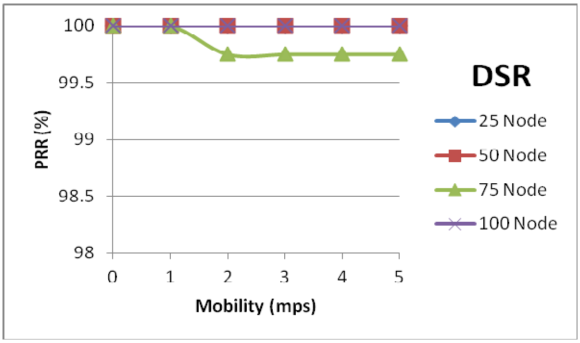


Fig. 3. PRR Vs Mobility for DSR Routing Protocol with varying Node Density

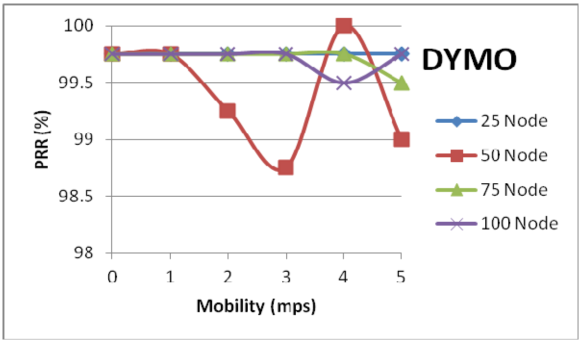


Fig. 4. PRR Vs Mobility for DYMO Routing Protocol with varying Node Density

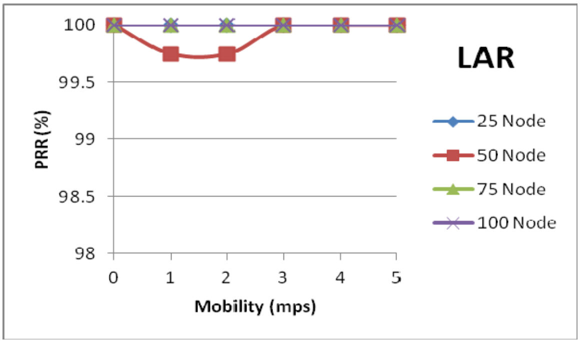


Fig. 5. PRR Vs Mobility for LAR routing protocol with varying node density

4.2 Throughput

Throughput of a network is a measure of the average rate of successful message delivered over a communication channels. In WSNs scenario the reliability depends upon the throughput also. A higher value of throughput is required for better network reliability and capability. Figure 6, 7, 8 and 9 shows the effect of mobility on the

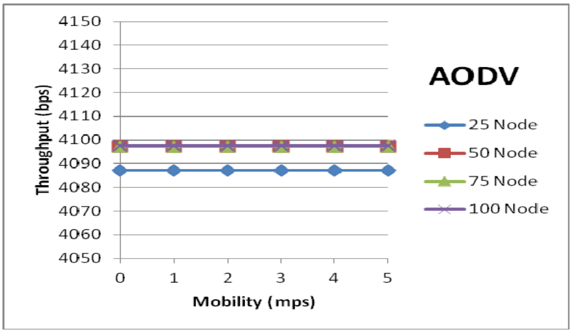


Fig. 6. Throughput Vs Mobility of AODV routing protocol with varying node density

varying network density with AODV, DSR, DYMO and LAR routing respectively. The result shows that with the selected range of mobility AODV, DSR and LAR performs better than DYMO routing but the performance of LAR is outstanding. It provides maximum throughput value.

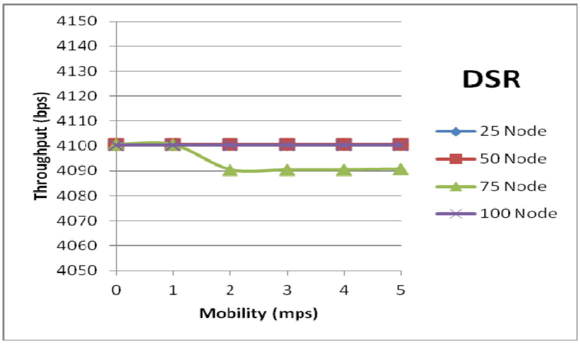


Fig. 7. Throughput Vs Mobility of DSR routing protocol with varying node density

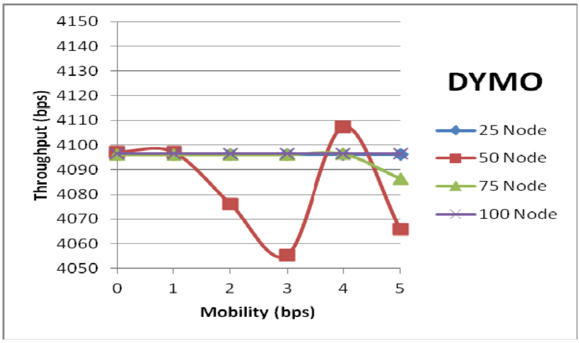


Fig. 8. Throughput Vs Mobility of DYMO routing protocol with varying node density

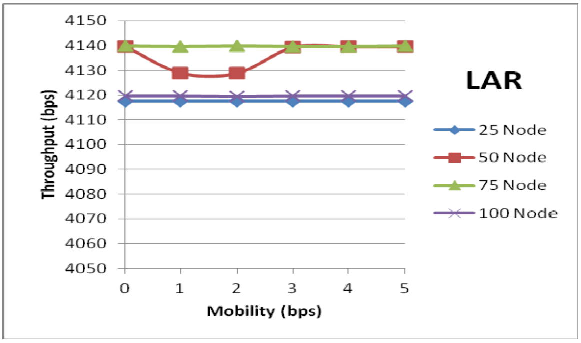


Fig. 9. Throughput Vs Mobility of LAR routing protocol with varying node density

4.3 End to End Delay

The end to end delay refers to the time taken by information to move from source node to destination node. It includes time taken by information at transmitter and receiver end as well as at the channel for propagation. The network performance for dead line delivery of data end to end delay is considered [14]. Results shown in figure 10-13

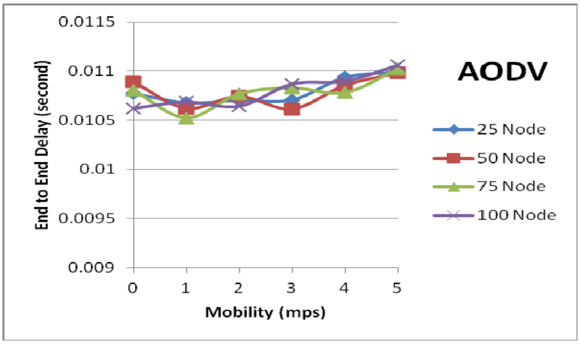


Fig. 10. End to end delay Vs Mobility for AODV Routing with varying node density

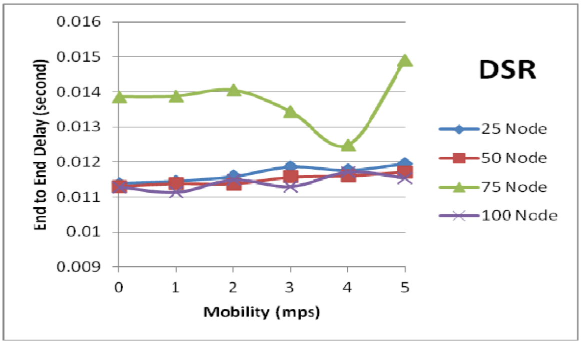


Fig. 11. End to end delay Vs Mobility for DSR Routing with varying node density

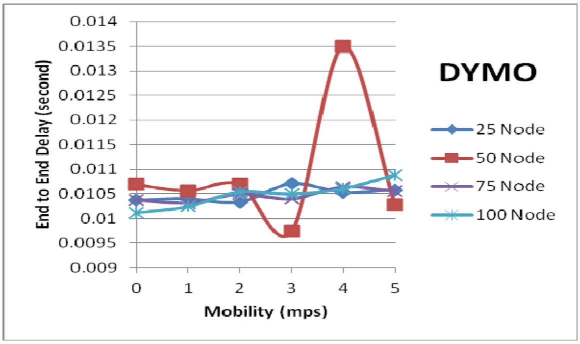


Fig. 12. End to end delay Vs Mobility for DYMO Routing with varying node density

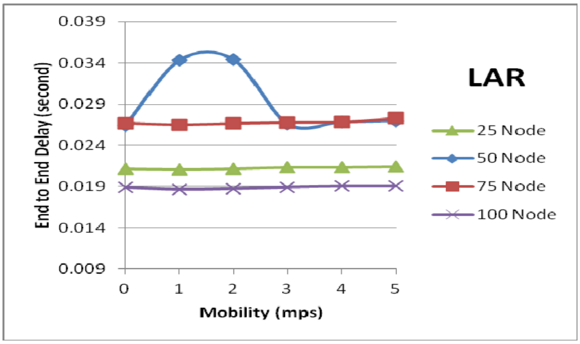


Fig. 13. End to end delay Vs Mobility for LAR Routing with varying node density

shows that DYMO routing protocol performs better than other routing protocols because it requires minimum end to end delay but AODV routing protocol requires slightly more delay for smooth performance which increases slightly with mobility.

4.4 Jitter

Jitter is inter packet delay variation which occurs in packet switch networks. It is variation in latency, effects real time applications, its significant value affect network performance also. Figure 14 to 17 represents jitter variation with mobility in AODV, DSR, DYMO and LAR routing protocols with increasing node density of network. Result show that as per property, AODV and DYMO performs well and jitter increases with mobility. Out of AODV and DYMO, AODV shows slightly more jitter than DYMO.

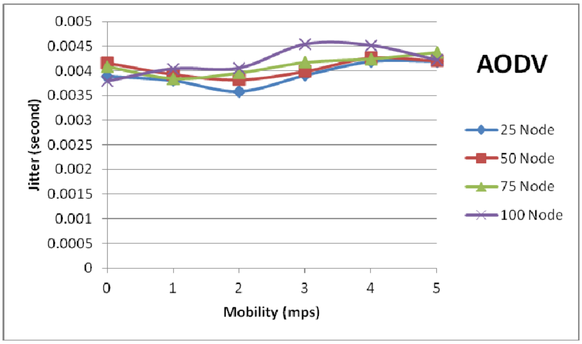


Fig. 14. Jitter Vs Mobility for AODV Routing with varying node density

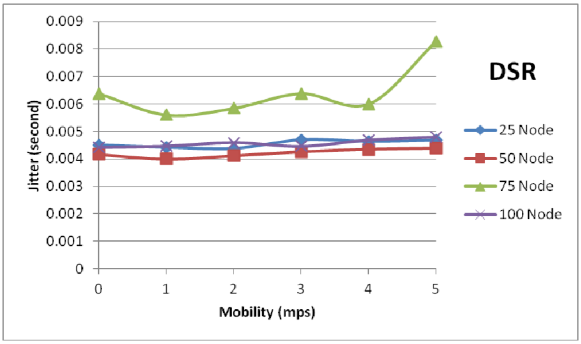


Fig. 15. Jitter Vs Mobility for DSR Routing with varying node density

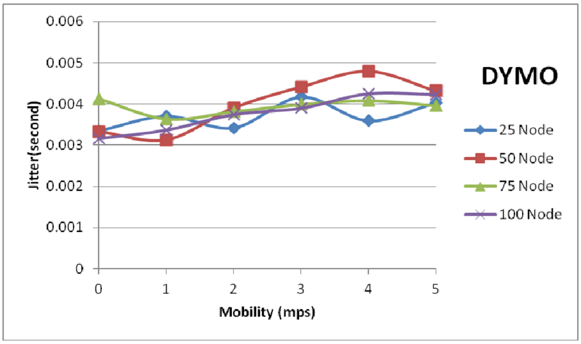


Fig. 16. Jitter Vs Mobility for DYMO Routing with varying node density

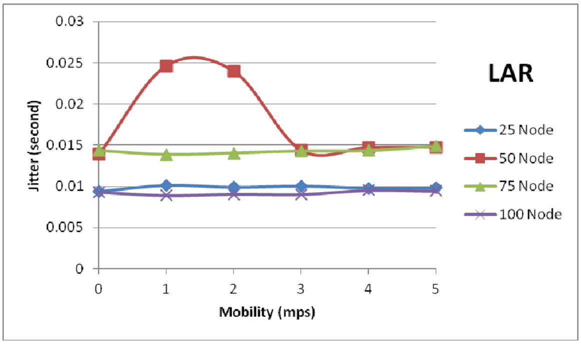


Fig. 17. Jitter Vs Mobility for LAR Routing with varying node density

5 Conclusion

Result shows that as for as PRR and end to end delay performance is concern the AODV routing protocol performs best among other routing protocols. As for as throughput is concern the LAR performs best but for the jitter point of view DYMO routing protocol may perform best. In nut shell tradeoff is required between various considered performance parameters. AODV may be considered best as for as PRR, end to end delay and jitter is concerned. From throughput point of view, LAR routing may perform well.

The reliable packet delivery always depends upon PRR and throughput, real time packet delivery depends upon end to end delay. For real time streaming function end to end delay and jitter play an important role. Real time in Industry automation and control shows that measured parameters and control information should arrive at destination node in acceptable delay limit. The late delivery of information may degrade or damage the performance and functioning. So depending upon the application requirement and criticality the routing protocol may be selected.

Future work will include the gradient setup considering the various metric for each link used to compute the path according to some objective function such as throughput, latency etc that may improve network reliability for real time data delivery in IWSNs.

References

1. Gungor, V.C., Hancks, G.P.: Industrial wireless Sensor Networks: Challenges Design Principles and Technical Approaches. IEEE Trans. Industrial Electronics 56, 4256–4265 (2009)
2. Akyildiz, F., Su, W., Subramaniam, Y.S., Cayirci, E.: Wireless Sensor Networks: A Survey. Computer Networks 38, 393–422 (2002)
3. Christin, D., Mogre, P.S., Hollick, M.: Survey on Wireless Sensor Network Technologies for Industrial Automation: The Security and Quality of Service Perspectives. Future Internet 2, 96–125 (2010)
4. Heo, J., Hong, J., Cho, Y.: EARQ: Energy Aware Routing for Real-Time and Reliable Communication in Wireless Industrial Sensor Networks. IEEE Trans. Industrial Informatics 5, 3–11 (2009)
5. Quang, P.T.A., Kim, D.-S.: Enhancing Real-Time Delivery of Gradient Routing for Industrial Wireless Sensor Networks. IEEE Trans. Industrial Informatics 8, 61–68 (2012)
6. Stojmenovic, I.: Handbook of Sensor Networks: Algorithms and Architectures. Wiley, New York (2005)
7. Royer, E., Toh, C.: A Review of Current Routing Protocols for Ad-hoc Mobile Wireless Networks. Personal Communications, 46–55 (1999)
8. Perkins, C.E., Royer, E.M.: Ad-hoc On-Demand Distance Vector Routing. In: Proc. 2nd IEEE Wksp. Mobile Comp. Sys. and Apps., pp. 90–100 (February 1999)
9. Perkins, C., Belding-Royer, E., Das, S.: Ad hoc on-demand distance vector (AODV) routing. Internet experimental. RFC 3561 (2003)
10. Johnson, D., Hu, Y., Maltz, D.: The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4. RFC 4728 (2007)

11. Broch, J., Johnson, D.B., Maltz, D.A.: The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks. IETF Internet draft, draft-ietf-manet-dsr-01.txt (1998)
12. Chakeres, I., Perkins, C.: Dynamic MANET On-demand (DYMO) Routing. draft-ietf-manet-dymo-05 (June 2006)
13. QualNet simulator, `http://www.scalable-networks.com`
14. Kim, M.-K., Phong, N.H.: Reliable Message Routing Protocol for Periodic Message on Wireless Sensor Networks. In: Proc. 5th Int. Conf. on Ubiquitous Information Technologies and Applications (CUTE), Sanya, pp. 1–6 (2010)
15. Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad-hoc Networks 3, 325–349 (2005)

Mobile Based Attendance System in Distributed LAN Server

Ratnesh Prasad Srivastava[1], Hardwari Lal Mandoria[2], and Rajesh Nautiyal

[1] Department of Information Technology College of Technology, GBPUAT
Pantnagar. 263145, India
write2ratnesh@gmail.com

[2] UniversityD́epartment of Information Technology College of Technology, GBPUAT
Pantnagar. 263145, India
drmandoria@gmail.com

Abstract. With recent advancement in mobile communication, there has been a rise in the number of applications for mobile and its users are also continuously growing throughout the universe. Most of the applications developed, are for en-tertainment and internet. Apart from this many of the utilities have been devel-oped to make handing of mobile devices easy. Numbers of people using mobiles and wireless devices are growing rapidly, but still mobile application market isn't growing at the expected rate. Number of people having mobile are far more than people having computer or laptop. And if we provide application for mobile that will assist people in their day-to-day routine and official work, it can become a po-tential area of development. Considering the ubiquity wireless devices and their ability to be used anywhere and at any time, This paper presents a platform for the intercommunication between applications and load balancing access of dis-tributed LAN servers with implementation issues of application development.

Keywords: Kannel, WAP, WAP Gateway, WAP Services.

1 Introduction

Information is one of the key factors to facilitate the economic growth of a country. Information is a base for performing different activities in sev-eral disciplines. For instance, information regarding the demand of a par-ticular good is essential for a producer to decide when and where to sell the product. As a result of which the producer would be able to maximize profit.

And leads towards a better economic growth of a countryInternet is one way of retrieving information from different areas through the web. Most of the technology developed for the Internet has been designed for desktop and larger computers supporting medium to high bandwidth connectivity on reliable data networks. In recent times, technologies that enable handheld wireless devices to retrieve information have been developed. However, these handheld wireless

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 261–272, 2013.

devices present a more constrained computing environment compared to desktop computers. In addition, providing Internet and WWW (World Wide Web) services on a wireless data network presents many challenges.

And at any time is an interest of everybody, because getting accurate information from anywhere on time is an important element for making a better decision.

Retrieving information through wireless terminal on mobile creates new business opportunities for corporations by providing additional channel for the existing services. The possibility of this additional service is that it can reach cus-tomers 24 hours a day wherever they are. Since mobile services are based on open protocol, it provides the same technology to all vendors regardless of the network system. This common standard offers a better market scale that encourages manufacturers, application developers and content providers to in-vest in developing products that are compatible with mobile. At the present time, accessing web content from anywhere and at any time using mobile devices is possible using WAP technology.

It is thus possible to reliably and efficiently communicate data over wireless WANs (Wide Area Net-works). Retrieval of information using local language is also another critical is-sue for facilitating different tasks for those who work in their local languages. Thanks to the Unicode and other standards, developing multilingual WAP services and getting WAP enabled mobile terminals that support multilingual character set is no more a problem.

This paper is organized as followed: section II introduce the review of related work , introduces the brief description of about the WAP services, section IV introduces the concept of WAP Gateway, section and languages in support for designing the project. Section III introduces conclusion and future work.

2 Review of Related Work

In this paper, we introduce WAP in general terms, and explain the role of the gateway in WAP, outlining their duties and features. It also explains why the Kannel project was started in the first place, and why it is open source.

2.1 WAP Technology

The wireless communication and the Internet are the rapidly growing industries that are gaining more and more customers every day. The WAP intention is to combine these two markets and met the new demands in the field. This and other reasons initiate some of the largest vendors to unite and create the WAP Forum, the standardizing organization of the WAP.

WAP specifies an application framework and network protocols for wireless devices such as mobile phones, pagers and PDA (Personal Digital Assistants).

WAP's specifications extend existing mobile networking technologies and some Internet technologies such as XML (extensible Markup Language) and scripting content formats.

The WAP platform is an open specification that addresses wireless net-work characteristics by adapting existing network technologies (and intro-ducing new ones where appropriate) to the special requirements of hand-held wireless devices . Therefore, WAP intends to standardize the way wire-less devices (mobile phones, PDA, and so forth) access Internet data and ser-vices. WAP's reuse of existing Internet protocols will ease the development of WAP services for Java and other Web developers.

Facilitating the delivery of Internet data to wireless devices will certainly lead to the introduction of new technologies. Wireless handheld devices preset a more constrained computing environment and platforms, compared to desktop computers which most of the Internet technology was devel- oped for. The handheld devices tend to have less powerful CPU's, less memory, very restricted power consumption, smaller and variant displays, phone keypads etc. Furthermore, the wireless networks present additional constraints as communication infrastructures. They have less bandwidth, more latency and less connection stability and unpredictable availability. WAP intends to overcome these difficulties by being interoperable, have scalable quality of service, efficient in the mobile network resources, reliable and se-cure.

WAP allows carriers to strengthen their service offerings by providing subscribers with the information they want and need while on the move. Infrastructure vendors will deliver the supporting network equipment. Application developers and content providers delivering the value added services are contributing to the WAP specification. Enabling information access from handheld devices requires a deep understanding of both technical and market issues that are unique to the wireless environment. The WAP specification was developed by the industry's best minds to address these is-sues.

Nowadays web pages are browsed on mobile terminals using the WAP. Because, WAP is a standardized way for delivering Internet data over wireless networks and capable of addressing the unique characteristics of mobile terminals and wireless networks[1].

2.2 WAP Architecture

The WAP standard defines two essential elements: an end-to-end applica-tion protocol and an application environment based on a browser. The application protocol is a communication protocol stack that is embedded in each WAP-enabled wireless device (also known as the user agent). The server side implements the other end of the protocol, which is capable of communicating with any WAP client. The server side is known as a WAP gate-way and routes requests from the client to an HTTP (Hyper Text Transfer Protocol) (or Web) server. The WAP gateway can be located either in an Operator premises (Figure 1) or in WAP application provider premises with the web server (Figure 2). Figure 3 illustrates an example structure of a WAP network. In the WAP net-work the client communicates with the WAP gateway in the wireless net-work. The WAP gateway translates WAP requests to WWW requests, so the WAP client is able

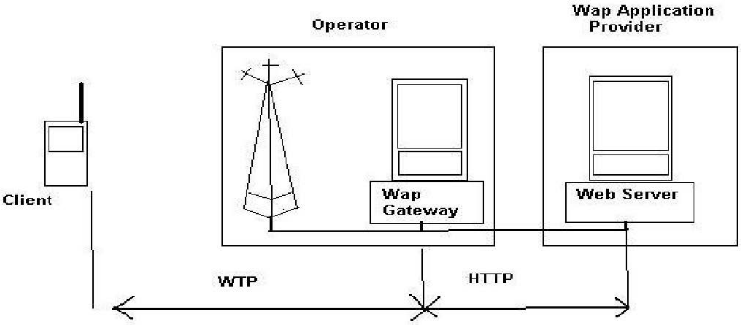


Fig. 1. Gateway equipment in the operator premises

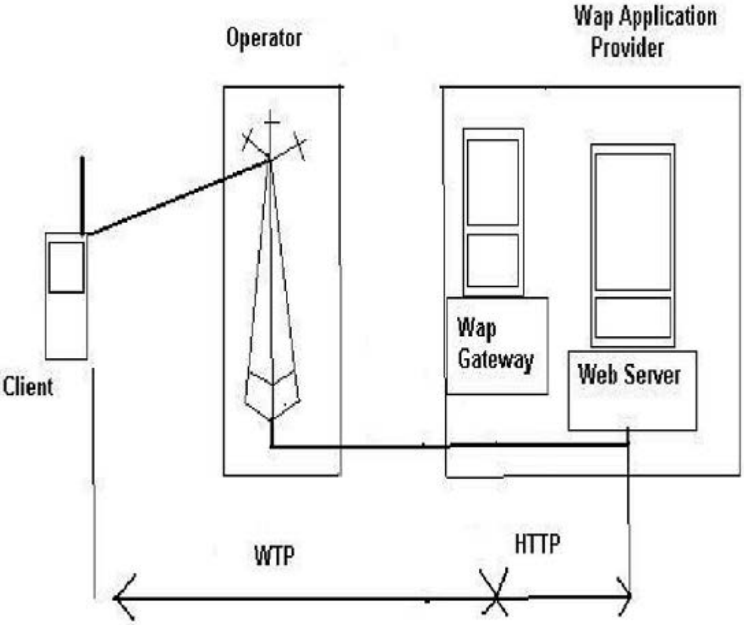


Fig. 2. WAP Gateway in the WAP application provider premises

to submit requests to the Web server. Also, the WAP gate-way translates Web responses into WAP responses or a format understood by the WAP client.

The wireless application environment provides WAP micro browser for interaction between WAP (web applications) and wireless devices. This browser relies on WAP Markup languages such as WML (Wireless Markup Language), WML Script and XHTML MP (Extensible Hypertext Markup Language Mobile Profile).

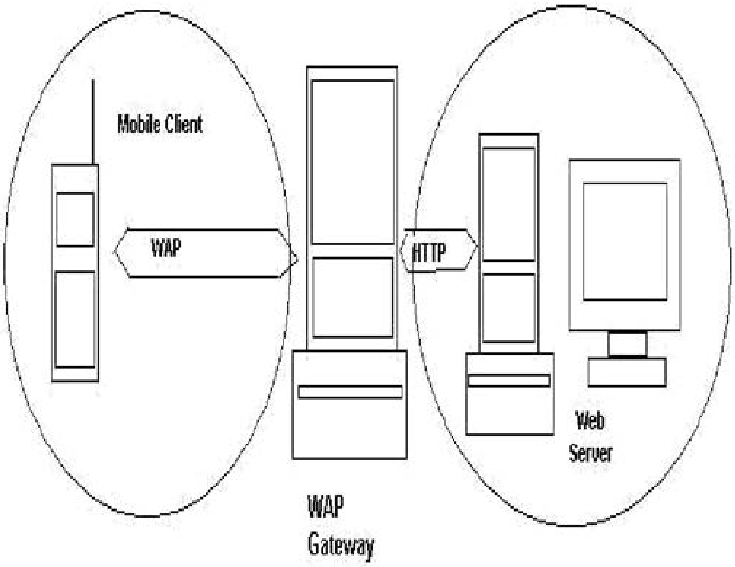


Fig. 3. The WAP network structure

2.3 WAP Programming Model

The WAP programming model is similar to the Web programming model with matching extensions, but it accommodates the characteristics of the wireless environment. The WAP programming model is based heavily on the Web programming model. But how does the WAP gateway work with HTML (Hyper Text Markup Language)? In some cases, the data services or content located on the Web server is HTML-based. Some WAP gateways could be made to convert HTML pages into a format that can bedisplayed on wireless devices. Be- cause HTML was not really designed for small screens, the WAP protocol de-fines its own markup language.

WML, WML Script, and XHTML MP are the languages that are specifically de-signed to develop WAP applications for Mobile devices. These languages adhere to the XML standard and are designed to enable powerful applica-tions within the constraints of handheld devices . In most cases, the actual application or other content located on the Web server will be native WAP contents created with WML (XHTML MP) or generated dynamically usingWML Script, Java Servlets or JSP (Java Server Page), or other server side programming languages. WML is an XML-based markup language that was designed especially to present WAP content on a wireless terminal. WML can preserve the content of variables between different WML pages. The basic unit of WML is the card that specifies a single interaction between the user and the user agent.

Multiple cards are grouped together in decks, which is the top most element of a WML file. When the user agent receives a deck, it activates only the first card in the deck. There are no functions to check the validity of user input or to generate messages and dialog boxes locally in using WML. Therefore, to overcome this limitation, WML Script was developed.

WML Script, which is based on ECMA Script (the standard for java script), is a language that can be used to provide programmed functionality to WAP applications. It was defined to enable the execution of scripts on WAP devices. The goal of using WML Script is to reduce the number of turn around between the client and the server. It is part of the WAP specification, and it can be used to add script support to the client. Its difference from ECMA Script is that it is compiled into byte code before it is sent to the client. The main reason for this is to cope up with the narrowband communication chan nels and to keep client memory requirements to a minimum. XHTML is a markup language used to create richer web content on an ever increasing range of platforms including mobile handsets. It is similar with HTML in its tag definition and syntax, but it adds modularity and enforces strict adherence to language rules. It brings a clear structure to web pages, which is especially important for the small screens and limited power of mobile devices. The XHTML MP is a mobile adaptation of XHTML by excluding those features not appropriate for devices with small screens. It is a strict subset of XHTML that includes additional elements and attributes that are useful in mobile browsers with additional presentation elements and support for internal style sheets.

Mobile browsing technology is evolving from WAP 1.x to WAP 2.0, by introducing different enhancements for mobile content development. Espe-cially WAP 2.0 provides support for protocols such as IP, TCP and HTTP. This

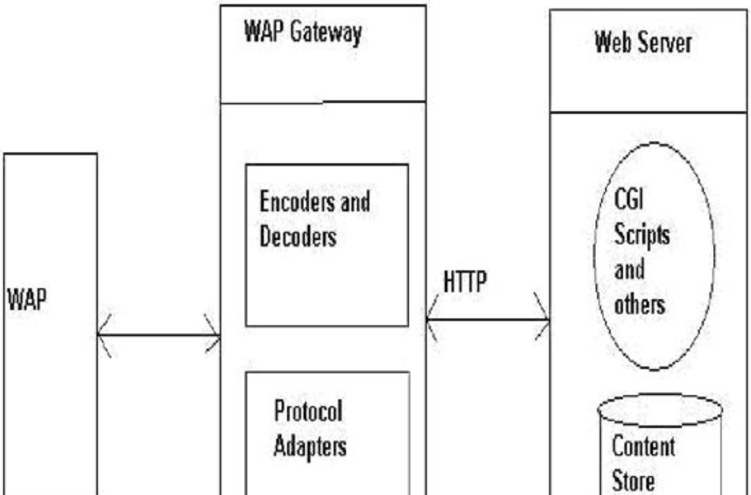


Fig. 4. The WAP programming model

provides interoperable optimizations suitable to the wireless environment and to the environment that permits wireless devices to utilize existing Internet technologies. WAP 2.0 also provides different application environ-ment, which enables delivery of information and interactive services to wireless devices.

WAP standard defines the future of wireless browsing technology based on the WML, XHTML MP and WAP CSS (WAP Cascading Style Sheet). Both WML and XHTML MP are a reformulation of the XML. XML is a language for marking up structures in text documents and supports the UTF-8 (8 bit Unicode Trans-formation Format) coding standard. The UTF-8 coding standard supports sev-eral languages character set including Ethiopic. So WML and XHTML MP can be used to create WAP pages that are encoded as UTF-8. Browsing from wireless terminals supporting UTF-8 encoding becomes possible.

2.4 WAP Communication Model

WAP contents usually reside on WWW servers on the Internet. The WAP gateway placed between the mobile network and the web servers. It receives WAP requests using the binary WAP communication protocols, and translates the requests to the text based WWW protocols and forwards to the content servers using the TCP/IP network protocol[7].

The WAP gateway also waits for the WWW text protocol reply from the content server, and receives using the TCP/IP protocol. It formats the message to binary WAP protocol, and then sends the reformatted response to the WAP client via WDP (WAP Datagram Protocol), which is almost equivalent to UDP (User Datagram Protocol) of the Internet. WDP makes no attempt to confirm delivery, resend lost packets, or correct errors in transmission like the UDP.

The WAP protocols are designed to operate over a variety of different bearer services, including short messages, circuit-switched data and packet data. Each bearer offers different levels of quality of service with respect to throughput, error rate and delays. The WAP protocols are designed to compen-sate for or tolerate these varying levels of service.The WDP specification lists the bearers that are supported and techniques used to allow WAP protocols to run over each bearer.

3 WAP Services

WAP provides two types of services, which are pull and push (Figure 5). In pull service, the user can request and fetch for the WAP site, for the informa-tion that he needs to browse. In the Push service, origin server or PI (Push Ini-tiator) initiates the connection and sends the push messages to the mobile de-vice. The push service can be, for example, messaging, stock price and traffic update alerts.

As shown in Figure 5, the client requests a server for a response in a pull ser-vice, and a server initiates a connection with a client for a push service.

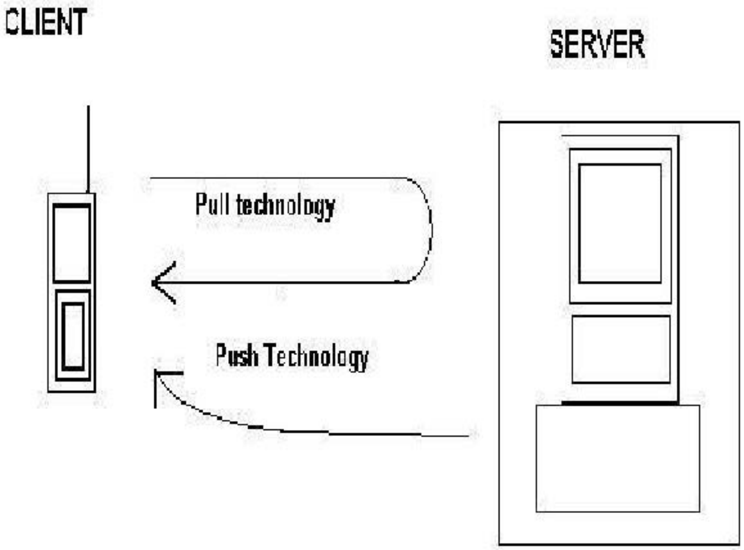


Fig. 5. The structure of a WAP pull and push service

3.1 Pull Service

WAP pull is a traditional WAP service. It works similar to the normal client/server model. In the normal client/server model, a client requests a service or information from a server, which then the server responds by transmitting

information to the client. This is known as ”pull” technology. That is, the client pulls information from the server. The WWW is the best example of pull technology, where a user enters a URL (Uniform Resource Locator) (the request), which is sent to a server, and the server answers by sending a web page (the response) to the user. Similarly, inWAP pull, the user agent requests content from a WAP server. Then the WAP server reply through the WAP proxy to the user agent and the user agent can access the content[3].

When developing an application, which has connection with a database server, a pull service of WAP is essential. For example banking (transfer money between accounts), finance (buy and sell stocks, exchange rate), shopping (buy books, searching for record), and Ticketing (cinema tickets, concert tickets) are some of the applications, which can be developed using the pull technology of WAP.

3.2 Push Service

WAP push, available since WAP 1.2, allows WAP content to be pushed to the mobile handset with minimum user intervention. A WAP push is basically a spe-cially encoded message that includes a link to a WAP address. It can be deliv-ered over WAP or SMS (Simple Message Service) bearer. On receiving

WAP push messages, WAP enabled handsets can automatically give the user the op-tion to access the WAP content.

The push functionality is especially relevant to real-time applications that send notifications to users. Without the push functionality applications would require the devices to poll application servers for new information or status. In wireless environments such polling activities would constitute inefficient and wasteful use of resources of wireless networks. WAP's push functionality provides control over the lifetime of pushed messages, store and forward capabilities at the push proxy and control over bearer choice for deliv-ery.

The WAP push service directs the end user to a WAP address where particular content may be stored, which is ready for viewing or downloading to the handset that enhances usability[5]. Operators and content providers can utilize push data transfer to deliver content that is relevant to user groups, instead of relying on the traditional pull model. WAP push advances existing messaging models (SMS, Smart Messaging) by integrating them into the WAP application environment. In consumer segment, users can subscribe to content they are truly interested in or to receive push messages that add value (promotions and discounts). The push message can inform them to follow links to more details or to complete a transaction. In a corporate segment, the company can send important news and users can access vital, time-critical information easily.

A push operation in WAP is accomplished by allowing a PI to transmit push con-tent and delivery instructions to a PPG (Push Proxy Gateway), which then delivers the push content to the WAP client. The PI is typically an application that runs on an ordinary web server. It communicates with the PPG using the Push Access Protocol (PAP). The PPG uses the push Over-the-Air (OTA) protocol to deliver the push content to the client. The architecture of push can be thought of similarly to that of SMS, except that with push, the PPG is found in the middle rather than the SMSC. Push can be sent over SMS or GPRS.

4 WAP Gateways

WAP gateway acts as mediator between Cellular device and HTTP or HTTPS web server. WAP gateway routes requests from the client (Cellular Phones) to an HTTP (or Web) server. The WAP gateway can be located either in a telecom network or in a computer network.

4.1 Kannel

Kannel[2] is an open source WAP gateway. It attempts to provide this essential part of the WAP infrastructure freely to everyone so that the market potential for WAP services, both from wireless operators and specialized service provid-ers, will be realized as efficiently as possible. Kennel's quality has been recognized on March 7, 2001 when it was certified by WAP Forum as the first WAP 1.1 gateway in the world. Greater quality recognition are the quantity of companies using Kannel to successful connect to a variety of SMSC protocols in lots of countries.

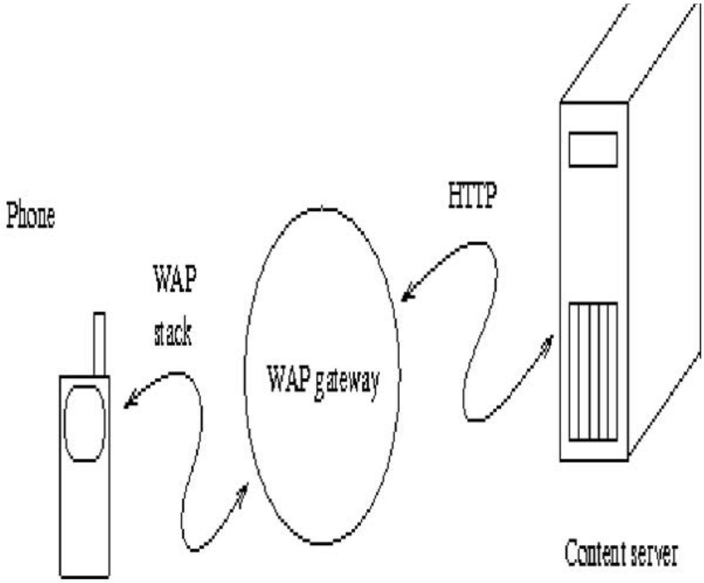


Fig. 6. Logical positions of WAP gateway (and PPG) between a phone and a content server

Kannel gateway architecture:

The gateway divides the processing load between the following two hosts:

I. The bearer box, which connects to the SMS (Short Message Service) centers and CSD (Circuit Switched Data) routers,providing a unified interface tothem for the wapbox. The bearer box does this by imple-menting the WDP (Wireless Datagram Protocol) layer of the WAP stack.

II. The wapbox, which hosts the upper layers of the WAP stack. Each ses-sion and its transactions are handled by the same wapbox.

Working of the system:

The bearer box receives UDP (User Datagram Protocol) packets from CSD rout-ers, inspects them to see whether they are WAP packets, and then routes themto the WAP box. This simple design allows the bearer box to do a minimum of processing per packet. The bearer box also sends the UDP packets that the otherboxes generate, which adds some more routing processing. The wapboxes implement the WTP (Wireless Transaction Protocol) and WSP (Wireless SessionProtocol) layers. These take HTTP-like requests from the phones and make the actual HTTP requests to content servers, compress the respons-es, and sendthem back to the terminals. (Sessions are maintained to make as much use of the limited radio bandwidth as possible.)

4.2 Nokia Mobile Internet Toolkit

Nokia Mobile Internet Toolkit (NMIT) [5] consists of a set of editors that you can use to learn how to create various types of mobile Internet content. NMIT lets you display this content on multiple phone SDKs.Phone SDKs are installed separately. NMIT detects installed, supported phone SDKs at startup and lists these in its SDK Control Panel. You can display content you author on any sup-ported phone SDK by simply clicking a Show button within an editor. Many NMIT editors are used for creating XML-based content types defined by Docu-ment Type Definitions (DTDs). These editors employ content validation to check content against a DTD, and they provide features for easily selecting elements and attributions for insertion based on current cursor position. In addition, NMIT provides a DTD Manager through which you can import new DTDs for use by NMIT editors[4] . NMIT is a software simulation for mobile, integrated with WAP server. It is gen-erally used to test WAP sites. We also used it to test WAP content.Major features of NMIT include:

" Browsing Editors " Push Content Editors " Messaging Editors " Deployment Editors

4.3 WAP

WAP Proxy is a high performance WAP Gateway that is designed to meet the needs of WAP 2.0 and multimedia applications, especially MMS (Multimedia Messaging Service) and Java downloads. With WAP providing the underlying protocol support for multimedia object delivery to mobile clients[8]. The explosive growth of MMS services is placing heavy demands on conventional WAP gate-ways. WAP Proxy was designed to meet the needs of new MMS services, as well as legacy WAP applications. Now WAP Proxy provides full WAP gateway support for WAP v1.1, v1.2, v1.2.1, v1.3 and v2.0 WAP clients. WAP Proxy can support WAP 2.0 clients using either Wireless Profiled HTTP/TCP (W-HTTP and W-TCP), or the WSP protocol stack[6] .

WAP is proprietary software, designed to run on Windows platform. Since it's not free and a trial version of one month is available. We tested trial version to host WAP content.

5 Conclusion

Our designed system is able to solve the issues of integration related with open source WAP Gateways through service based interfaces. The orchestration of services is done by using Service oriented architecture approach. We tested that the content created by NMIT kit supports the various SDK's of mobile phones of different vendors by using XML based content configuration, which makes it portable across different vendor specific devices.

References

1. Mehta, N.: Mobile Web Development. Packt Publishing (2008) ISBN: 1847193439
2. Wizenius, L.: KannelArchitecture and Design. A Research Paper (March 2000)
3. Evans, C., Bilal, M.: Developing a WAP Application for Mobile Retail Customers. In: 2nd International Conference on RPervasive Computing and Applications, ICPCA 2007, July 26-27, pp. 328–332 (2007)
4. He, F., Fan, J., Fu, X.: Research and application of a WAP-based mobile learning system. In: IEEE International Conference on Communications Technology and Applications, ICCTA, October 16-18, pp. 864–868 (2009)
5. Yue, H.: Research and development of the mobile library based on WAP technology. In: 2010 2nd International Conference on Industrial Mechatronics and Automation, ICIMA, pp. 191–194 (May 2010)
6. Reynolds, F.: Web 2.0-In Your Hand. IEEE Pervasive Computing 8(1), 86–88 (2009)
7. Sierra, K., Bates, B.: Head First JSP Servlet, 3rd edn., pp. 162–359. Orelly Publication
8. Lin, T.-H., Wang, K.: An efficient load balancing strategy for scalable WAP gateways. In: Proceedings of the Ninth International Conference on Parallel and Distributed Systems, December 17-20, pp. 625–630 (2002)

An Improved Method for Contrast Enhancement of Real World Hyperspectral Images

Shyam Lal[1], Rahul Kumar[1], and Mahesh Chandra[2]

[1] ECE Department, Moradabad Institute of Technology, Moradabad (U.P.), India
[2] ECE Department, Birla Institute of Technology, Mesra, Ranchi (J.H.), India
{shyam.mtec,rrkmalik}@gmail.com, shrotriya@bitmesra.ac.in

Abstract. This paper proposed an improved method for contrast enhancement of real world hyperspectral images**.** The proposed method consists of two stages: In first stage the poor quality of image is processed by adaptive histogram equalization in spatial domain and in second stage the output of first stage is further processed by adaptive filtering for image enhancement in frequency domain. Simulation and experimental results on benchmark real world hyperspectral image database demonstrates that proposed method provides better results as compared to other state-of-art contrast enhancement techniques such as alpha rooting (AR), multi contrast enhancement (MCE), multi-contrast enhancement with dynamic range compression (MCEDRC), brightness preserving dynamic fuzzy histogram equalization (BPDFHE). Proposed method performs better for different dark and bright real world hyperspectral images by adjusting their contrast very frequently. Proposed method is simple and efficient approach for contrast enhancement of real world hyperspectral images. This method can be used in different applications where images are suffering from various contrast problems.

Keywords: Adaptive Histogram Equalization, real world hyperspectral image, Image processing, Contrast Enhancement.

1 Introduction

The contrast enhancement techniques are commonly used in various applications where subjective quality of image is very important. The objective of image enhancement is to improve visual quality of image depending on the application circumstances. Contrast is an important factor for any individual estimation of image quality. It can be used as controlling tool for documenting and presenting information collected during examination.

The contrast enhancement of image refers to the amount of color or gray differentiation that exists between various features in digital images. It is the range of the brightness present in the image. The images having a higher contrast level usually display a larger degree of color or gray scale difference as compared to lower contrast level. The contrast enhancement is a process that allows image features to show up more visibly by making best use of the color presented on the display devices.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 273–282, 2013.

During last decade a number of contrast enhancement algorithms have been developed for contrast enhancement of images for various applications. Histogram equalization [1], global histogram equalization [2], local histogram equalization [3], adaptive histogram equalization and contrast limited adaptive histogram equalization [4]-[5], other histogram equalization based algorithms [6]-[14] and other contrast enhancement methods [15][16] have been proposed by various researchers. One of the most widely used algorithms is global histogram equalization, the basic idea of which is to adjust the intensity histogram to approximate a uniform distribution. It treats all regions of the image equally and, thus often yields poor local performance in terms of detail preservation of image.

The outline of this paper is as follows. Section 2 describes literature review. Section 3 describes proposed method for contrast enhancement of real world hyperspectral images. Section 4 gives simulation results and discussions to demonstrate the performance of proposed method. Finally, conclusion is drawn in section 5.

2 Literature Review

The existing contrast enhancement techniques for mobile communication and other real time applications is fall under two broad categories that is contrast shaping based methods and histogram equalization based methods [1]. These methods are derived from digital image processing. These methods may lead to over-enhancement and other artifacts such as flickering, and contouring. The contrast shaping based methods are work by calculating an input-output luminance curve defined at every luminance level. The shape of the curve must depend on the statistics of the image frame being processed. For example, dark images would have a dark stretch curve applied to them. Although contrasts shaping based methods are the most popular methods used in the consumer electronics industry but they cannot provide a localized contrast enhancement which is desirable. For example, when a dark stretch is performed, bright pixels become brighter. However, a better way to enhance darker images is to stretch and enhance the dark regions, while leaving brighter pixels untouched [1, 17].

A very popular technique for contrast enhancement of image is histogram equalization technique [3, 1, 2]. A histogram equalization is a technique that generates gray map which change the histogram of image and redistributing all pixel values to be as close as possible to user specified desired histogram. This technique is useful for processing images that have little contrast with equal number of pixels to each the output gray levels. The histogram equalization (HE) is a method to obtain a unique input to output contrast transfer function based on the histogram of the input image which results in a contrast transfer curve that stretches the peaks of the histogram (where more information is present) and compresses the troughs of the histogram (where less information is present) [1]. Therefore it is a special case of contrast shaping technique. As a standalone technique, histogram equalization is used extensively in medical imaging, satellite imagery and other applications where the emphasis is on pattern recognition and bringing out of hidden details. Thus histogram equalization results in too much enhancement and artifacts like contouring which is unacceptable in consumer electronics [4, 5]. During last decade a number of techniques have been proposed by various researchers to deal with these problems. In [10], the histogram is divided into two parts based on the input mean and each part is equalized, separately. This preserves the mean value of image to a certain extent. In [7], each peak of the histogram is equalized separately. An adaptation of HE, termed as contrast limited adaptive histogram equalization (CLAHE) [5] divides the input image into a number of equal sized blocks and then performs contrast limited histogram equalization on each block. The contrast limiting is done by clipping the histogram before histogram equalization. This tends to tone down the over enhancement effect of histogram equalization and gives a more localized enhancement. However it is much more computationally intensive than histogram equalization. If the blocks are non-overlapping, an interpolation scheme is needed to prevent blocky artifacts in the output picture. Therefore overlapping blocks can solve this problem (every pixel is replaced by the histogram equalization output using a neighborhood) but it is more computationally intensive than using non-overlapping blocks. So the CLAHE also requires a field store. Finally one more contrast enhancement method that is homomorphic filter is proposed in spatial domain [1]. In this filter images normally consist of light reflected from objects. The basic nature of the image may be characterized by two components: (1) the amount of source light incident on the scene being viewed, and (2) the amount of light reflected by the objects in the scene but this method does not provide good image quality [4]. Another method is histogram specification (HS) which takes a desired histogram by which the expected output image histogram can be controlled [1]. However specifying the output histogram is not a smooth task as it varies from image to image.

Since past few years various researchers have also focused on improvement of histogram equalization based contrast enhancement techniques such as mean preserving bi-histogram equalization (BBHE) [10], dualistic sub-image histogram equalization (DSIHE) [14] and minimum mean brightness error bi-histogram equalization (MMBEBHE) [18]. The BBHE separates the input image histogram into two parts based on input mean. After separation, each part is equalized independently. This method tries to overcome the brightness preservation problem. The DSIHE method uses entropy value for histogram separation. The MMBEBHE is the extension of BBHE method that provides maximal brightness preservation. Though these methods can perform good contrast enhancement, but they also cause more annoying side effects depending on the variation of gray level distribution in the histogram. Therefore, recursive mean-separate histogram equalization (RMSHE) [8] is proposed which provides better contrast results over BBHE. This algorithm is the improvement in BBHE. However, it also has some side effects. In [15] Hassan and Norio is proposed new approach for contrast enhancement using sigmoid function. The objective of this new contrast enhancer is to scale the input image by using sigmoid function. However this method is also having some side effects. In order to improve the performance of above mentioned algorithm, exact histogram specification (EHS) [9] is used for contrast enhancement of images. In order to provide better result another technique named as brightness preserving dynamic fuzzy histogram equalization (BPDFHE) is proposed [12]. This technique is the modification of the brightness preserving dynamic histogram equalization technique to improve its brightness preserving and contrast enhancement abilities while reducing its computational complexity. This technique uses fuzzy statistics of digital images for their representation and processing. Therefore, representation and processing of images in the fuzzy domain enables the technique to handle the inexactness of gray level values in a better way which results in improved performance. In [15] Celik and Jahjadi proposed contextual and variational contrast enhancement for image. This algorithm enhances the contrast of an input image using interpixel contextual information. This algorithm uses a 2-D histogram of the input image constructed using a mutual relationship between each pixel and its neighboring pixels. A smooth 2-D target histogram is obtained by minimizing the sum of frobenius norms of the differences from the input histogram and the uniformly distributed histogram. The enhancement is achieved by mapping the diagonal elements of the input histogram to the diagonal elements of the target histogram. This algorithm produces better enhanced image results as compared to other existing state-of-the-art algorithms. On the other hand various researchers also proposed many algorithms for contrast enhancement in DCT based compressed domain such as alpha rooting (AR) [6], multi contrast enhancement (MCE) [13], Multi contrast enhancement with dynamic range compression (MCEDRC) [11] and wavelet based domain (ACEWD) [19].

3 Proposed Method

The proposed method consists of two stages: In first stage the poor quality of image is process by adaptive histogram equalization in spatial domain and in second stage the output of first stage is further processed by adaptive filtering for image enhancement in frequency domain. The proposed method is abbreviated as Adaptive Contrast Enhancement Based on Histogram Equalization (ACEBHE). The model of proposed method is shown in Fig. 1. A two dimensional original image is denoted by a function f(x, y). The amplitude of ‘f’ at spatial coordinates (x, y) is a positive scalar quantity whose physical meaning is determined by the source of image. When an image is generated from a physical process, its values are proportional to energy radiated by a physical source such as electromagnetic waves and infrared waves. As a consequence, f(x, y) must be non zero and finite. The two dimensional function f(x, y) may be characterized by two components: (i) The amount of source illumination incident on the scene being viewed (ii) The amount of illumination reflected by objects in scene. First one is called illumination component and it is denoted by i(x, y) and second one is called reflectance component and it is denoted by r(x, y).

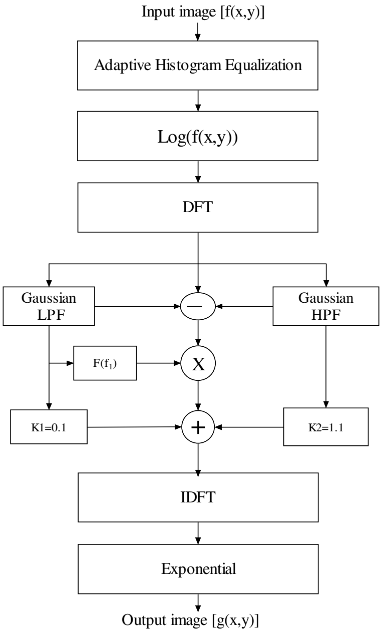


Fig. 1. Block diagram of proposed Adaptive Contrast Enhancement Based on Histogram Equalization

These two components are combined as a product to form two dimensional function f(x,y). Therefore it is given by

$$f(x,y)=i(x,y)*r(x,y)$$

The nature of i(x,y) is determined by illumination source, and r(x,y) is determined by the characteristics of the imaged objects. The function f(x,y) cannot be used directly to operate separately on the frequency components of illumination and reflectance because the Fourier transform of the product of two function is not separable. However if we define

$$z(x,y)=\ln[f(x,y)]=\ln[i(x,y)]+\ln[r(x,y)]$$

Then

$$F\{z((x,y)\}=F\{\ln[f(x,y)]\}=F\{\ln[i(x,y)\}+F\{ \ln[r(x,y)]\}$$

The illumination component of two dimensional images is characterized by slow spatial variation, while the reflectance component tends to vary abruptly, particularly at the junction of dissimilar components. These characteristics lead to associating low frequencies of the Fourier transform of the logarithm of an image with illumination and the high frequencies with reflectance. A good deal of control can be gained over the illumination and reflectance components by defining a filter function that affects low and high frequency components of the Fourier transform in different ways. The filter function should be such that it tends to decrease the contribution made by the low frequencies (illumination) and amplify the contribution made by high frequencies (reflectance). The net result is simultaneous dynamic range compression and contrast enhancement. In our method we have suppressed the low frequency components by 90% and increased the clearly visible high frequency components by 110%. Therefore the hidden frequency components are locally enhanced depending on illumination of that particular region. The hidden frequency components are convolved with the function F(f_l) where F is defined as

$$F(f_l)=1+kf_l$$

Where the value of k is different for each 17x17 block

Now the modified high frequency components, low frequency components, and hidden frequency components are added together to give new enhanced two dimensional images in the frequency domain. After that inverse discrete Fourier transform is taken to get the enhanced image in spatial domain. Finally, as z(x,y) was formed by taking the logarithm of the original image f(x,y), the inverse (exponential) operation yields the desired new enhanced image.

3.1 Implementation of Proposed Method

Step(1): Read the input image.
Step(2): Convert input image into gray scale image if it is color image.
Step(3): Apply adaptive histogram equalization algorithm.
Step(4): Find DFT of output from Step(3)
Step(5): Perform adaptive filtering for image enhancement.
Step(6): Find IDFT after adaptive filtering operation.
Step(7): Find exponential of IDFT from Step(6)

4 Simulation Results and Discussions

In order to demonstrates the performance of proposed ACEBHE method, it is tested on different gray scale real world Hyperspectral images with dimension Ml×M2 (=512×512) [20]. In order to obtain simulation and experimental results of proposed ACEBHE method and other existing algorithms, MATLAB software (MATLAB 7.6, release 2008a) is used. Therefore, two experiments have been conducted on different gray scale real world hyperspectral images. In the first experiment the quality metrics is presented and in the second experiment visual quality of image has been presented. In order to judge the performance of proposed ACEBHE method the quality parameters such as measure of enhancement (EME) and measure of enhancement factor (EMF) are the automatic choice for the researchers. Therefore, a better value of EME and EMF implies that the visual quality of the enhanced image is good. The measure of enhancement (EME) and measure of enhancement factor (EMF) are defined in equation (3) and equation (4) respectively for gray scale real world

Hyperspectral images. These image quality metrics are used to compare the performance of proposed ACEBHE method and other existing contrast enhancement techniques such as Alpha Rooting (AR) [6], Multi contrast enhancement (MCE) [13], Multi-contrast enhancement with dynamic range compression (MCEDRC) [11], Brightness preserving dynamic fuzzy histogram equalization (BPDFHE) [12]. The real world hyperspectral images used for experimental tests are available on the website http://vision.seas.harvard.edu/hyperspec/explorei.html

The measure of enhancement (EME) [17, 21] of image I(i, j) with dimensions MlxM2 pixels is defined as:

$$EME_{k_1 k_2} = \frac{1}{k_1 k_2} \sum_{l=1}^{k_1} \sum_{k=1}^{k_2} \left[20 * ln \left(\frac{I_{max,k,l}}{I_{min,k,l}} \right) \right] \tag{3}$$

where an image (I) is divided into $k_1 \times k_2$ blocks, $I_{max,k,l}$ and $I_{min,k,l}$ are the maximum and minimum values of the pixels in each block.

The measure of enhancement factor (EMF) between output image and input image is defined as:

$$EMF = \frac{\text{EME of output image}}{\text{EME of input image}} \tag{4}$$

4.1 Experiment 1

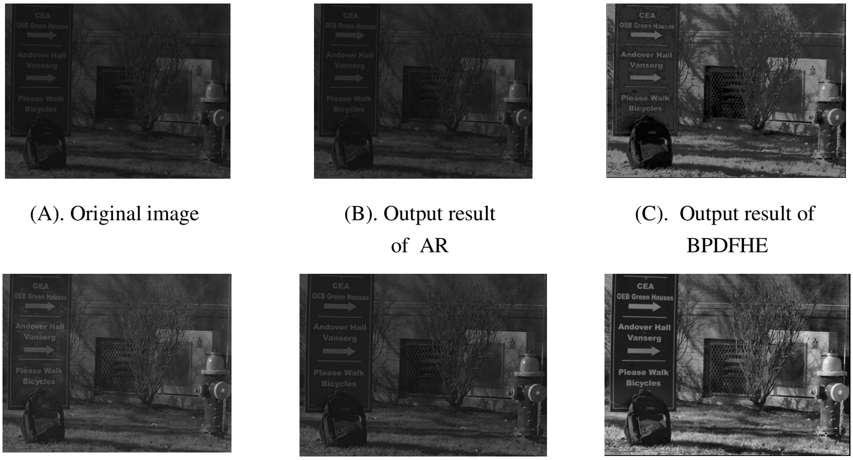
In this experiment the performance of proposed ACEBHE method is tested on different gray scale real world Hyperspectral images. The performance of proposed ACEBHE method and many existing contrast enhancement techniques has been evaluated for image1 and image2 in terms of quality parameters such as measure of enhancement (EME) and measure of enhancement factor (EMF). For image1, and image2, the performance of proposed ACEBHE method has been compared with many existing contrast enhancement techniques. The measure of enhancement (EME), measure of enhancement factor (EMF) and CPU processing time of proposed ACEBHE method and many existing contrast enhancement techniques for image1, and image2 have been given in Table 1. It can be noticed from Table 1 that the proposed ACEBHE method provides better results as compared to other state-of-art contrast enhancement techniques.

4.2 Experiment 2

In order to perform the superiority of proposed ACEBHE method another experiment has been conducted on different gray scale real world Hyperspectral images. This experiment visualizes subjective image enhancement performance, the enhanced contrast of image1 and image2 have been compared with result of proposed ACEBHE method and many existing contrast enhancement techniques. The visual contrast enhancement results of proposed ACEBHE method and many existing contrast enhancement techniques have been given from Figure 2 to Figure 3. Therefore, it can be noticed from Figure 2(B) to Figure 2(F), and Figure 3(B) to Figure 3(F) that proposed ACEBHE method gives better contrast enhancement results as compared to other existing contrast enhancement techniques.

Table 1. Comparative performance of different methods and gray-scale image

Method / Parameters	AR	BPDFHE	MCEDRC	MCE	PROPOSED ACEBHE
image1.tif					
EME(Original)	10.22	10.22	10.22	10.22	**10.22**
EME(Output)	10.88	17.77	10.64	12.84	**20.12**
EMF	1.06	1.74	1.04	1.26	**1.97**
CPU Time (second)	0.38	0.23	1.75	0.38	**2.74**
image2.tif					
EME(Original)	2.47	2.47	2.47	2.47	**2.47**
EME(Output)	2.77	3.36	2.52	3.02	**4.89**
EMF	1.12	1.36	1.02	1.22	**1.98**
CPU Time (second)	0.38	0.19	1.77	0.38	**0.85**



(A). Original image

(B). Output result of AR

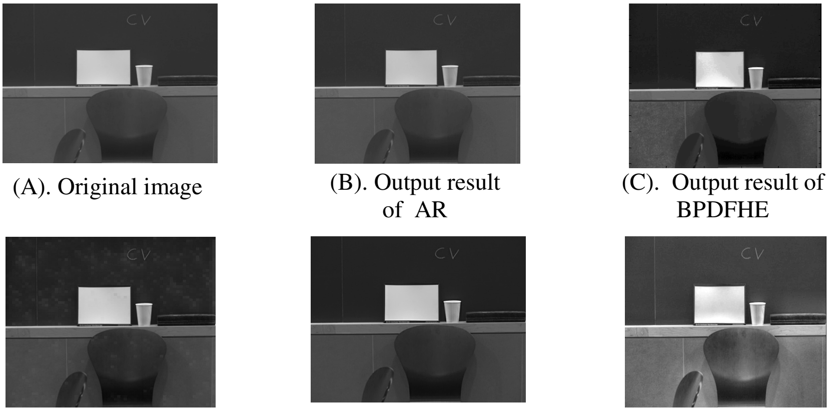
(C). Output result of BPDFHE

(D). Output result of MCEDRC

(E). Output result of MCE

(F). Output result of ACEBHE Method

Fig. 2. Visual Enhancement results of different algorithms for image1. tif



(A). Original image

(B). Output result of AR

(C). Output result of BPDFHE

(D). Output result of MCEDRC

(E). Output result of MCE

(F). Output result of ACEBHE Method

Fig. 3. Visual Enhancement results of different algorithms for image2.tif

5 Conclusion

In this paper an improved contrast enhancement method was proposed for image enhancement purpose for various applications. This method was tested on different gray scale real world Hyperspectral images. The qualitative and subjective enhancement performance of proposed ACEBHE method was evaluated and compared to other state-of-art contrast enhancement techniques. The performance of proposed ACEBHE method was evaluated and compared in terms of EME, EMF and Execution time. The simulation results demonstrated that the proposed ACEBHE method provided better results as compared to other state-of-art contrast enhancement techniques for different gray scale real world Hyperspectral images. The visual enhancement results of proposed ACEBHE method were also better as compared to other state-of-art contrast enhancement techniques. Therefore, proposed ACEBHE method performed very effectively for contrast enhancement of gray scale real world Hyperspectral images. The proposed ACEBHE method can also be used for many other images such as remote sensing images, electron microscopy images and even real life photographic pictures suffer from poor contrast problems during its acquisition.

References

[1] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Addison-Wesley Publishing Company (1992)

[2] Stark, J.A.: Adaptive Contrast Enhancement Using Generalization of Histogram Equalization. IEEE Transactions on Image Processing 9(5), 889–906 (2000)

[3] Caselles, V., Lisani, J.L., Morel, J.M., Sapiro, G.: Shape Preserving Local Histogram Modification. IEEE Transactions on Image Processing 8(2), 220–230 (1998)

[4] Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer Vision, Graphics and Image Processing 39(3), 355–368 (1987)

[5] ZuiderveldK: Graphics Gems IV. In: Contrast Limited Adaptive Histogram Equalization, vol. 5, ch. VIII, pp. 474–485. Academic Press, Cambridge (1994)

[6] Aghagolzadeh, S., Ersoy, O.K.: Transform Image Enhancement. Optical Engineering 31, 614–626 (1992)

[7] Chen, S.D., Ramli, A.R.: Preserving Brightness in Histogram Equalization Based Contrast Enhancement Techniques. Digital Signal Processing 14(5), 413–428 (2004)

[8] Chen, S.D., Ramli, A.R.: Contrast Enhancement Using Recursive Mean-Separate Histogram Equalization For Scalable Brightness Preservation. IEEE Transactions on Consumer Electronics 49(4), 1301–1309 (2003)

[9] Coltuc, D., Bolon, P., Chassery, J.M.: Exact Histogram Specification. IEEE Transactions on Image Processing 15(5), 1143–1151 (2006)

[10] Kim, Y.T.: Contrast Enhancement Using Brightness Preserving Bi-histogram Equalization. IEEE Transactions on Consumer Electronics 43(1), 1–8 (1997)

[11] Lee, S.: An Efficient Content-Based Image Enhancement In The Compressed Domain Using Retinex Theory. IEEE Transactions on Circuits Systems and Video Technology 17(2), 199–213 (2007)

[12] Sheet, D., Garud, H., Suveer, A., Mahadevappa, A.M., Chatterjee, J.: Brightness Preserving Dynamic Fuzzy Histogram Equalization. IEEE Transactions on Consumer Electronics 56(4), 2475–2480 (2010)

[13] Tang, J., Peli, E., Acton, S.: Image Enhancement Using A Contrast Measure In the Compressed Domain. IEEE Signal Processing Letter 10(10), 289–292 (2003)

[14] Wang, Y., Chen, Q., Zhang, B.: Image Enhancement Based on Equal Area Dualistic Sub-Image Histogram Equalization Method. IEEE Transactions on Consumer Electronics 45(1), 68–75 (1999)

[15] Celik, T., Tjahjadi, T.: Contextual and Variational Contrast Enhancement. IEEE Transactions on Image Processing 20(12), 3431–3441 (2011)

[16] Hassan, N., Akamatsu, N.: A New Approach For Contrast Enhancement Using Sigmoid Function. The International Arab Journal of Information Technology 1(2), 221–225 (2004)

[17] Agaian, S., Silver, B., Panetta, K.: Transform Coefficient Histogram-Based Image Enhancement Algorithms Using Contrast Entropy. IEEE Transactions on Image Processing 16(3), 741–757 (2007)

[18] Chen, S.D., Ramli, A.R.: Minimum Mean Brightness Error Bi-Histogram Equalization in Contrast Enhancement. IEEE Transactions on Image Processing 49(4), 1310–1319 (2003)

[19] Lal, S., Chandra, M., Upadhyay, G.K.: Contrast Enhancement of Compressed Image in Wavelet Based Domain. In: The Proceedings of International Conference on Signal Recent Advancements in Electrical Sciences, ICRAES 2010, Tiruchengonde (TN) India, January 08-09, pp. 479–489 (2010)

[20] Chakrabarti, A., Zickler, T.: Statistics of Real-World Hyperspectral Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 193–200 (2011)

[21] Panetta, K., Zhou, Y., Agaian, S., Jia, H.: Nonlinear Unsharp Masking for Mammogram Enhancement. IEEE Transactions on Information Technology in Biomedicine 15(6), 918–928 (2011)

Wavelet Analysis of Electrical Signals from Brain: The Electroencephalogram

Rohtash Dhiman, Priyanka, and J.S. Saini

D.C.R. University of Science & Technology, Murthal, Sonipat, Haryana, India
rohtash.k@gmail.com

Abstract. The Electroencephalogram (EEG) is a measure of neural activity and is used to study cognitive processes, physiology, and complex brain dynamics. The analysis and processing of EEG data and to extract information from it, is a difficult task. The EEG signals are non-stationary signals. So, only transformation of these signals from time to frequency domain does not serve the purpose, it is required to know the time domain information too associated with the frequency domain information. Wavelet transform is one such tool being used recently for such analysis of non-stationary signals like EEG. In this paper, wavelet packet decomposition of EEG signals is presented. Feature extraction from EEG signal is also introduced in this paper.

Keywords: Electroencephalogram (EEG), Wavelet transform (WT), Wavelet Packet decomposition (WPD).

1 Introduction

Richard Canton, an English physician, discovered electrical currents in the brain in 1875. German psychiatrist Hans Berger[11], in 1929, first recorded these currents, and named them as Electroencephalogram (EEG) [3]. EEG is the electrical signal due to electrical activity of neurons in human brain. The EEG signals are very helpful in studying the physiological, psychological aspects and complex dynamics of brain. The electrical activity of brain changes during different conditions like sleep, awake-state, coma, eye open, eye closed, corresponding to different physical activities of body and different physiological & psychological disorders. From EEG, different distinct patterns can be identified corresponding to these conditions, but even a trained and experienced neurologist is sometimes incapable of identifying all patterns corresponding to all conditions. With the advancement in computing technology, the processing and analysis of these signals have opened new avenues for researchers to use EEG signals for automated diagnosis, brain computer interface, brain controlled prosthesis and understanding complex brain dynamics. This type of analysis and processing of bio-signals had came into use since 1960's; this paved the way for providing means towards accurate and precise diagnosis by physicians [2]. Time domain analysis is always very difficult for feature extraction and classification [1]. So frequency domain approaches, such as Fourier Transform and Fast Fourier

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 283–289, 2013.

Transform etc., are used for analysis and processing of EEG signals. But in case of these transforms, only frequency domain information is provided by the transformed domain, which is insufficient information for processing of non-stationary signals like EEG. So wavelet transform is used for analysis and processing of non-stationary signals. In this paper, wavelet analysis is presented for EEG signals, which are non-stationary in nature. The features can be extracted from the coefficients of decomposition; by calculating some statistical and non statistical properties for each node of decomposition.

2 Wavelet Transforms

This section provides a primer on wavelet transforms. Mathematical transforms are applied to signals to extract further information and that too in forms amenable for further processing. Often signals intended to be processed are in time-domain, but in order to process them and extract some information of interest, the frequency domain transformation is required. Mathematical transforms translate the information of signals into different representations. For example, the Fourier Transform provides the information about how many frequency components are there in a signal but cannot provide information about at which time these frequencies occur, because time and frequency is viewed independently. To solve this problem, the Short Term Fourier Transform (STFT) introduced the idea of windows through which different parts of a signal are viewed. For a given window in time, the frequencies can be viewed. However Heisenberg's Uncertainty Principle states that, as the resolution of the signal improves in the time-domain, by zooming on different sections, the frequency domain resolution gets worse. Ideally, a method of multi-resolution is needed, which allows certain parts of the signal to be resolved well in time, and other parts to be resolved well in frequency. This multi-resolution is the basis of wavelet transforms. It provides the time frequency representation. A wavelet is a little part of a wave, a wave that is only non-zero in a small region. In wavelet analysis, rather than examining the entire signal through the same window, different parts of the wave are viewed through different sized windows. High frequency parts of the signal use a small window to give good time resolution while low frequency parts use a big window to get good frequency information. Hence, Wavelet transform is capable of providing the time and frequency information simultaneously.

Continuous as well as discrete forms of wavelet transforms are available. The continuous wavelet transform was developed as an alternative approach to the Short Time Fourier Transform (STFT), to overcome the resolution problem. The wavelet analysis is done in a similar way to the STFT analysis, in the sense that the signal is multiplied with a function and the transform is computed separately for different segments of the time domain signal.

$$CWT_x^{\psi}(\tau, s) = \Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t)\psi^* \left(\frac{t-\tau}{s}\right) dt$$

As seen in the above equation, the transformed signal is a function of two variables, τ and s, the translation and scale parameters, respectively. Ψ is the transforming function, and it is called the mother wavelet. The continuous wavelet transform is computed by changing the scale of the analysis window, shifting/translating the window in time, multiplying by the signal, and integrating over all times. In the discrete case, filters of different cutoff frequencies are used to analyze the signal at different scales. The signal is passed through a series of high pass filters to analyze the high frequencies, and then it is passed through a series of low pass filters to analyze the low frequencies. The resolution of the signal, which is a measure of the amount of detailed information in the signal, is changed by the filtering operations, and the scale is changed by up-sampling and down-sampling operations.

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k].h[n-k]$$

The sequence is denoted by x[n], where n is an integer. The procedure starts with passing this signal sequence through a half band digital lowpass filter with impulse response h[n]. Filtering a signal corresponds to the mathematical operation of convolution of the signal with the impulse response of the filter. The convolution operation in discrete time operates in this manner: a half band lowpass filter removes all frequencies that are above half of the highest frequency in the signal. The wavelet decomposition corresponds to passing of a signal through filters successively and designating the divided signal into detailed coefficients and approximate coefficients. At every level of decomposition, the data is filtered and then the approximation and detailed coefficients are produced from this filtered data. The same process is followed for further levels of decomposition. Suppose we have a signal S, after the decomposition we get A_1 approximation coefficient and D_1 Detailed coefficient. If we do the decomposition of the A_1 then we get AA_2 and DA_2 as the approximation and detailed coefficients of A_1. The same procedure is followed for D_1 and so on. In this way, a decomposition tree can be obtained for any number of decomposition levels, as shown in Fig.1.

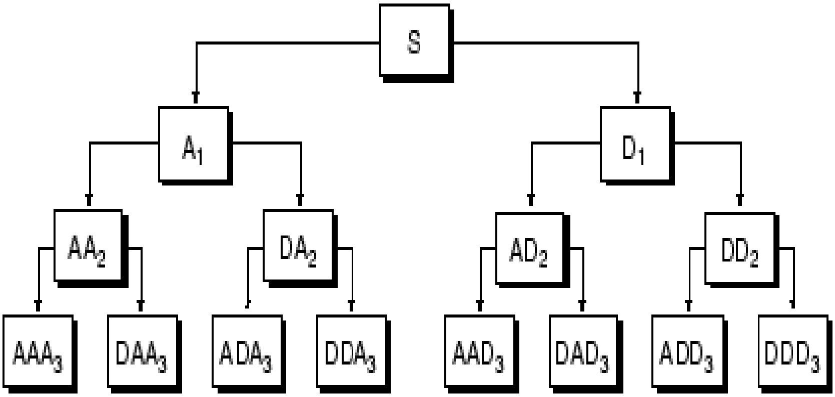
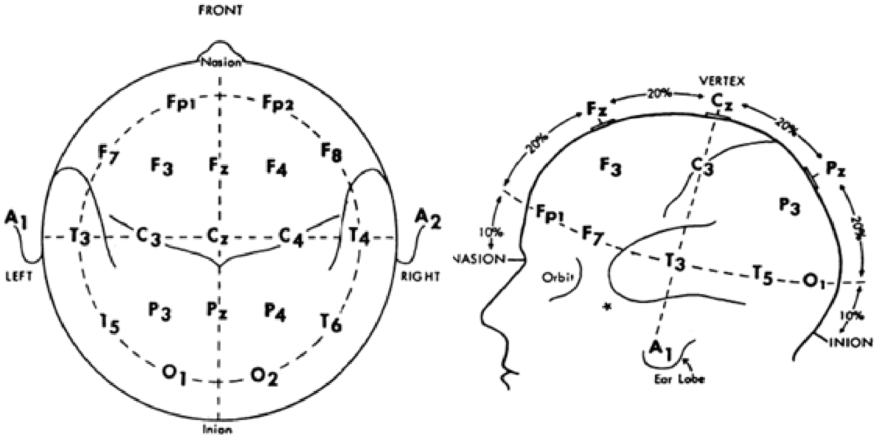


Fig. 1. The Wavelet Decomposition tree

3 Electroencephalogram Data

EEG is usually recorded according to the international 10-20 electrode placement system [4] shown in Fig.2. The 10-20 system was developed to standardize the collection of EEG and facilitate the comparison of studies performed in different parts of the world. When only a few channels of EEG are collected, the electrodes are placed at a subset of the sites. Fig.3 shows plot of EEG data from a normal subject and Fig.4 shows EEG recordings from a patient in ictal state. This data is taken from the source as given in [5]. Two sets of data are used, one from healthy subjects with eyes open and another for patients in ictal state, meaning thereby that the patient is under seizure during the recording of data. These two data sets are used in the subsequent section to demonstrate wavelet decomposition EEG of signals.



(a) (b)

Fig. 2. 10/20 International system of electrode placement. (a) Top view (b) Left side view.

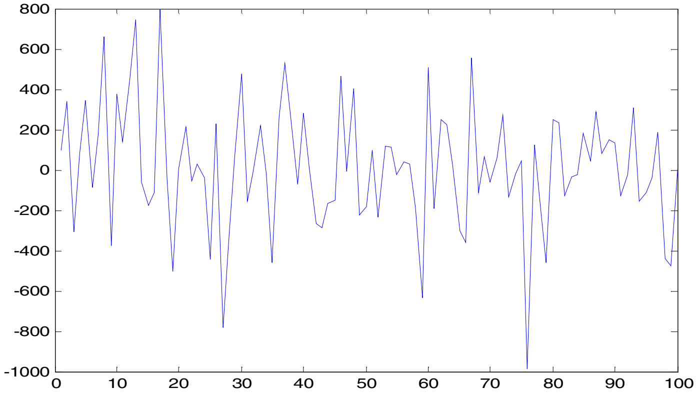


Fig. 3. EEG Signal from a normal subject

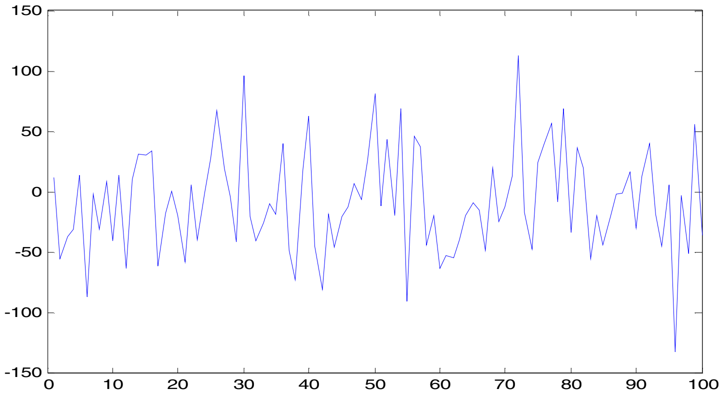


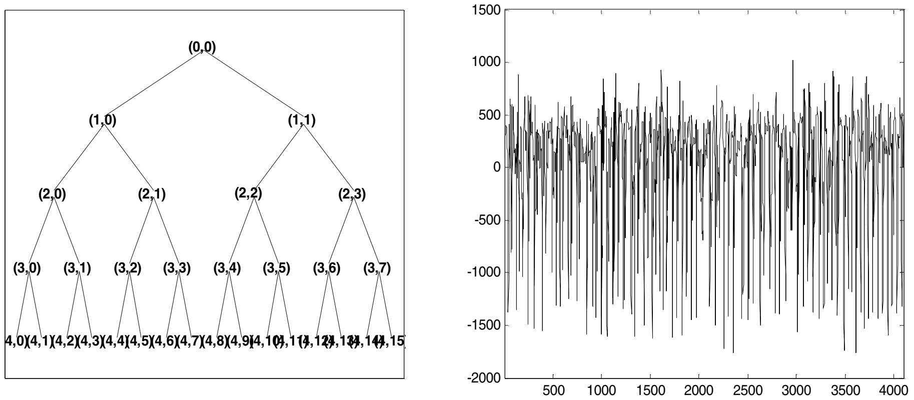
Fig. 4. EEG Signal from a patient in ictal state

4 Wavelet Decomposition of EEG

Feature extraction and classification of the signals are required for both diagnostic and brain computer interface purposes. However, Feature extraction always suffers from critical problems in time domain analysis [1]. The main purpose of feature extraction, is to extract salient characteristics from digitized data collected from the data acquisition phase [6] follows classification based on the extracted features [7, 8]. The features of the signal are derived from its linear expansion coefficients; the most common linear expansion method used is Fourier transform [10]. Fourier transform is suitable only for stationary signals. However, EEG signals are characterized as non-stationary signals; so, if processed with Fourier transform, would not yield the best result. Hence, for such non-stationary signals, a time–frequency representation is required, in order to derive meaningful features [9]. So wavelet transforms are one of such good choices to process these kind of signals. There are family of wavelet transforms which can be used for different purposes here Wavelet packet decomposition of EEG signals is demonstrated in subsequent section.

Wavelet Packet Decomposition of EEG Signals Wavelet packet analysis is used for extracting features of non-stationary signals, and is appropriate for extracting features from EEG signals which are non stationary in nature. Wavelet packet analysis is a generalized form of the DWT, wherein a signal is split into approximation and detailed coefficients. The approximation is again split into a second-level approximation and detailed coefficients, and the process is repeated. There are n + 1 possible ways to encode the signal for the n-level decomposition. The signal is passed through a series of low-pass and high-pass filters called quadrature mirror-filters.

In wavelet packet analysis, the approximation as well as detailed coefficients can be split like complete binary tree structure. A 4th level decomposition of an EEG signal from a normal subject is demonstrated in this paper. In Fig.5(a), the decomposition tree for a 4th level decomposition and in Fig.5(b), the data at node first of the 4th level of decomposition is shown. The features can be extracted from the coefficients of decomposition; by calculating some statistical and non statistical properties for each node of decomposition.



(a) (b)

Fig. 5. (a) Wavelet packet analysis decomposition tree up to 4th level decomposition (b) Data at node (4, 1)

5 Conclusions and Future Directions

Wavelet packet decomposition using Matlab with wavelet and bioinformatics tool box, is carried out. A 4th level decomposition using wavelet packet decomposition is demonstrated. The displayed data in Fig 5(b) are the data corresponding to the coefficient at node 1st of 4th level decomposition. For n-level decomposition, we have 2^n coefficients, we can generate characteristic features from these coefficients for the purpose of classification. For both diagnostic and brain computer interface we need an efficient and accurate strategy of classification of EEG signals. So this kind of decomposition may be used for extraction of features for different kinds of classifier systems. Further we can try and compare the efficiency of the classification system for different types of wavelet transforms and can find out the best wavelet transform this purpose.

References

[1] Shi, Y., Zhang, X.: A Gabor atom network for signal classification with application in radar target recognition. IEEE Transactions on Signal Processing, 2994–3004 (2001)

[2] Jahankhani, P., Kodogiannis, V., Revett, K.: EEG signal classification using wavelet feature extraction and neural networks. In: IEEE John Vincent Atanasoff International Symposium on Modern Computing (2006)

[3] Litt, B., Echauz, J.: Prediction of epileptic seizures. Lancet Neurol. 1, 22–30 (2002)

[4] Jasper, H.H.: The ten-twenty electrode system of the international federation. Electroencephalography and Clinical Neurophysiology 10, 371–373 (1958)

[5] Andrzejak, R.G., et al.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Physical Review E 83 64, 061907, doi:10.1103/Physreve.64.061907

[6] Yong, L., Shenxun, Z.: The application of wavelet transformation in the analysis of EEG. Chinese Journal of Biomedical Engineering, 333–338 (1998)

[7] Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Mixed band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. IEEE Transactions on Biomedical Engineering, 1545–1551 (2007)

[8] Ghosh-Dastidar, S., Adeli, H.: Improved spiking neural networks for EEG classification and epilepsy and seizure detection. Integrated Computer-Aided Engineering 14(3), 187–212 (2007)

[9] Kalayci, T., Ozdamar, O.: Wavelet preprocessing for automated neural network detection of EEG spikes. IEEE Engineering in Medicine and Biology Magazine, 160–166 (1995)

[10] Dickhaus, H., Heinrich, H.: Classifying biosignals with wavelet networks. IEEE Engineering in Medicine and Biology Magazine, 103–111 (1996)

[11] Berger, H.: Über das Elektroenkephalogram des Menschen. Arch. f. Psychiatry 87, 527–570 (1929)

Design and Analysis of Small Planar Antenna Based on CRLH Metamaterial for WSN Application

Sanjeev Jain[1], Indrasen Singh[1], Vijay Shanker Tripathi[1], and Sudarshan Tiwari[2]

[1] Department of Electronics and Communication Engineering
Motilal Nehru National Institute of Technology
Allahabad, 211004, India
[2] National Institute of Technology, Raipur 492010, India
{snjece,erindrasen}@gmail.com, {vst,stiwari}@mnnit.ac.in

Abstract. In this paper a small antenna is proposed for wireless sensor network applications in the frequency band of 5 GHz to 15 GHz. Wireless Sensor Networks (WSN) is having significant attention due to their numerous features nowadays. The most important component in sensor is its TX/RX antenna. In this design a new efficient CRLH based DNG metamaterial antenna is proposed. The design of the unit cell of these antennas is based on the composite right/left-handed (CRLH) transmission line model. Antenna1 used two patches and obtained results are compared with antenna2 with four patch. In this paper, we introduce a new method of design of CRLH metamaterial antennas to tackle the above problem using CRLH mushroom structure. The results of antenna2 show improvements over antenna1 in gain and bandwidth. By employing CRLH geometry, an overall size reduction of 65% was achieved compared to the conventional rectangular patch antenna. The proposed antenna can be easily built in a miniaturized wireless sensor network (WSN). RogersRT/Duroid 5880 is taken as substrate with thickness 1.572 mm and relative permittivity 2.2. The bandwidth of this antenna less than 10 dB is 700 MHz and the percentage of the bandwidth is 6.1%. The antenna characteristics, such as return loss and VSWR achieved by the proposed structure are plotted.

Keywords: Return loss, WSN, VSWR, CRLH, Bandwidth.

1 Introduction

Metamaterial structures were first introduced by Veselago in 1968 [1]. Later on, the realization of such material was done by [2–4], where periodic array of wires with specific radios and spacing is used to produce negative permittivity [2] or capacitive loaded stripes can also be used to produce the same effect [4], whereas the negative permeability was induced using the split ring resonators [3]. These metamaterials have many interesting properties, one of these that the double negative metamaterials are able to match the intrinsic capacitance (reactance) of an electrically small antenna. This property was used in [5] toward producing an efficient electrically small wire antenna.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 290–297, 2013.

One of the other metamaterial realizations was presented using transmission line (TL) theory based on the equivalent circuits of the right-handed and left-handed propagation in the unit cell of the TL [8–11]. This model is called composite right/left-handed (CRLH) met materials. The CRLH-TL can be realized using the mushroom structure [9] which was originally introduced by Seivenpiper [12] to produce high impedance surfaces. Park [13] emphasize on using the CRLH antenna structures at the zeroth-order resonance.

The Wireless Sensor Network (WSN) is a flexible and scalable paradigm that is drawing increasing attention due to its potential utilization in many civilian and military domains effect [4], A WSN utilizes a significant number of sensor motes and automatically builds a network topology. People staying at the coverage of the network can receive the signals sent by the sensors in time and thus swiftly arrived at the venue. Based on this requirement, in this paper our aim is to develop a novel small antenna that is compatible with the WSN. The novel antenna should have a compact size, low profile, low power, low cost, easy fabrication, and sufficient transmission range. They are suitable to develop a miniature WSN, combine individual wireless sensors, and become an information network for universal transmission of information.

2 CRLH Transmission Line Model

The equivalent electrical circuit models of a purely RH (PRH), purely LH (PLH), and CRLH lossless Transmission Line are shown in Figure 1(a), (b), and (c), respectively. Transmission Line theory has long been a powerful analysis and design tool for conventional RH materials. By modeling a CRLH metamaterial as an equivalent TL, TL theory can be used to analyze and design 1-D, 2-D, or even 3-D CRLH metamaterials. In this section the TL approach of CRLH metamaterials has developed. First, the CRLH metamaterial will be represented by an equivalent electrical homogeneous CRLH TL to gain immediate insight into its fundamental characteristics. Then, an LC network implementation of the TL will be developed, since homogeneous CRLH structures do not appear to exist in nature. The LC network pro-vides a realistic description of the CRLH metamaterial. Finally, physical realizations of the LC network will be discussed. For simplicity, only the lossless TL will be examined. This circuit model shows that the structure is a CRLH material, as explained in [14]. The unit cell consists of a ground plane, substrate, and patch on the top of the substrate. The via connects the patch and the ground plane, and there is a gap between the adjacent patches in the whole structure. The study of the TL model of his structure is vital to know the effect of the inductances and capacitances on the guided and radiating waves.

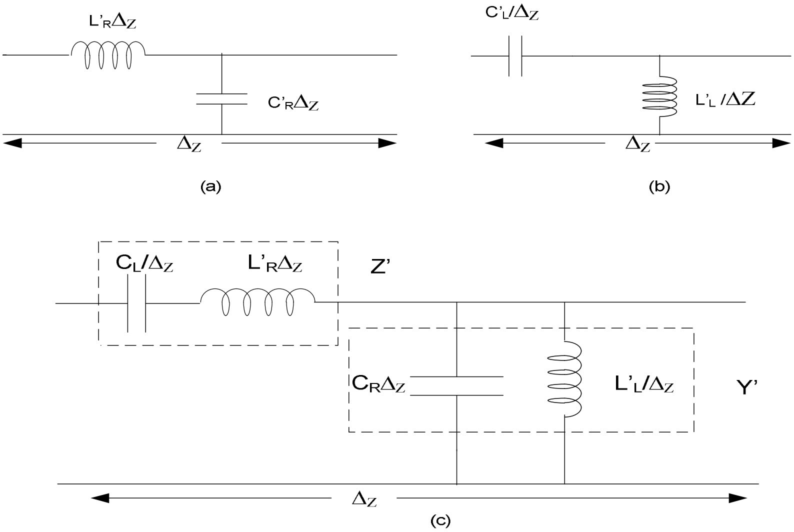


Fig. 1. Equivalent circuit model of (a) purely RH TL (b) Puruly LH TL (c) Purely CRLH TL

The propagation constant of a TL is given by $\gamma = \alpha + j\beta = \sqrt{Z'Y'}$, where Z' and Y' are, respectively, the per-unit length impedance and per-unit length admittance. In the particular case of the CRLH TL, Z' and Y' are defined as

$$Z' = j\omega L'_R - \frac{1}{\omega C'_L}$$

$$Y' = j\omega C'_R - \frac{1}{\omega L'_L} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The advantages of using CRLH structures in the antenna design are that the resonances in the LH part of the dispersion diagram have lower resonant frequencies compared to the conventional antennas. This planar antenna has a zeroth-order resonant frequency that depends mainly on the values of the equivalent inductances and capacitances of the unit cell. Based on the open ended TL model of the CRLH, the zeroth-order resonance is given by

Based on the open ended TL model of the CRLH, the zeroth-order resonance is given by

$$fo = \frac{1}{2\pi\sqrt{L_L C_R}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

and the bandwidth ratio at this resonance is proportional to $\sqrt{L_L/C_R}$.

3 Antenna Design with Two Patches

The geometry of the proposed CRLH based small antenna with two patches is shown in Fig. 2.

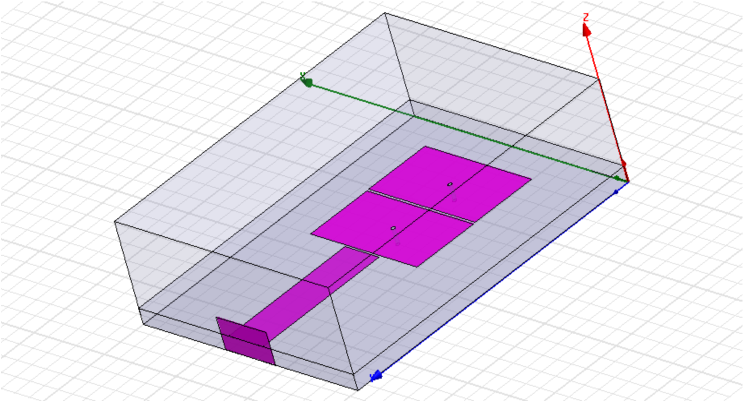


Fig. 2. CRLH metamaterial patch antenna1

The antenna is printed on a thin substrate with dielectric constant er=2.2 and dimensions of 30x35x1.575 mm3. Two patches are printed on the substrate; each

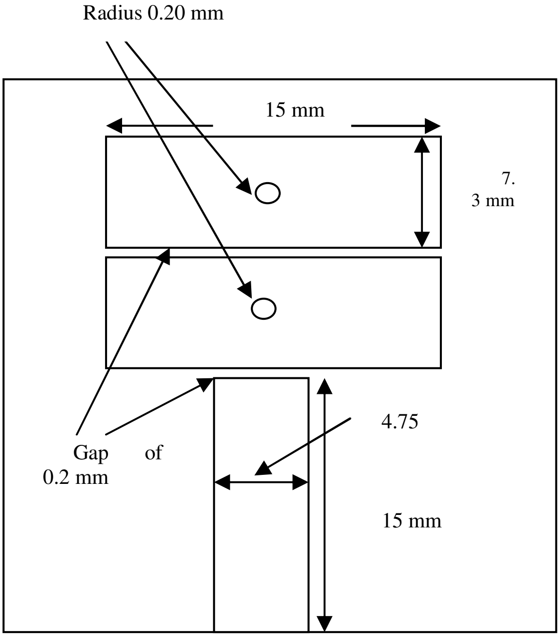


Fig. 3. Top view of CRLH antenna with two patches

patch has dimensions of 15x7.3 mm2 with a gap between the two patches of 0.2 mm. A micro strip line, of 4.75 mm width, is used to feed the first patch through a gap of 0.2 mm. A via with radius 0.2mm is used in each unit cell to introduce the shunt inductance L_L. The gap between the patches is used to introduce the series capacitance C_L. The antenna is simulated using HFSSv12 and the results plotted Figure 4 and 5. The results of the measurement agree well with the results of the simulation in the band 5–15 GHz.

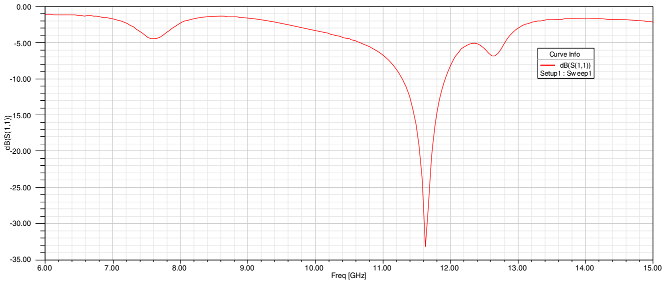


Fig. 4. Simulated Return loss (S11) plot for proposed antenna with two patches

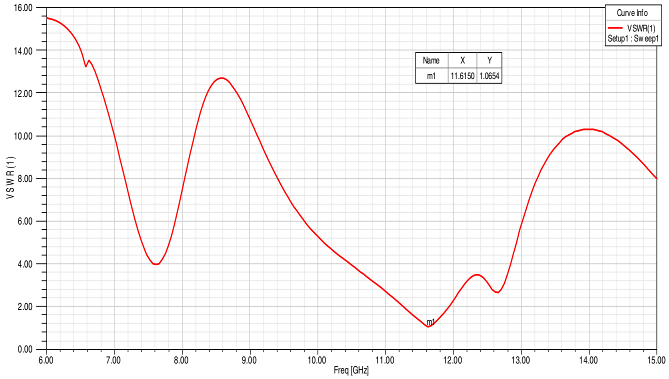


Fig. 5. Simulated VSWR plot for proposed antenna with two patches

4 Antenna Design with Four Patches

The geometry of the proposed CRLH based small antenna with four patches is shown in Fig. 5. The antenna is printed on a thin substrate with dielectric constant er = 2.2 and dimensions of 30x35x1.575 mm3. Four patches are printed on the substrate; each patch has dimensions of 15x5 mm2 with a gap between the two patches of 0.2 mm. A micro strip line, of 4.75 mm width, is used to feed the first patch through a gap of 0.2 mm. due to increase of patches over substrate the performance of antenna increases.

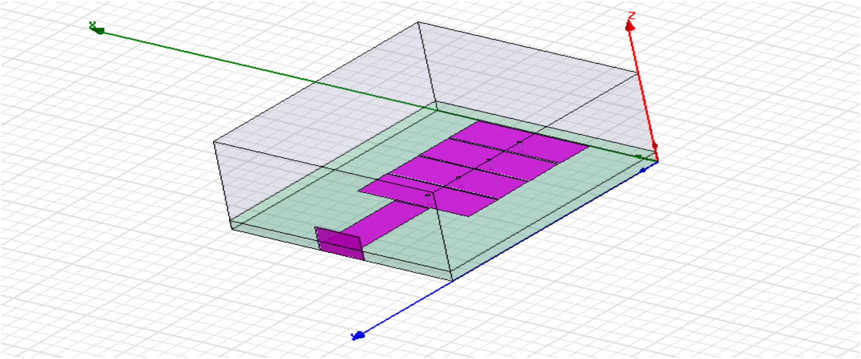


Fig. 6. CRLH metamaterial antenna with four patches

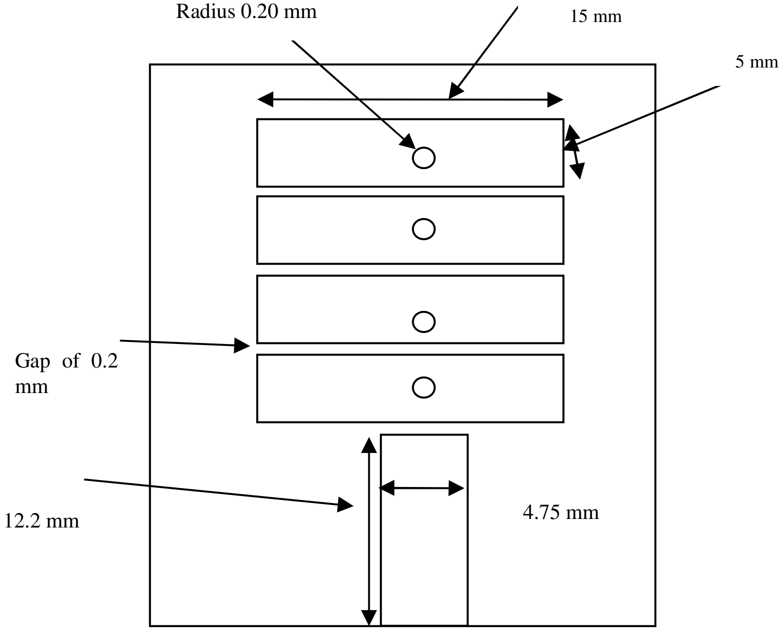


Fig. 7. Top view of CRLH antenna with four patches printed on thin substrate

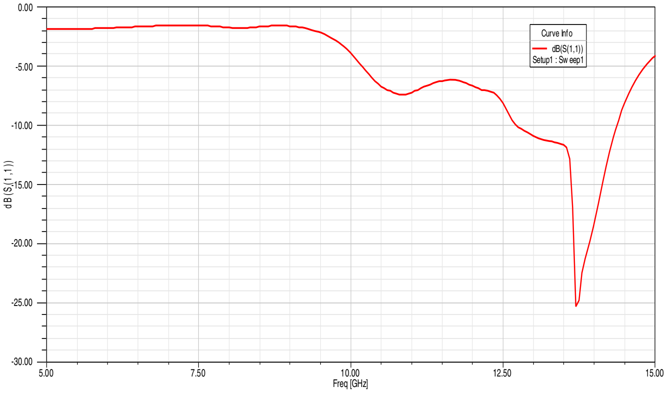


Fig. 8. Simulated Return loss (S11) plot for proposed antenna with four patches

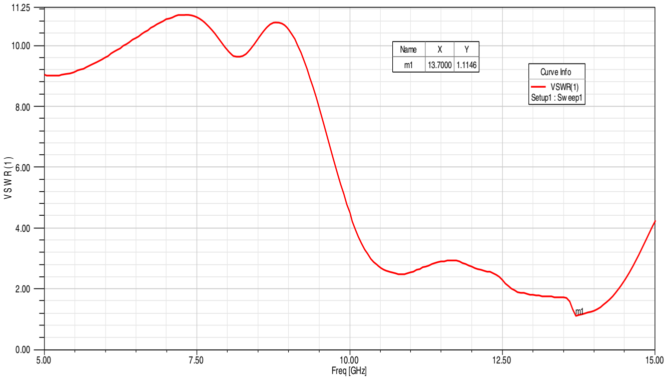


Fig. 9. Simulated VSWR plot for proposed antenna with four patches

5 Conclusions

The small size planner antenna designed for used in WSN is mainly intended to reduce power, achieve miniaturization, and achieve significant transmission distances. In this paper, we select the CRLH based double negative metamaterial for antenna miniaturization. We can shift the center frequency of the antennas by adjusting the gap between two CRLH unit and radius of via which is connected to ground and patch. Because of the advantage of the DNG Metamaterial it increases the bandwidth; moreover, the characteristics of the antenna can be improved, while its size can be significantly reduced and applicable for WSN application. By using four patches performance of antenna increases significantly. The above mentioned advantages have help to achieve the goal of antenna application for WSN. The designed structures use the concept of the CRLH-TL to achieve negative permeability and permittivity in the LH region.

References

1. Veselago, V.G.: The electrodynamics of substances with simultane-ously negative values ofland e. Sov. Phys. Uspekhi 10, 509–514 (1968)
2. Pendry, J.B., Holden, A.J., Stewart, W.J., Youngs, I.: Extremely low frequency plasmons in metallic mesostructures. Phys. Rev. Lett. 76, 4773–4776 (1996)
3. Pendry, J.B., Holden, A.J., Robbins, D.J., Stewart, W.J.: Magne-tism from conductors and enhanced nonlinear phenomena. IEEE Trans. Microwave Theory Tech. 47, 2075–2084 (1999)
4. Ziolkowski, R.W.: Design, fabrication, and testing of double negative metamaterials. IEEE Trans. Antennas Propag. 51, 1516–1529 (2003)
5. Ziolkowski, R.W., Kipple, A.D.: Application of double negative materials to increase the power radiated by electrically small antennas. IEEE Trans. Antennas Propag. 51, 2626–2640 (2003)
6. Alu, A., Engheta, N.: Coaxial-to-waveguide matching with e-near-zero ultranarrow channels and bends. IEEE Trans. Antennas Propag. 58, 328–339 (2010)
7. Jin, P., Ziolkowski, R.: Broadband, efficient, electrically small metamaterial-inspired antennas facilitated by active near-field reso-nant parasitic elements. IEEE Trans. Antennas Propag. 58, 318–327 (2010)
8. Sanada, A., Caloz, C., Itoh, T.: Planar distributed structures with negative refractive index. IEEE Trans. Microwave Theory Tech. 52 (2004)
9. Caloz, C., Itoh, T.: Electromagnetic Metamaterials: Transmission Line Theory and Microwave Applications. Wiley, Hoboken (2006)
10. Lai, A., Caloz, C.: Itoh, T. Composite right/left-handed transmission line metamaterials. IEEE Microwave Mag., 33–50 (2004)
11. Qureshi, F., Antoniades, M.A., Eleftheriades, G.V.: A compact and low-profile metamaterial ring antenna with vertical polarization. IEEE Antennas Propag. Lett. 4, 333–336 (2005)
12. Seivenpiper, D., Zhang, L., Broas, R.F., Alexopolous, N.G., Yablonovitch, E.: High-impedance electromagnetic surfaces with for-bidden frequency band. IEEE Trans. Microwave Theory Tech. 47, 2059–2074 (1999)
13. Park, J.H., Ryu, Y.H., Lee, J.G., Lee, J.H.: Epsilon negative zer-oth-order resonator antenna. IEEE Trans. Antenna Propag. 55, 3710–3712 (2007)
14. Abd-El-Raouf, H.E., Syed, S., Antar, Y.M.: Design of small antennas based on DNG metamaterials. In: IEEE International RF and Microwave Conference, Atlanta, GA (2008)
15. Abd El-Raouf, H.E., Yu, W., Farahat, N., Mittra, R.: Matched load truncation for feed line of patch antennas in the FDTD method. Microwave Opt. Technol. Lett. 49, 267–269 (2001)
16. Abd-El-Raouf, H.E., Syed, S., Antar, Y.M.: Design of double layered metamaterial antenna. In: European Conference on Antennas and Propagation (2009)
17. Syed, S., Abd-El-Raouf, H.E., Antar, Y.M.: CRLH Metamaterial Antennas Using Thick Substrate—Single and Double Layers. In: IEEE International Symposium on Antennas & Propagation and USNC/ URSI National Radio Science Meeting, Charleston, SC (2009)
18. Lim, S., Caloz, C., Itoh, T.: A reflecto-directive system using a composite right/left-handed (CRLH) leaky-wave antenna and heterodyne mixing. IEEE Microwave Wireless Compon. Lett. 14, 183–185 (2004)
19. Sanada, A., Caloz, C., Itoh, T.: Zeroth order resonance in com-posite right/left-handed transmission line resonators. In: Proc. Asia-Pacific Microwave Conf., Seoul, Korea, vol. 3, pp. 1588–1592 (2003)

Classification of Speech Dysfluencies Using Speech Parameterization Techniques and Multiclass SVM

P. Mahesha[1] and D.S. Vinod[2]

[1] Department of Computer Science and Engineering
S.J. College of Engineering
Mysore, Karnataka, India
maheshsjce@yahoo.com

[2] Department of Information Science and Engineering
S.J. College of Engineering
Mysore, Karnataka, India
ds.vinod@daad-alumni.de

Abstract. Stuttering is a fluency disorder characterized by the occurrences of dysfluencies in normal flow of speech, such as repetitions, prolongations and interjection and so on. It is one of the serious problems in speech pathology. The goal of this paper is to present experimental results for the classification of three types of dysfluencies such as syllable repetition, word repetition and prolongation in stuttered speech. The three speech parameterization techniques :Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC) are used as speech feature extraction methods. The performance of these parameterization techniques are compared using the results obtained by thorough experimentation. The speech samples are obtained from University College London Archive of Stuttered Speech (UCLASS). The dysfluencies are extracted from these speech samples and used for feature extraction. The multi-class Support Vector Machine (SVM) is employed for the classification of speech dysfluencies.

1 Introduction

Speech is one of the effective ways of communication between people. The basic purpose of speech is to send and receive a message in the form of language communication. Even speakers who are normally fluent experience dysfluencies due to emotional, physiological or psychological factors. Speaker is dysfluent when involuntarily repeating a word, prolonging a word, forgetting a word mid utterance, or interjecting too many “uh’s” and “um’s” during speech.

Stuttering is a sort of fluency disorder that sways the flow of speech. Approximately about 1% of the community is suffering from this disorder and has found to affect four times as many males as females[25,5,24,2]. The most perceptible attribute of this disorder is the production of certain types of delinquencies in

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 298–308, 2013.

Table 1. Type of Dysfluencies with Example

Type of Dysfluencies	**Example**
Repetition	
Whole word	“What-what-what are you doing”
Part word	“What t-t-t time is it?”
Prolongation	
Sound/ syllable	“I am Booooobbbby James”
Interjection (Filled pauses)	
Sound/syllable	“Um – uh -well, I had problem in morning”
Silent pauses	
Silent duration within speech considered normal	“I was going to the [pause] store”
Broken words	
A silent pause with in words	“it was won[pause]derful”
Incomplete phrase	
Grammatically in complete utterance	“I don't know how to . . . .
	let us go, guys”
Revisions	
Changed words, ideas	“There was a dog, no rat named Arthur”

the flow of speech. Dysfluencies are disturbances or breaks in the smooth flow of speech. This disorder is characterized by following major types of dysfluencies such as repetitions, prolongations, interjections, broken words etc. Examples of these dysfluencies are recorded in Table 1. Stuttering is the subject of interest to researchers from various domains like speech physiology, pathology, psychology, acoustics and signal analysis. Therefore, this subject is a multidisciplinary research field of science.

In conventional stuttering assessment method Speech Language Pathologists (SLP) classify and count the occurrence of dysfluencies manually by transcribing the recorded speech. These types of assessments are based on the knowledge and experience of speech pathologists. However, making such assessment are time consuming, subjective, inconsistent and prone to error [23,11,10,18,6]. Therefore, it would be sensible if stuttering assessment is often done through classification of dysfluencies using speech recognition technology and computational intelligence. The dysfluent speech processing is one of the areas, where research remains substantially ongoing.

In the last two decades, several studies [1,15,16,11,10,3] have been carried out on the automatic detection and classification of dysfluencies in stuttered speech by means of acoustic analysis, parametric and non-parametric feature extraction and statistical methods. Which facilitate SLPs for objective assessment of stuttering. In[1], author used Artificial Neural Network (ANN) and rough set to detect stuttering events yielding accuracy of 73.25% for ANN and about 91% for rough set. The authors of [15,16] proposed Hidden Markov Model (HMM) based classification for automatic dysfluency detection using MFCC features and achieved 80% accuracy. In [11], automatic detection of syllable repetition was presented for objective assessment of stuttering dysfluencies based on MFCC and perceptron features. An accuracy of 83% was achieved. Subsequently in[10],same author obtained 94.35% accuracy using MFCC features and SVM classifier. Authors of [3] achieved 90% accuracy with Linear Discriminant Analysis (LDA), k- Nearest Neighbor (k-NN) and MFCC features. In[13] the same author used similar classifiers, LDA and k-NN for the recognition of repetitions and prolongations with Linear Predictive Cepstral Coefficient (LPCC) as feature extraction method and obtained the best accuracy of 89.77%. In [19]our previous work, we have developed a procedure for classification of dysfluency using MFCC feature and k-NN classifier and obtained best accuracy of 97.78% for k=5.

Investigation of the literature shows that, different feature extraction and classification algorithms have been proposed. Most of these methods concentrate on classification of syllable repetition dysfluency. In few works[5,3,9] classification of prolongation is also considered. However, stuttering is characterized by different dysfluencies as listed in Table 1. There are no attempts in literature for classifying two forms of repetition such as syllable and word repetition.

Therefore, in this work we are proposing LPC, LPCC and MFCC based speech parametrization techniques to classify three types of dysfluencies such as syllable repetition, word repetition and prolongation. The Multiclass SVM is employed for classification of dysfluencies. The comparative analysis of the these parameterization techniques are presented.

2 Database

The speech database was acquired from UCLASS[17,22]. It contains recordings of stuttered speech. This database is freely available to assist people doing research in the area of stuttered speech. We have selected 20 sound recordings for experimentation. These twenty speech files include 10 male and 10 female speakers with age ranging from 11 to 20 years. The samples were selected with intent to cover a wide range of age and stuttering rate.

3 Methodology

The classification system aims to identify the different types of dysfluencies such as syllable repetition, word repetition and prolongation. The system performs thorough analysis of speech signal by extracting features which contain characteristic information of dysfluencies. The SVM classifier is used to classify different types of dysfluencies. Our classification system has four modules : segmentation, pre-processing, features extraction and classification as shown in Figure 1.

3.1 Segmentation

The collected speech samples of UCLASS database are analyzed to identify and segment the dysfluencies manually, which is tedious but straight forward

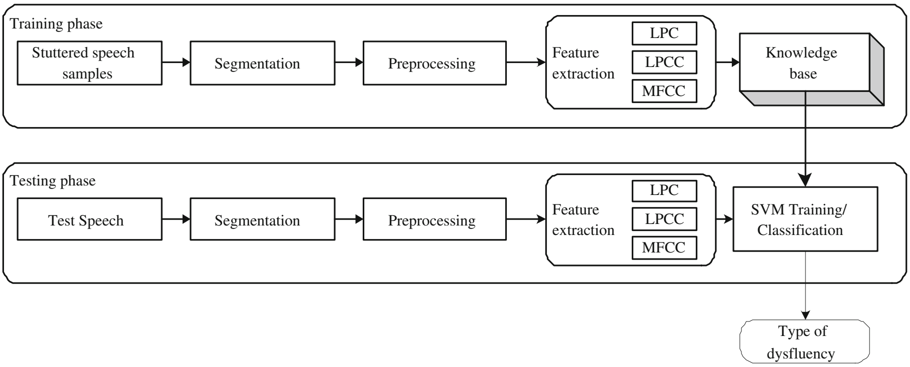


Fig. 1. Block diagram of classification

approach[[11,10]. The segmented speech syllables are subjected to feature extraction. We segmented three types of dysfluencies such as syllable repetitions, word repetitions and prolongations from speech samples.

3.2 Speech Signal Pre-processing

The speech signal pre-processing is performed to enhance the accuracy and efficiency of the feature extraction process. This phase is common for all the feature extraction methods as shown in Figure 2. This is carried to spectrally flatten the signal. A pre-emphasis filter is a simple first order high pass filter used to flatten the signal[12]. Typically the first order FIR filter is used as transfer function. The z-transform of the filter is given by

$$H(z) = 1 - \bar{a} * z^{-1}, \qquad 0.9 \leq a \leq 1.0 \tag{1}$$

The output of the pre-emphasis network $\bar{s}(n)$ is related to the input of network $s(n)$, by difference equation:

$$\bar{s}(n) = s(n) - \bar{a}s(n-1) \tag{2}$$

The output of pre-emphasized signal $\bar{s}(n)$ is divided into frames of N samples. Adjacent frames are sampled by M samples, in order to analyze each frame in the short time instead of analyzing the entire signal at once[9]. If $x_l(n)$is the l^{th}frame and there are L frames within entire speech signal, then

$$x_l(n) = s(Ml + n), \qquad n = 0, 1, \ldots, N-1 \; and \; l = 0, 1, \ldots, L-1 \tag{3}$$

The Hamming window is applied to each frame, which has the form :

$$w(n) = 0.54 - 0.46cos\left[\frac{2\pi n}{n-1}\right], \qquad 0 \leq n \leq N-1 \tag{4}$$

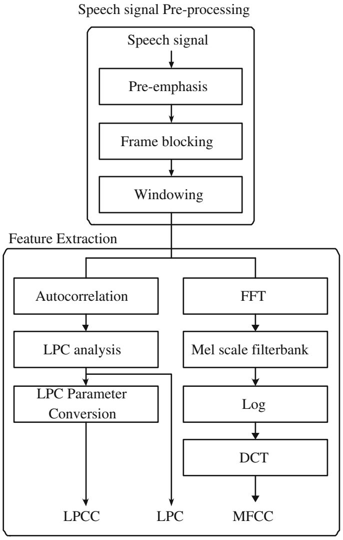


Fig. 2. Block diagram of speech signal pre-processing and feature extraction

3.3 Speech Parameterization

Speech parameterization is an important step in speech recognition systems. It is used to extract significant features from speech samples. Feature extraction is to convert an observed speech signal to some type of parametric representation for further investigation and processing. Three speech parameterization techniques were employed in this study, namely LPC, LPCC and MFCC.

LPC and LPCC. LPC is one of the most prevailing speech analysis technique. The steps involved in computation of LPC is shown in Figure 2. The LPC model is based on a mathematical approximation of the vocal tract represented by tube of a varying diameter. The key characteristic of LPC is, given the speech sample at time n, $\widehat{s}(n)$ can be predicted as linear combination of past p sample values. Where p represents order of the LPC[12].

$$\widehat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{5}$$

The prediction error $e(n)$ at any time is the difference between the actual and the estimated sample value, given by

$$e(n) = s_n - \widehat{s}(n) \tag{6}$$

$$= s_n - \sum_{i=1}^{p} a_i s(n-i) \tag{7}$$

In our work LPC with autocorrelation method is applied to each frame of windowed signal as given in [12], given by equation 8 and 9

$$r(m) = \sum_{i=1}^{N-1-m} x(n)x(n+m), \qquad m = 0, 1, \ldots p \tag{8}$$

where the autocorrelation function is symmetric, as a result the LPC equations can be stated as

$$\sum_{m=1}^{p} r(\mid m - k \mid)a_m = r(m), \qquad 1 \leq m \leq p \tag{9}$$

LPCC is Linear Prediction Coefficients (LPC) represented in the cepstrum domain [7]. These are the coefficients of the Fourier transform representation of the log magnitude spectrum. After obtaining LPC we compute Cepstral Coefficients(CC). LPCC can be derived directly from the LPC coefficients set. The recursion used is defined as follows :

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad 1 \leq m \leq p \tag{10}$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) \cdot c_k \cdot a_{m-k} \qquad m > p \tag{11}$$

c_m- Cepstral coefficients, a_m- Predictor coefficients, k - $1 < k < N - 1$, p - pth order.

MFCC. The MFCC is one of the popular speech parameterization technique and most commonly used feature for speech recognition. It produces a multidimensional feature vector for every frame of speech. In this study we have considered 12 MFCCs. The method is based on human hearing perceptions which cannot perceive frequencies over 1KHz. In other words, MFCC is based on known variation of the human ear's critical bandwidth with frequency[14]. The block diagram for computing MFCC is illustrated in Figure 2.

In first step, Fast Fourier Transform(FFT) is applied to pre-emphasized signal to convert each frame of N samples from time domain to frequency domain. Then, a set of triangular filters also called Mel-scale filters are used to compute a weighted sum of filter spectral components and the output of the process approximates to a Mel scale. The Mel frequency scale is linear up to 1000 Hz and logarithmic there after[20]. The Mapping of linear frequency to Mel scale is represented by the following equation (12). In final step log Mel spectrum is converted back to time domain using Discrete Cosine Transform (DCT). The outcome of conversion is called MFCCs.

$$mel(f) = 2595 log_{10} \left(1 + \frac{f}{700}\right) \tag{12}$$

In this study, 12 LPC, LPCC and MFCC were extracted to classify the three types of dysfluencies, namely syllable repetition, word repetition and prolongation in stuttered speech.

3.4 SVM Training and Classification

We have used SVM method for classifying three different types of dysfluencies. A SVM is a classification technique based on the statistical learning theory[21,4]. It is supervised learning technique that uses a labeled data set for training and tries to find a decision function that classifies best the training data. The purpose of the algorithm is to find a hyperplane to define decision boundaries separating between data points of different classes. It is commonly used in pattern recognition and classification problem. It gives good classification performance with limited training data compared to other classifiers. The hyper plane equation is given by

$$w^T x + b \tag{13}$$

where w is weight vector and b is bias.

Given the training labeled data set$\{x_i, y_i\}_{i=1}^{N}$with $x_i \in \mathbb{R}^d$ being the input vector and $y_i \in \{-1, +1\}$.Where $x_i$is input vector and y_i is its corresponding label[8]. SVMs map the d-dimensional input vector x from the input space to the d_h- dimensional feature space by non-linear function $\varphi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_h^d$. Hence hyperplane equation becomes

$$w^T \varphi(x) + b = 0 \tag{14}$$

with $b \in \mathbb{R}$ and w an unknown vector with the same dimension as $\varphi(x)$. The resulting optimization problem for SVM, is written as

$$\min_{w,\xi,b} \jmath_1(w, \xi) = \frac{1}{2} w^T w + c \sum_{i=1}^{n} \xi_i \tag{15}$$

such that

$$y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N \tag{16}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, N \tag{17}$$

The constrained optimization problem in equation 15, 16 and 17 is referred as the primal optimization problem. The optimization problem of SVM is usually written in dual space by introducing restriction in the minimizing function using Lagrange multipliers. The dual formulation of the problem is

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i, x_j) \tag{18}$$

subject to $\alpha_i \geq 0$ for all $i = 1, \ldots m$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

Thus, the hyperplane can be written in the dual optimization problem as :

$$f(x) = sgn\left[\sum_{i=1}^{m} y_i \alpha_i \left(x_i, x\right) + b\right] \tag{19}$$

Multiclass SVM. In this study multiclass SVM classifies given testing sample to one of the three classes. There are several methods for dealing with multiple classes. In the current work we use "one vs rest" method where, element that belong to a class are differentiated from the other. We calculate an optimal hyperplane that separates each class from the rest of the elements.

To get classification of N classes, a set of binary classifiers are constructed where each training separates one class from the rest. After that we combine them by doing the multiclass classification according to the maximal output before applying the *sgn* function, that takes a form

$$\arg \max_{j=1,\dots M} g^j(x), \quad where\ g^j(x) = \sum_{i=1}^{m} y_i \alpha_i^j k(x, x_i) + b^j \tag{20}$$

and

$$f^j(x) = sgn(g^j(x)) \tag{21}$$

This has a linear complexity as for N classes we compute N hyperplanes. In our study we have 3 classes and we compute 3 hyperplanes.

4 Experimental Results

As explained in the section 2, speech samples are selected from UCLASS database. From the selected speech samples, we have created 150 speech segments of syllable repetition, word repetition and prolongation. Using these set of segments, we created a training and a testing group. The 80% of the segment is used for training and 20% for testing. The Table 2 shows the distribution of speech segments for training and testing.

To extract features form the speech samples, we have considered three speech parameterization techniques namely LPC, LPCC and MFCC. The experiment is conducted independently for each of the features by considering the same data

Table 2. The speech data

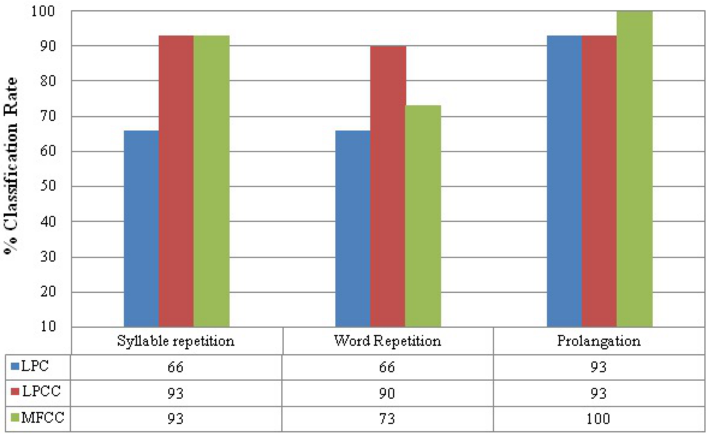
	Speech segments	Training	Testing
Syllable repetition	50	40	10
Word repetition	50	40	10
Prolongation	50	40	10
Total	150	120	30

as given in Table 2. We use SVM for the classification of dysfluencies and the total 60 speech segments are divided into 3 classes. In each experiment, we chose 80% of each class as the training set and remaining 20% as the testing data. The experiment was repeated 3 times, each time different training and testing sets were built randomly. The average accuracy of each type of dysfluencies were compared and reported in Figure 3.

Table 3 shows the classification result for syllable repetition, word repetition and prolongation with three different types of feature extraction techniques for three different set. It also shows the average accuracy of three dysfluencies for each set and overall average accuracy for LPC, LPCC and MFCC.

Table 3. The average classification accuracy of LPC, LPCC and MFCC using multiclass SVM

	LPC			**LPCC**			**MFCC**		
Dysfluencies	*Set 1*	*Set 2*	*Set 3*	*Set 1*	*Set 2*	*Set 3*	*Set 1*	*Set 2*	*Set 3*
Syllable repetition	60%	60%	80%	100%	80%	100%	100%	80%	100%
Word repetition	60%	80%	60%	80%	90%	100%	60%	80%	80%
Prolongation	80%	100%	100%	100%	100%	80%	100%	100%	100%
Accuracy	66.00%	80.00%	80.00%	93.00%	90.00%	93.00%	86.00%	86.00%	93.00%
Avg Accuracy	75.00%			92.00%			88.00%		



	Syllable repetition	Word Repetition	Prolangation
🟦LPC	66	66	93
🟥LPCC	93	90	93
🟩MFCC	93	73	100

Fig. 3. Average classification of each dysfluencies for LPC, LPCC and MFCC features

5 Conclusion

In this work, effectiveness of three speech parameterization techniques such as LPC, LPCC and MFCC are investigated in categorization of syllable repetition, word repetition and prolongation dysfluencies in stuttered speech. Multiclass SVM is used to perform classification. The recognition accuracy for parameterization techniques such as LPC, LPCC and MFCC is 75.00%, 92.00% and 88.00% respectively for all the three dysfluencies.

The experimental results demonstrate that the LPCC and SVM based system slightly outperforms because it has more discriminating capability. There is enough scope for extending the SVM with different kernel function to experiment on a larger corpus of dysfluent speech samples as part of future investigation.

References

1. Czyzewski, A., Kaczmarek, A., Kostek, B.: Intelligent processing of stuttered speech, vol. 21, pp. 143–171 (2003)
2. Bloodstein, O.: A handbook on stuttering. Singular Publishing Group,Inc., San-Diego (1995)
3. Chee, L.S., Ai, O.C., Hariharan, M., Yaacob, S.: MFCC based recognition of repetition and prolongation in stuttered speech using k-nn and lda. In: Procceedings of 2009 IEEE Student Conference on Research and Development (SCOReD), Malaysia (November 2009)
4. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press (2000)
5. Sherman, D.: Clinical and experimental use of the iowa scale of severity of stuttering. Journal of Speech and Hearing Disorders, 316–320 (1952)
6. Noth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., Wittenberg, T.: Automatic stuttering recognition using hidden markov models. Interspeech (2000)
7. Antoniol, G., Rollo, V.F., Venturi, G.: Linear predictive coding and cepstrumcoefficients for mining time variant information from software repositories. In: Proceedings of the 2005 International Workshop on Mining Software Repositories (2005)
8. Luts, J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., Suykens, J.: A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal. Chim. Acta 665, 129–145 (2010)
9. Proakis, J.G., Manolakis, D.G.: Digital signal processing. principles, algorithms and applications. MacMillan, New York
10. Ravikumar, K.M., Reddy, B., Rajagopal, R., Nagaraj, H.: Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. In: Proceedings of World Academy Science, Engineering and Technology, pp. 270–273 (2008)
11. Ravikumar, K.M., Rajagopal, R., Nagaraj, H.C.: An approach for objective assessment of stuttered speech using MFCC features. ICGST International Journal on Digital Signal Processing DSP 9, 19–24 (2009)
12. Rabiner, L., Juang, B.: Fundamentals of speech recognition. Prentice hall (1993)
13. Sin Chee, L., Chia Ai, O., Hariharan, M., Yaacob, S.: Automatic detection of prolongations and repetitions using lpcc. In: Proccedings of International Conference for Technical Postgraduates, TECHPOS (2009)
14. Lindasalwa, M., Begam, K.M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. Journal of Computing 2, 138–143 (2010)

15. Wisniewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., Suszynsk, W.: Automatic detection of disorders in a continuous speech with the hidden markov models approach. In: Computer Recognition Systems 2. ASC, vol. 45, pp. 445–453. Springer, Heidelberg (2008)
16. Wisniewski, M., Kuniszyk-Jozkowiak, W., Smolka, E., Suszynski, W.: Automatic detection of prolonged fricative phonemes with the hidden markov models approach. Journal of Medical Informatics & Technologies 11 (2007)
17. Howell, P., Huckvale, M.: Facilities to assist people to research into stammered speech. Stammering Research, 130–242 (2004); an Online Journal Published by the British Stammering Association
18. Howell, P., Sackin, S., Glenn, K.: Development of a two stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: Ii. ann recognition of repetitions and prolongations with supplied word segment markers. Journal of Speech, Language, and Hearing Research 40, 1085 (1997)
19. Mahesha, P., Vinod, D.S.: Automatic classification of dysfluencies in stuttered speech using MFCC. In: Proccedings of International Conference on Computing Communication & Information Technology (ICCCIT), Chennai, India (June 2012)
20. Prahallad, K.: Speech technology: A practical introduction topic: Spectrogram, cepstrumand mel-frequency analysis. Technical report, JCarnegie Mellon University and International Institute of Information Technology, Hyderabad
21. Schoslkopf, B., Smola, A.: Learning with kernals, support vector machines. MIT Press, London (2002)
22. Devis, S., Howell, P., Batrip, J.: The UCLASS archive of stuttered speech. Journal of Speech (April 2009)
23. SAwad, S.: The application of digital speech processing to stuttering therapy. In: Proceedings of Instrumentation and Measurement Technology Conference: IEEE Sensing, Processing, Networking, pp. 1361–1367 (1997)
24. Cullinan, W.L., Prathe, E.M., Williams, D.: Comparison of procedures for scaling severity of stuttering. Journal of Speech and Hearing Research, 187–194 (1963)
25. Young, M.A.: Predicting ratings of severity of stuttering (monograph), pp. 31–54 (1961)

RoboGardner: A Low-Cost System with Automatic Plant Identification Using Markers

Reema Aswani and N. Hema

Department of Computer Science Engineering
Jaypee Institute of Information Technology, Noida
Uttar Pradesh, India
reema_aswani@hotmail.com, hema.n@jiit.ac.in

Abstract. In this modern era, automation is inevitable. With the fast paced and busy lifestyles, there is no time even for day to day household activities. There is a need of automation in each and every small activity performed by humans. In this paper we present an autonomous system that caters to the need of automation in gardening by providing a low cost, portable and efficient system for watering indoor potted plants at home and offices. The system comprises of a mobile ROBO equipped with a camera for auto-location and identification of the plants and a sensing circuitry for analyzing the watering needs. The RoboGardner is designed to ease out human workload and performs all the functions without any human intervention. It even provides an automatic feedback mechanism to inform the user about its daily performance. The water level of the RoboGardner's on-board reservoir is also scrutinized automatically by it and the user is alarmed to refill when required. The paper outlines the complete architecture, functional modules and detailed implementation supported with the design circuits. It concludes with the system performance of the RoboGardner along with graphical analysis.

Keywords: Marker Identification, Zigbee technology, Mobile Robot, Obstacle Avoidance, Temperature Humidity Sensor.

1 Introduction

The presence of plants in our life is indispensible. As we know, plants add beauty to the environment, soothe the eyes, provide us oxygen and relieve stress preventing the rapidly growing diseases. Moreover, studies have consistently shown that by simply looking at environments with greenery like flowers and plants as compared to artificially created places lacking nature like rooms and buildings is significantly more effective in promoting recovery from stress [1] [2]. Research also suggests that by viewing places with plants or other nature for a few minutes can promote measurable restoration even in hospital patients who are acutely stressed.

Studies in environmental psychology also depict that the physical benefits of contact with nature and greenery are vanishing in towns and cities due to the disengagement from the natural environment because of the fast paced and modern

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 309–324, 2013.

lifestyle [3] [4]. It is claimed that the modern society and way of living especially in urban areas has insulated people from outdoor environment and regular contact with nature [3] [5].

In spite of the innumerable benefits of these plants people are reluctant to either keep these beautiful possessions or are not able to take care of them properly if planted. Unfortunately, the plants die due to insufficient water and care provided to them by their owners. Even this is the result of the busy routine life in today's modern society where time and user ease play an extremely important role in how one adopts their lifestyle.

In this paper we provide automation in the field of plant watering, so that these beautiful possessions can be taken care of automatically without being a burden on their owners [6]. The paper is an improvisation to the earlier system introduced [7], "Plant Watering Autonomous Mobile Robot", the system discussed in the paper had a lot of design constraints and assumptions to reduce complexity of the design. It followed a predefined path indicated by a black line using LDR sensors, the plants were identified using RFID (Radio Frequency Identification) Technology, and each plant was equipped with an RFID Tag. Although the system was scalable, the problem with it was that with each increasing plant the cost of the system increased as it had to be equipped with a tag. Moreover, the range of the RFID module was just 4-inches, beyond that the plant was not recognized, this made the system prone to missing the plants and thus inefficient. Since it was a line follower robot, a 5 cm thick black path had to be laid around the plants which spoiled the beauty of the environment and restricted the movement. The system lacked the ability to act intelligently in real time situations when it encountered an obstacle on its way and stopped without any movement further waiting for the user to clear the path.

The current system overcomes all the disadvantages of the previous design and provides a reliable, efficient, cost-effective and a completely autonomous mobile robot that is capable of taking care of the watering needs of the potted plants without creating any disturbance and inconvenience to the users around the area.

The system is now capable of avoiding real time obstacles autonomously that might come on its way while watering the potted plants. It also checks the water level and sends feedback to the user depicting the watering report. This saves a lot of time and makes the system more reliable and self sustainable. The system now uses a camera module for path planning and mobilization to the plants is achieved using the visual marker tags being identified which act as guidelines for the RoboGardner.

Section 2 discusses the related literature survey with a description of the existing plant watering systems. Section 3 gives an overview of the entire system with the architecture and system behavior. Further section 4 depicts the detailed implementation of the system with different modules for plant identification, watering operation, obstacle avoidance, and water level checker and feedback mechanism of the RoboGardner. The section also highlights the main features with their detailed description that distinguish our approach from already existing ones. Section 5 discusses the results of the proposed system supported with suitable statistics. The paper concludes with the system evaluation which presents the efficiency and reliability of the system and conclusion in Section 7.

2 Related Work

This section reviews the already existing systems designed for watering the plants. The problems faced by each of the portable system is object identification, path planning and obstacle avoidance.

A system designed by Sikorski, as a part of Intel Research performs the exactly same functionality [8]. The system had certain limitations which the RoboGardner tries to overcome. The system used high cost laser range finders to locate pot plants in the lab environment, which failed to work sometimes in the Lab as it contained glass surfaces where the laser finders bounced off. The system used high end hardware which made it very costly. Moreover, apart from that, the system also uses wireless network of Intel Research Lab. Lastly, the potted plants to be watered by the system were considered to be circles when detected by the range finder, with an assumption that only circular base pots were used, but in real life scenario the shape of the pots can vary.

Angelopoulos *et al.* presented a similar system for watering garden plants, the system comprises of TeleosB WSN motes attached to each plant, and each plant is also equipped with water valves that are triggered when the soil moisture of the corresponding plant goes down beyond a certain level [9]. The system performs the same action like the RoboGardner but is not portable, the valves are permanently attached to the plants and create disturbance to the user. Moreover, the valves laid all over spoil the natural beauty of the garden and give it an artificial look.

Zhou *et al.* proposed an approach for watering fields, a field irrigation system using the Zigbee Technology, the system used a star network topology and sensors were placed in different areas of the field along with Xbee devices [10]. The sensors continuously communicated with the base station Xbee attached to the portable controller, actuation is achieved by the valves and water pump attached at the in-field nodes. The system proposed is stagnant and applicable for field irrigation and cannot be used indoors.

The system designed by Kawakami *et al.* is an interactive plant pet where the plants are equipped with the sensors and wheels to achieve locomotion [11]. The plants remind their owners to water them by performing predefined actions. Although the system is scalable but it is not cost effective as each plant has to be placed on a separate mobile robot equipped with expensive hardware, increasing the cost of building the system to a great extent. Moreover, although these plants are equipped with ultrasonic sensors to prevent them from colliding with each other and the objects in the surroundings, if these moving plants are very large in number, moving here there, this might even disturb the user and might come on the way. Lastly, the system doesn't reduce the work load of the users; rather it is just like an alarm to remind them to take of their possessions.

The system proposed by Correll *et al.* is yet another robot gardener system which comprises of a mobile robot equipped with camera and water supply [12]. It is used to take care of potted cherry tomato plants. Each plant has an embedded Linux system to monitor the humidity and fruit ripening status along with watering operation. The whole system is centrally controlled by a central coordinator notebook which allocates tasks to all other components by receiving signal messages from the plants for water and harvesting and accordingly directs the mobile robot.

3 System Model and Architecture

This section states the system model of the system along with architecture and design details. Fig. 1 outlines the System Behavior of RoboGardner. The robot moves independently and locates the plants, waters them and moves to the next plant on its way. On encountering obstacles, it automatically takes up an alternative path to avoid any hindrance in progress.

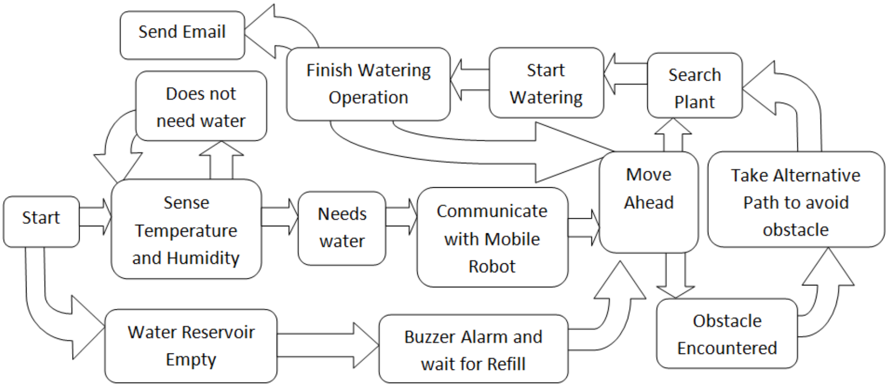

Send Email
Finish Watering Operation
Start Watering
Search Plant
Does not need water
Move Ahead
Take Alternative Path to avoid obstacle
Start
Sense Temperature and Humidity
Needs water
Communicate with Mobile Robot
Water Reservoir Empty
Buzzer Alarm and wait for Refill
Obstacle Encountered

Fig. 1. System Behavior

The design of the RoboGardner is implemented using the Arduino Simulator and MATLAB Simulink [13], besides it requires the interfacing of the hardware components described in this section to get the desired functionality [14] [15]. Fig. 2 presents the system architecture of the Autonomous System.

A. Xbee Series 2

XBee is a wireless communication module that Digi built to the 802.15.4/Zigbee standard. The main feature of the Zigbee is that it uses free bandwidth at 2.4GHz for wireless communication [16] . It is used to transmit the temperature and humidity of the plant to the mobile vehicle wirelessly. Since we are using a Xbee Series 2, the Outdoor RF line-of-sight range is 120 m and the Indoor/Urban range is 40 m which is quite sufficient for the system developed. The Xbee are configured to communicate with each other using XCTU software by setting the PAN ID and address.

B. Arduino Duemilanove

The Arduino Duemilanove controller which has an Atmega AT168 microcontroller that has a pre-installed boot loader is used, so one can download code to the board using an A-B USB cable. The USB cable is also used for serial communication while communicating with the image processing MATLAB module.

C. Temperature and Humidity Sensor Nodes

The system uses the temperature and humidity sensor module, the module comprises of the HSM-20g sensor attached with capacitors and resistors [16]. The input and output voltage is D.C. 5.0V ± 0.2V and D.C. 1.0V-3.0V respectively which is provided easily by the microcontroller.

D. Ultrasonic Distance Sensor

The ultrasonic distance sensor provides precise distance measurements from 2 cm to 3 m. The sensor works by transmitting an ultrasonic burst and provides an output pulse corresponding to the time required for the burst to return to the sensor. By measuring the echo width, the distance to target is calculated. In this scenario, the ultrasonic sensor is used for obstacle avoidance as well as automatic water level checker.

E. Water Pump and Relay

Keeping in mind the portability and to minimize the weight and bulkiness of the robotic system, a DC 12V water pump is used in the design of RoboGardner, and the system is plugged to a 12V DC adaptor. The water pump is triggered using a 6V relay which acts like a switch but instead of physically touching it to switch it on/off we supply voltage to toggle it [17].

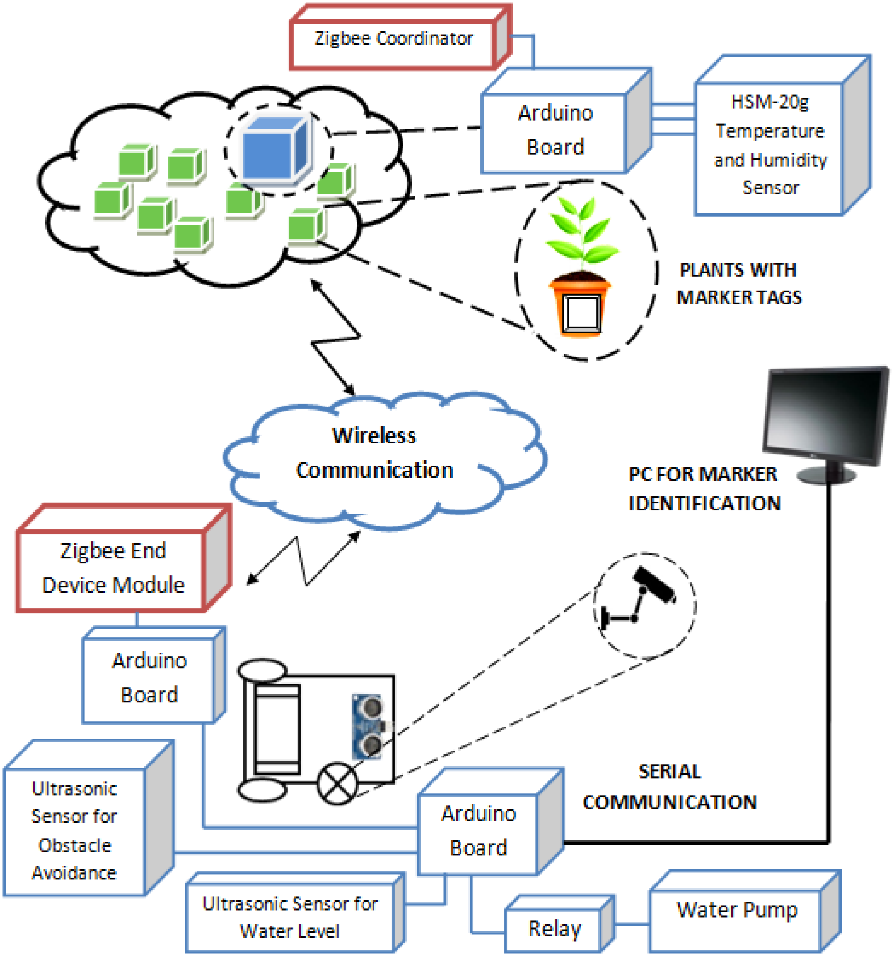


Fig. 2. RoboGardner System Architecture

4 System Implementation

4.1 Working Environment

The region to be watered by the RoboGardner is a square/rectangular field i.e. any compound area at home or any office balcony, entrance or gallery where potted plants are placed along the sides of the area. The camera placed on the RoboGardner can also be mounted on a half rotation servo motor so that it can identify and water plants placed anywhere in the environment rather than just along the sides.

The path to be followed by the robot needs to have some friction and no bumps. There should be direct view of the plant from the mobile robot's camera i.e. there should be nothing between the line of sight of the camera and the plant marker so that it can detect the marker properly.

The system can carry only limited amount of water in one go since the water carrying capacity is 1 liters to start with the demonstration due to the weight the DC motors can drive. This also restricts the number of pots and the size of potted plant to be watered. Fig. 3 depicts the improvised system layout and the working environment of the autonomous system.

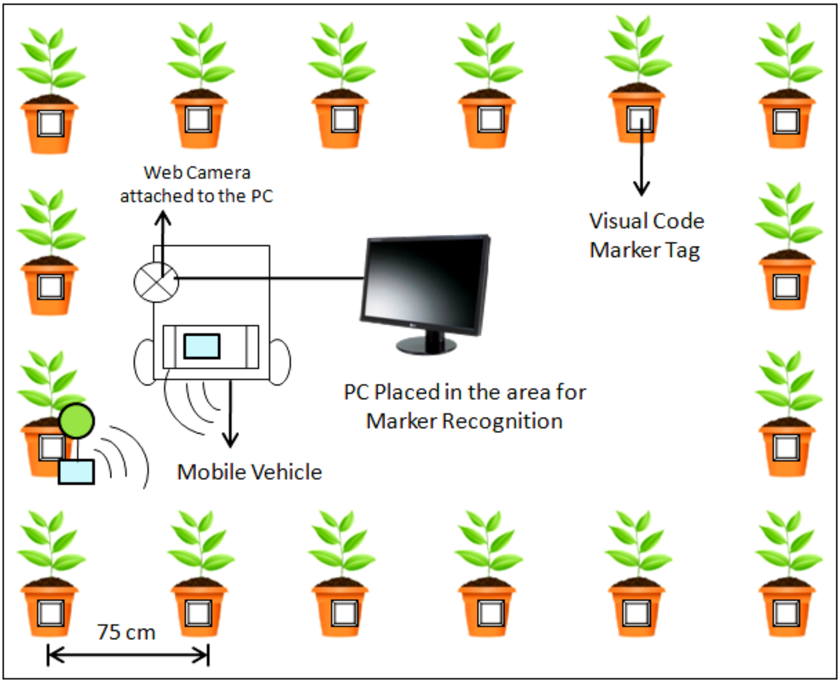

Web Camera attached to the PC
Visual Code Marker Tag
PC Placed in the area for Marker Recognition
Mobile Vehicle
75 cm

Fig. 3. Improvised System Layout of RoboGardner

The temperature and humidity sensor module is to be placed on the potted plant so that desired values can be obtained for proper functioning of the system. Since the area to be watered is a room, balcony or office gallery, the temperature and humidity of the area is assumed to be the same and thus only one temperature humidity sensor is placed along with Zigbee Coordinator to monitor the watering requirements of the plants. In case of a large area, or indoor/outdoor differences we can use more than one sensor to monitor the area.

The obstacle detected by ultrasonic sensor is taken to be of fixed length and width and accordingly, the timing for turns have been given for alternative path to be followed.

The height of the potted plant is taken as per the height of the mobile robot so that water can be dispensed easily from the on-board water tank to the plant using the water pipe.

4.2 Functional Modules of the System

The RoboGardner performs different functionalities to complete the watering process i.e. locating the plants, finding its way, plant identification, checking the water level, sending alerts to the system. This section gives a brief description of each of the five modules into which the RoboGardner can be broadly classified.

A. Plant Location and Identification
For plant identification visual code marker tags are placed at each potted plant. While moving, the Autonomous robot detects the marker tags placed on the plant pots via the attached web camera unlike in [18] where ultrasonic sensor is used. The idea of these visual code marker tags has been taken from augmented reality applications but a basic image processing algorithm [19] has been used instead of complex AR identification as the need here is just single tag identification. The tag being used in the design is shown in Fig. 4. Fig. 5 shows the snapshots of the image processing phases. The marker identification algorithm involves three main phases:

(I). Pre-processing:
The pre-processing step produces a binary image from the input RGB image captured by the web camera attached at the mobile vehicle that has an accurate representation of the visual code marker in the image. In this way we do not lose any relevant information in the process and it is a simple, computationally efficient algorithm to read the embedded data bits of the captured image [20].

(II). Shape Detection:
Once the image is preprocessed, and a binary image is obtained, we use shape detection algorithm to detect the corner stone circles.

(III). Visual Marker Detection:
After detecting the corner stones from the pre processed image, the centroid of each of the region is calculated and the distance between them is measured so that the orientation can be decided. The orientation, number of pixels and the area of the identified regions in the marker tag are the parameters to identify the tag.

B. Watering Operation
As soon as the robot reaches a plant and the visual marker tag is identified the robot stops and dispenses the water to the plant via the attached water pump and the on-board water reservoir.

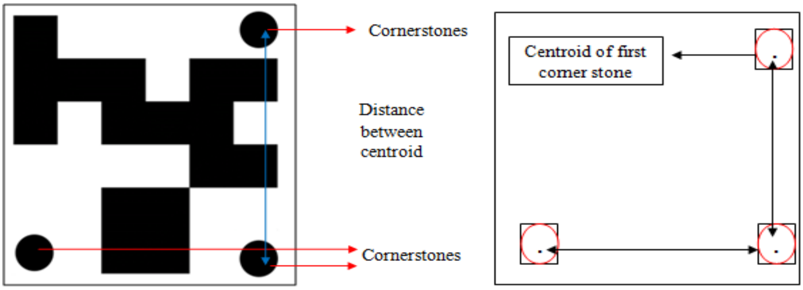


Fig. 4. Visual Code Marker Tag

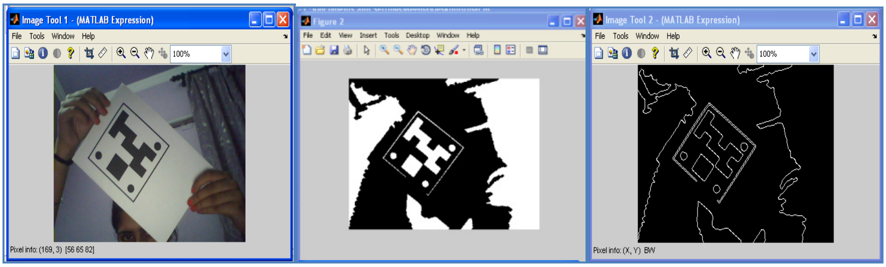


Image Tool 1 - (MATLAB Expression)
File Tools Window Help
100%
Pixel info: (169, 3) [56 65 82]
Figure 2
File Edit View Insert Tools Desktop Window Help
Image Tool 2 - (MATLAB Expression)
File Tools Window Help
100%
Pixel info: (X, Y) BW

Fig. 5. (a) Captured Image **(b)** Binary Image after Thresholding **(c)** Edge Detection

C. Obstacle Avoidance

Since the RoboGardner has to move autonomously, its movement might get hindered due to obstacles in its path. The obstacles might come on the way real time, the mobile vehicle has to be programmed for handling the obstacles that might come accidently. For this, we need obstacle avoidance which enables the robot to take an alternative path to avoid the hindrance, so that its movement is not hampered.

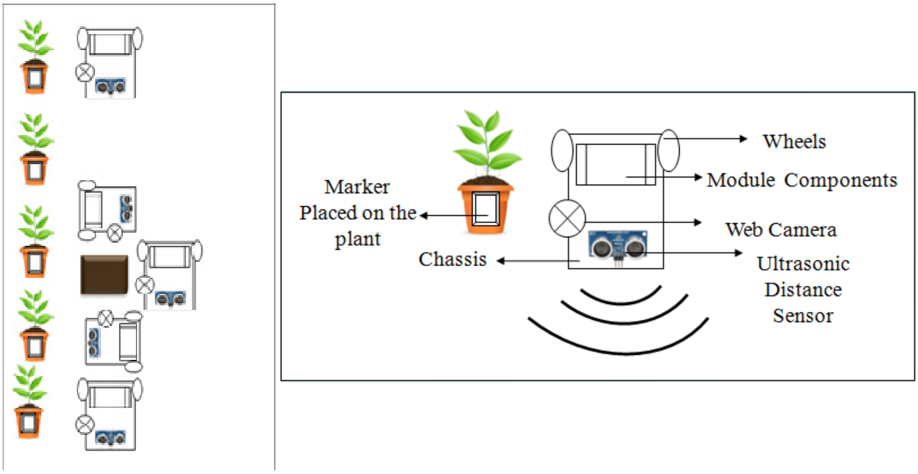


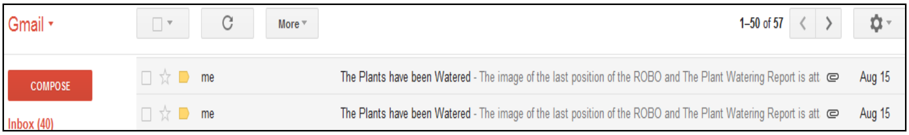
Fig. 6. Obstacle avoidance Mechanism

The ultrasonic distance sensor is used for this purpose of obstacle avoidance as discussed by Borenstein and Koren in [21]. It is placed in the front of the mobile robot, so that as it moves forward, the sensor keeps on detecting the obstacles which helps the robot to take an alternative path to avoid the same.

The Fig. 6 shows the approach taken by the ROBO to avoid obstacles coming in its way. The ROBO senses the obstacle from a distance of 50 m and starts diverting its path by taking a right turn for a predefined distance and then covering the same distance back to reach its original path to continue the watering operation.

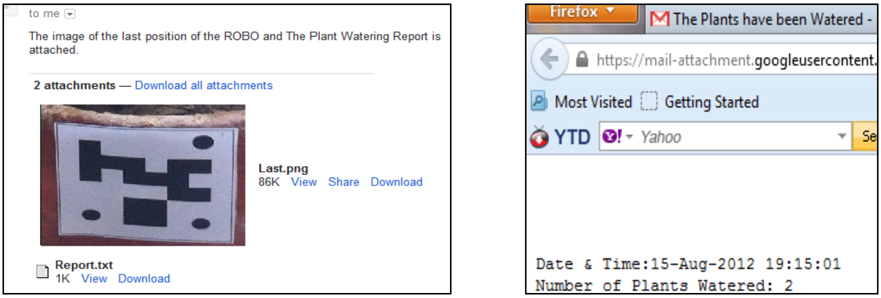
D. Automatic Feedback Mechanism

The RoboGardner also has an automatic feedback mechanism. As soon as it completes one cycle of its watering operation, it dynamically generates a feedback report comprising of a text file and an image. The text file comprises of the date and time when the watering operation was completed and the number of plants watered in the cycle. The image depicts the last position of the robot. Fig. 7 and Fig. 8(a) shows the snapshot of the Automatic Email Mechanism.



me	**The Plants have been Watered** - The image of the last position of the ROBO and The Plant Watering Report is att	Aug 15
me	**The Plants have been Watered** - The image of the last position of the ROBO and The Plant Watering Report is att	Aug 15

Fig. 7. Automatic Email Sent by the System


to me
2 attachments — Download all attachments
Last.png
86K View Share Download
Report.txt
1K View Download
Firefox
The Plants have been Watered -
https://mail-attachment.googleusercontent.
Most Visited
Getting Started
YTD
Yahoo
Se

The image of the last position of the ROBO and The Plant Watering Report is attached.

```
Date & Time:15-Aug-2012 19:15:01
Number of Plants Watered: 2
```

Fig. 8. (a) Email Details **(b)** Plant Watering Report

This report is automatically mailed to the user so that the user gets to know the complete watering scenario by just seeing the report at the end of the day. The text file shown in Fig. 8(b) is dynamically created each time the robot completes watering a set of plants, further, the report is attached and mailed to the user automatically by the RoboGardner. This automatic mailing action is performed using the MATLAB interface which configures the Gmail account of the user where the email address and the password are defined in the code. As the system keeps on sending the mail updates, it simultaneously creates an excel file using MATLAB and Microsoft Excel interfacing, and a graph depicting the performance of the RoboGardner after each cycle, the graph after 10 cycles is simultaneously sent to the system developer for performance evaluation.

E. Automatic Water Level Checker

As the robot autonomously moves around to locate and water plants, it is equipped with a sensor to detect the water level of the on-board reservoir. Since the water carrying capacity of the robot is limited, the water in the reservoir has to be refilled by the user manually. The system generates an alarm using a buzzer attached to it as the water is about to get over.

For detecting the water level, the Ultrasonic distance sensor is used which is placed at the brim of the reservoir container and faces the water reservoir's base. The distance of the water from the sensor is sensed continuously as it moves around to water the plants. A threshold is maintained and programmed in the Arduino Duemilanove board. As the robot keeps on watering the plants, the water in the reservoir keeps on decreasing and the distance values sensed by the ultrasonic sensor keep on increasing as the distance between the water surface and the sensor increases.

As the value reaches the threshold, the buzzer alarm turns on and the robot stops its movement, simultaneously an email is sent to the user stating that the water reservoir needs refilling.

4.3 Complete Working Model

This section focuses on the detailed description of the working system model of the RoboGardner. Apart from the main functional modules discussed in the previous section, the system can be broadly modularized into two modules, one is the mobile vehicle and the second one is the sensing module placed at the plants. Fig. 10 depicts the complete Circuitry of the RoboGardner.

Sensing Module

The sensing module placed at the plant comprises of the Zigbee Coordinator, Arduino Duemilanove board and Temperature and humidity sensor HSM-20g. The temperature and humidity sensor sends a continuous stream of data via the Zigbee Module about the watering needs of the plants. Fig. 9(a) depicts the Sensing Module that is placed at the plant.

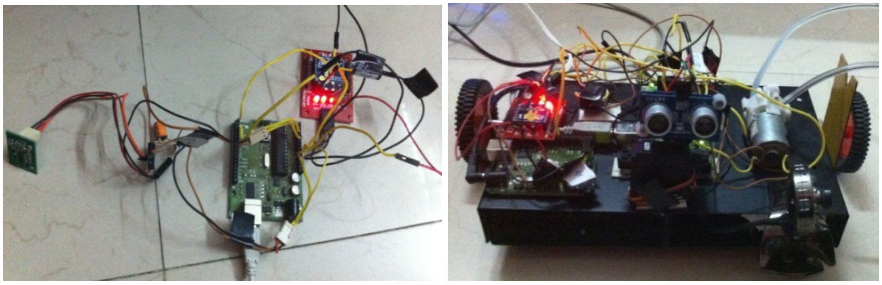


Fig. 9. (a) Sensing Module **(b)** Mobile Vehicle

A preset threshold is maintained, the values sensed by the sensor are compared to the threshold, if the values do not fulfill the threshold condition of the temperature and humidity, the Arduino board attached to the sensor sends a 'Y' to the mobile vehicle via the Zigbee Coordinator device indicating that the plants require water.

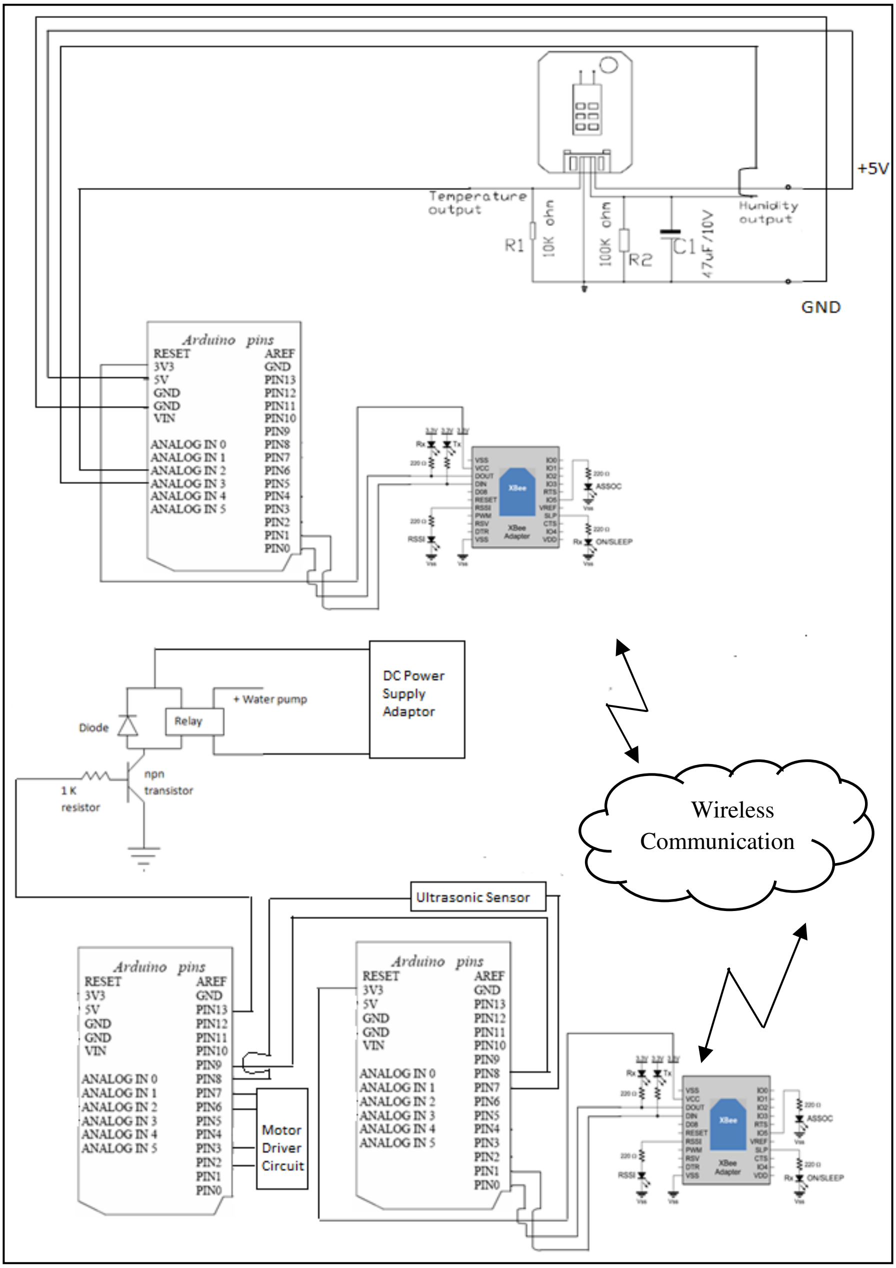


Fig. 10. RoboGardner Complete Circuit

In case, the values fulfill the threshold criterion, an ‘N’ is sent. The Zigbee sends the ‘Y’/’N’ continuously to the mobile vehicle depending upon the need of the plants being taken care of.

Mobile Vehicle

The mobile vehicle is equipped with the Zigbee Receiver, configured as the End Device, Full rotation DC motors, two Arduino Duemilanove boards, Web camera, Relay circuit, two ultrasonic distance sensors, DC Water pump and the on-board water reservoir. The Zigbee End Device keeps on receiving data from the Zigbee Coordinator placed at the plant, as soon as it receives a ‘Y’ the RoboGardner starts moving towards the plant to water them. Fig. 9(b) shows the mobile vehicle equipped with the sensors and other hardware components.

Each plant is equipped with a visual code marker tag; the camera attached on the mobile vehicle continuously takes snapshot while travelling towards the plant, each image captured by the camera is sent to the PC via serial communication. The image is processed used appropriate program written in MATLAB as discussed in Section 4.2 (A).

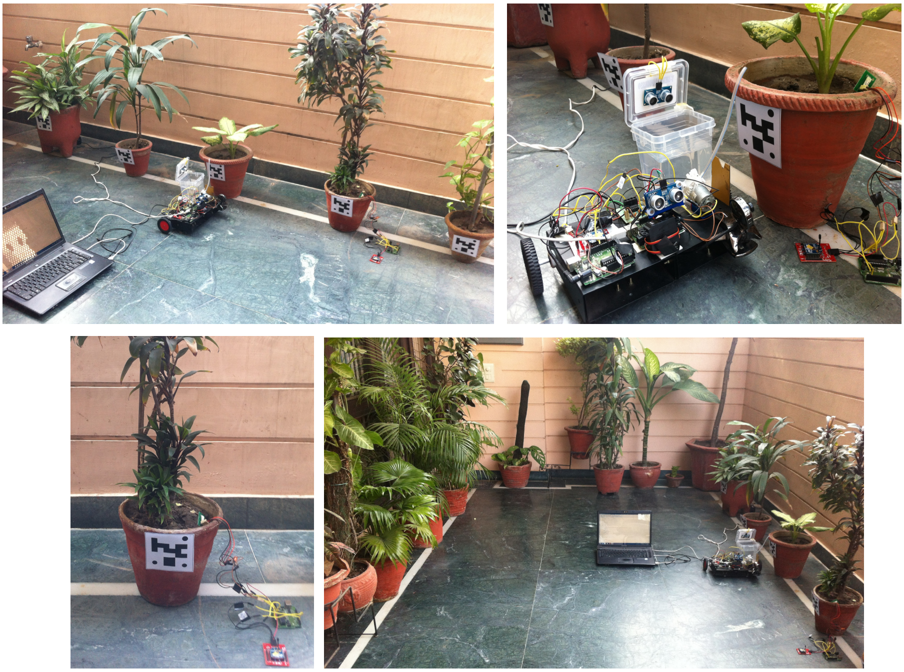


Fig. 11. RoboGardner performing the watering operation in real time scenario

Based upon the image captured, if the desired marker tag is present, an ‘S’ signal is sent to the microcontroller, indicating that a plant is detected. The robot stops its motion, turns on the relay as discussed in Section 4.2 (B) and triggers the attached DC water pump to dispense the water. The watering process requires 10 seconds after which the robot starts its locomotion again in search of the next plant to water. The same process repeats as it keeps on detecting the plants by identifying the attached marker tags. The last plant of each row is equipped with a different set of marker tag, as soon as the RoboGardner waters the last plant of each row an ‘R’ signal is serially communicated to it, indicating that a right turn is to be taken. In this way the RoboGardner covers the entire square/rectangular region.

5 Results and System Evaluation

This section outlines the results and depicts the performance of the system. The RoboGardner is designed as an improvisation of the previously designed system and to overcome the shortcomings of the already existing systems. The system is far more cost-effective, portable, light weighted, maintainable and efficient when compared to different systems on the stated parameters. The system is reliable and waters the indoor plants autonomously in no significant time.

The Fig. 11 depicts the RoboGardner performing the watering operation in a real time environment of compound area at home. As shown in the Fig the plants are equipped with visual code marker tags and the first plant is equipped with sensing module, the RoboGardner keeps on watering the plants and moves to the other one. The water dispensed is taken to be 100 ml per plant for experiment purpose which can be increased according to the design as and when required.

Time Analysis

It uses a 60RPM 12V D.C. motor. The system is evaluated on the basis of the time required to reach the plant, the watering operation and the number of plants successfully watered. The results are supported with the appropriate statistics and graphical analysis.

$$D=RPM \times dR \tag{1}$$

Where, D= distance travelled in per minute
RPM= Rotations per minute of the DC motor (60 in this case)
dR= distance travelled in one rotation (15 cm)

$D=60 \times 15 = 900$ cm/min
Distance between the centers of two plants is assumed to be 75 cm.
Number of plants to be reached in 1 min without stopping = 900/75= 12
Time required to reach 1 plant $T_R = 60/12$ seconds = 5 seconds.
Time consumed in each watering operation $T_W = 10$ seconds.
Total time required to water 1 plant = $T_R + T_W = 15$ seconds.

Based on the above mentioned statistics, the report generated by the robot depicts its performance in 10 consecutive watering rounds. The Fig. 12(a) shows the number of plants watered in each round in the given number of rounds. The Fig. 12(b) depicts the total time to water the given number of plants corresponding to the stated number in Fig. 12(a) , Time Analysis, the graph shows the linear curves for the time to reach the plant, time to water the plants and the total time spent on the complete watering process in each specified watering round. The X-axis in both curves depicts the watering round number. The y- axis in Fig 12(a) shows the time in seconds.

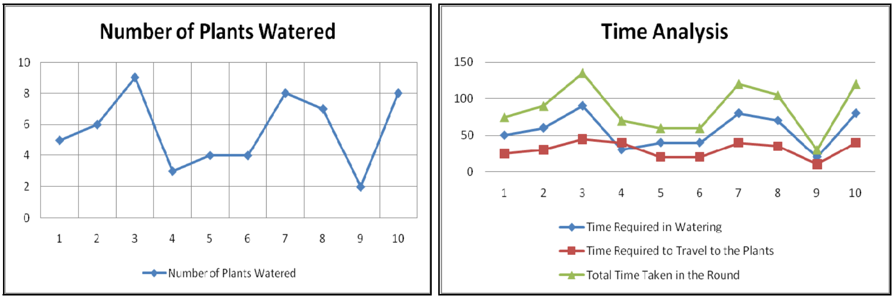


Fig. 12. (a) Watering in each Round **(b)** Time Analysis of RoboGardner's Complete Process

Cost Analysis

Table 1 depicts the comparison of the RoboGardner with the Intel's PlantCare System in terms of cost. The total cost of Intel System's major components comes out to be approximately $9000 whereas the RoboGardner in $190. The system presented in this paper is approximately 50 times more cost-effective and is capable of automatically locating, identifying and watering plants according to the watering needs.

Table 1. Cost Analysis and Comparison of RoboGardner and Intel PlantCare System

Intel Research PlantCare System[1] (Major Components)			RoboGardner: A Low-cost System with Automatic Plant Identification using Markers		
Component	Quantity	Cost [2]	Component	Quantity	Cost
SICK Laser Range Finder	1	$5000	Arduino Duemilanove	3	Rs. 1600
Pioneer 2-Dxe Robot	1	$3995	Web Camera	1	Rs. 500
Lead-Acid Batteries	3	$68.44	Dc Motors	2	Rs. 300
Sonar Sensor	8	$236	Relay	1	Rs. 50
			Xbee Series 2	2	Rs. 5000
			Hsm-20g sensor	1	Rs. 800
			Water pump	1	Rs. 250
			Ultrasonic Distance Sensor	2	Rs. 2000
Total Cost (approx.)		$9299.4			Rs.10,500 ($188) [3]

[1] Apart from this it uses the wireless network of Intel Research Lab.

[2] Prices as per ActiveMedia pricelist June 2002.

[3] According to the conversion of Rupees to Dollar (1$=Rs 56).

6 Conclusion

In this paper, we presented a low-cost autonomous plant watering system which comprised of a temperature and humidity sensing module and a mobile vehicle for plant identification and watering. The RoboGardner is designed keeping in mind the need of automation in the watering of plants at homes and offices. It is successfully implemented to water indoor potted plants placed on plain even surface. It senses the temperature humidity of the plants being taken care of, locates and identifies them autonomously and waters them without any human involvement. Apart from the watering operation, the system includes different functional modules like obstacle avoidance, water level checker and automatic feedback mechanism to enhance the usability and reliability of the system.

7 Future Work

In the future, the functionality of the RoboGardner will be further enhanced to ensure a completely reliable and efficient autonomous system development. The web camera which communicates with the PC/ Laptop for marker identification will be replaced with a Serial JPEG Camera Module-TTL connection which is compatible with Arduino Duemilanove board. With this replacement, the image processing would be directly done at the microcontroller without the need of an attached PC. The work in this area is ongoing.

Moreover, in this paper the system is designed keeping in mind that the plants are placed along the sides of the desired watering area in a straight line and the robot is aware of the plant positions and is initially placed in close proximity to them. Work is ongoing in implementing an algorithm for path planning where the RoboGardner will not have a priori knowledge of working space. It would be designed to work viewed as a grid with plants being viewed as points.

The obstacle avoidance would also be improvised using this path planning algorithm for unknown environment and the alternative path to be taken would also be either clockwise or anticlockwise depending upon the real time scenario.

The robot can be equipped with a mechanical arm to increase the usability so that it can water any plant regardless of the height of the pot.

References

1. Ulrich, R.S.: Communicating with the healthcare community about plant benefits. In: Shoemaker, C. (ed.) Proceedings of the Sixth International People Plant Symposium, Chicago Botanic Garden, Chicago (2002)
2. Ulrich, R.S.: Effects of gardens on health outcomes: Theory and research. In: Cooper-Marcus, C., Barnes, M. (eds.) Healing Gardens: Therapeutic Benefits and Design Recommendations, pp. 27–86. John Wiley, New York (1999)
3. Katcher, A., Beck, A.: Health and Caring for Living Things. Anthrozoos 1, 175–183 (1987)

4. Stilgoe, J.R.: Gone Barefoot Lately? American Journal of Preventative Medicine 20, 243–244 (2001)
5. Axelrod, L.J., Suedfeld, P.: Technology, Capitalism, and Christianity: Are They Really the Three Horsemen of The Eco-Collapse? Journal of Environmental Psychology 15, 183–195 (1995)
6. Marti, M., Anastasiades, B.: Thesis: If plants could walk. Autonomous Systems Lab (2010)
7. Nagaraja, H., Aswani, R., Malik, M.: Plant Watering Autonomous Mobile Robot. IAES International Journal of Robotics and Automation 1(3) (2012) ISSN: 2089-4856
8. Sikorski, K.: Thesis- A Robotic PlantCare System. University of Washington. Intel Research (2003)
9. Marios Angelopoulos, C., Nikoletseas, S., Constantinos Theofanopoulos, G.: A Smart System for Garden Watering using Wireless Sensor Networks. In: MobiWac 2011 Proceedings of the 9th ACM International Symposium on Mobility Management and Wireless Access, pp. 167–170. ACM, New York (2011)
10. Zhou, Y., Yang, X., Wang, L., Ying, Y.: A wireless design of low-cost irrigation system using Zigbee technology. In: IEEE International Conference on Networks Security, Wireless Communications and Trusted Computing (2009)
11. Kawakami, A., Tsukada, K., Kambara, K., Siio, I.: PotPet: Pet-like Flowerpot Robot. In: Tangible and Embedded Interaction 2011, pp. 263–264. ACM, New York (2011)
12. Correll, N., Arechiga, N., Bolger, A., Bollini, M., Charrow, B., Clayton, A., Dominguez, F., Donahue, K., Dyar, S., Johnson, L., Liu, H., Patrikalakis, A., Robertson, T., Smith, J., Soltero, D., Tanner, M., White, L., Rus, D.: Building a Distributed Robot Garden: Intelligent Robots and Systems. IEEE Journal/RSJ International Conference (2009)
13. MathWorks India - MATLAB and Simulink for Technical Computing, `http://www.mathworks.in`
14. Wu, K., Eng, B.: Thesis: Realtime Control of a Mobile Robot Using Matlab. The University of Applied Science Hamburg (2004)
15. UC Hobby, Visualizing sensor data with Arduino processing, `http://www.uchobby.com/index.php/2009/03/08/visualizing-sensor-with-arduino-and-processing/`
16. XBEE Tutorial, `https://sites.google.com/site/xbeetutorial/example`
17. Humidity sensor calibration, `http://msl.irl.cri.nz/research/temperature-and-humidity/humidity-sensor-calibration`
18. Harper, N., McKerrow, P.: Recognizing Plants with Ultrasonic Sensing for Mobile Robot Navigation. Robotics and Autonomous Systems 34(2) (12), 71–82 (2001)
19. Kulji, B., János, S., Tibor, S.: Determining Distance and Shape of an Object by 2D Image Edge Detection and Distance Measuring Sensor. In: IEEE 6th International Symposium on Intelligent Systems and Informatics, SISY (2008)
20. Bruce, J., Balch, T., Veloso, M.: Fast and Inexpensive Color Image Segmentation for Interactive Robots. In: IEEE/RSL International Conference on Intelligent Robots and Systems, vol. 3, pp. 2061–2066 (2000)
21. Borenstein, J., Koren, Y.: Obstacle Avoidance with Ultrasonic Sensors. IEEE Journal of Robotics and Automation 4(2) (1988)

6H-SiC Based Power VJFET and Its Temperature Dependence

Amir Khan, Mohd. Hasan, Anwar Sadat, and Shamsuz Z. Usmani

Dept. of Electronics Engineering, Aligarh Muslim University
Aligarh (U.P.), India
{amiramu10,anwart7039}@gmail.com

Abstract. In this paper 6H-SiC VJFET has been shown and the s device characteristics are also shown. Further the optimization is also carried out with respect to the breakdown voltage. 6H-SiC VJFETs breakdown characteristic is also plotted and the parameter dependence of breakdown voltage is also shown. Electric field across the channel and across the channel and across the device length is also plotted.

Keywords: Breakdown voltage, VJFET, channel, doping, SiC.

1 Introduction

Silicon Carbide (SiC) is a new material in the field of power devices. Before the advent of SiC this field was mainly occupied by the Silicon (Si). Now this ubiquitous material (Si) has to share the spot light. Since the Si limit has almost reached in the power applications so the advent of the SiC is good news for power devices industry. For the power applications we need higher band gap energy and higher value of critical electric field which results in higher value of breakdown voltage (BV) and at the same time we need lower value of specific on resistance which further demands higher value of critical electric field and higher value of carrier mobility. SiC has almost 3 times band gap energy than Si and also almost 10 times higher value of critical electric field than Si and these make SiC as material of choice for power devices. Also the SiC has thermal conductivity almost 3 times than that of Si. For the same value of breakdown voltage SiC needs thinner device than Si. All these properties make SiC superior than Si. Many SiC devices like Schottky Diode, VJFET and MOSFET have already arrived to the market. But MOSFETS and IGBT etc., of Si dominate over SiC. Despite the recent advances, the quality of fabricated oxide layer in SiC technology is inferior to that in Si, which adversely affects the mobility of carriers in the channel of the device and hence increases the on-state resistance. Also the reliability of oxide at elevated temperature is still a point of concern. Because of the oxide reliability problem SiC loses the importance of operation at higher temperature.

SiC JFETs, on the other hand, are free of gate oxides. Therefore, they can fully benefit from the superior properties of SiC without being compromised with the poor

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 325–332, 2013.

quality of material interfaces. In recent years a number if designs have been developed for the power JFET structures in SiC. A number of these involve a horizontal current control channel above the drift region whilst others use a purely vertical structure. There are two types of JFET namely normally-on and normally-off using the depletion and enhancement mode channel respectively. Normally-on devices are unsuitable for the applications where the power is applied at the start up, such as power supplies and low power drives, where normally-off devices are preferred. The development of enhancement mode VJFETs is necessary for high temperature and high power applications. SiC is said to have more than 200 poly type. The poly type whose wafer is available in the market are 4H, 6H and 3C and the 4H poly type is most matured in terms of device fabrication because of higher value of mobility parallel to c-axis than all other poly type. Table 1 compares the properties of Si and three SiC poly type. 6H-SiC has high value of critical electric field than 4H-SiC and all other poly type and material.

The band gap energy (E_G) in 6H-SiC, as a function of temperature, is approximated by equation (1). Reduced band

$$E_G(T) = 3.0\text{-}3.3\text{x}10^{-4}\,(T\text{-}300)\text{ eV} \qquad (1)$$

gap at higher temperature results in larger intrinsic carrier densities, larger leakage currents in p-n junctions, poorer junction rectification in power switching devices, and poorer device isolation by reverse-biased junctions.

2 VJFET

This is a new device for the new material SiC but not for the Si. Because of the reliability issues related to the SiC/SiO$_2$ interface at elevated temperature SiC based MOSFETs are not matured fully. So VJFET is a suitable candidate for switch.

Table 1. Electrical and Physical properties of silicon and silicon carbide

Property	Material			
	Si	*3C SiC*	*6H SiC*	*4H SiC*
Dielectric Constant	**11.8**	**9.7**	**9.7**	**9.7**
Energy gap (eV)	**1.1**	**2.39**	**3.03**	**3.26**
Critical field E_c (MV/cm)	**0.3**	**1.5**	**3.2**	**3.0**
Electronics mobility μ_n(cm^2/Vs)	**1400**	**`750**	**370**	**800**
Electron Drift Velocity v_{sat} (×=10^7 cm/s)	**1**	**2.5**	**2**	**2**
Thermal conductivity $_k$ (W/cm K)	**1.5**	**5**	**4.9**	**4.9**

operation. Similar to the MOSFET depletion mode VJFET, operates at zero gate bias and negative gate to source voltage is needed to turn it off. The enhancement mode device (JFET) has channel has no channel when V_{GS} = 0V because the depletion regions formed by the gates reduce the conducting channel width and a positive bias turns the device on by creating the channel in the device.

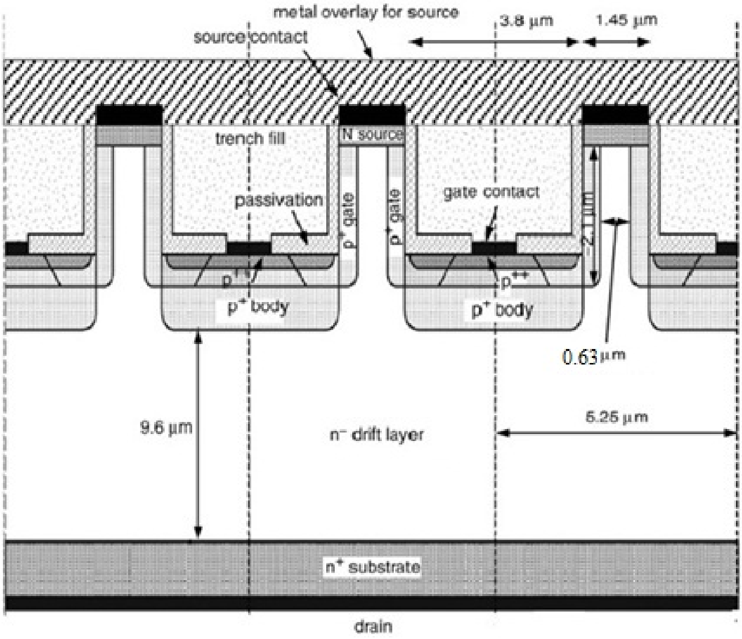


Fig. 1. Cross section of TIVJFET

In order to maintain the unipolar conduction and avoid the significant gate leakage in the on state the maximum allowable gate to source bias should not be more than the 3V which is the built in potential in the case of SiC.

Enhancement mode VJFET is normally-off and is mainly used in the fail-safe operation. In the ON-state majority carriers (electrons) flow vertically from source to drain. To control current through the device gate terminal is subjected to voltage which modulates the depletion width at the junction of n-type channel and p-type gate.

A. Breakdown Voltage (BV)

The breakdown phenomenon ascertains the highest voltage that can be applied and also limits the maximum power it could withstand. In the normally-off device BV is the maximum voltage that can be applied to drain under the condition that gate voltage is kept at 0V, after which the device starts conducting heavily. Breakdown voltage depends on the device parameter like length of the device. Channel width, doping concentration etc. Increasing the length of the device causes the increase in the BV because the JFET sustain this voltage across the drift region. Also reduction in the doping of the drift region causes the increase in the breakdown voltage as drift region has lower number carriers and breakdown voltage increases. Reduction in the width of channel causes the increment in the breakdown voltage because of the reduction in carriers supplied by source for same drain voltage. These all parameters need to design carefully to fully optimize the device.

B. Fabrication

The device parameters used for the fabrication of the SiC VJFET are to be selected carefully. For the demonstrated VJJFET the active area of the device is 320x293 μm^2. Drift region (n$^-$-type) doping is kept at 6.5x10^{15} cm^{-3}. To make the SiC n-type nitrogen and phosphorus are more preferred than other n-type dopant because of smaller size. Doping concentration of source and drain (n$^+$-type) is kept same at 2.0x10^{19}cm^{-3}. The trench region is comprised of 50-nm thick thermal oxide, followed by 200-nm thick silicon nitride layer and remaining portion is filled with oxide. The vertical channel is designed to have an opening of 0.63 μm and the length of the blocking layer (drift region) is kept approximately 9.4 μm. Designing of VJFET is done for these values and with the variation of these parameters and temperature optimization of VJFET will be carried out with respect to breakdown voltage.

3 Simulation Results

For the nominal values mentioned above the device simulation is carried out and the breakdown characteristic of is shown in the figure 2. Figure shows that the breakdown takes place at a drain voltage of 1150V. Before breakdown we can see that drain current density is almost zero making it near ideal switch.

As already discussed breakdown takes place under off condition for a gate to source voltage equal to 0V. As the gate voltage is zero there is no channel created and depletion width covers the channel and hence no current flow in the device. Also device can withstand larger voltage across the source and drain. As the drain voltage continue to increase, at the breakdown point suddenly a large current density is created in the device and device starts conducting heavily. For the actual value of the breakdown voltage we have included the impact ionization which is modeled by van Overstraeten-de Man for the 6H-SiC.

Figure 3 shows the I_d-V_d characteristic of the 6H-SiC VJFET for the said parameters for different values of gate voltage (source is kept grounded). Drain current is zero for zero gate voltage and increases with increase of gate voltage for same drain voltage. VJFET starts conducting sufficiently for gate voltage greater than 2V. Id-V_d curves are drawn for gate voltages greater than 3V because of drop in the contact resistances. Higher value of drain current is needed which is achieved at higher value of gate potential. Although the value of drain current is still small but this is sufficient to keep device in the on-state.

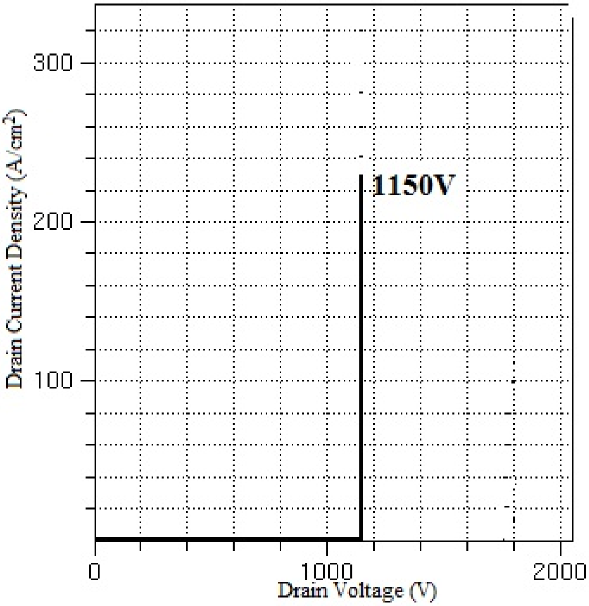


Fig. 2. Breakdown Characteristic of 6H-SiC VJFET

Figure 4 shows the variation of the gate current for different gate voltages. We want this leakage current as small as possible because it hinders the controllability of the drain current (main current) of the device. Logarithmic value of gate current is plotted with drain voltage for different gate voltages increment in the value of gate current is from the order of 10^{-4} to 10^{0} which is very large. The increment in the gate current is because of the forward biasing of p^{+}-n^{-} diode (formed between gate and drift region) becomes more and more forward bias. The increment in the gate voltage is advantageous for the drain current but the corresponding increase in the gate current is unwanted. We want higher value of drain to gate current ratio. For 6H-SiC VJFET at 300K this ratio comes out to be 1798 for a gate voltage of 2.5V. This ratio is good enough at 2.5V. But this ratio decreases with increase of gate voltage.

A. Temperature Dependence

The advantage of the SiC is its operability at elevated temperature. There are variations in the different parameters with temperatures but these variations are smaller than Si and also the range of temperature wider

Figure 5 shows the different I_d-V_d curves at different temperature at a constant gate voltage of 2.5V. Increment in the drain current with the increase of temperature is because of ionization of carriers increases with the increase of temperature as all impurity atoms are not ionized at room temperature. This increase in carriers results in increment of drain current.

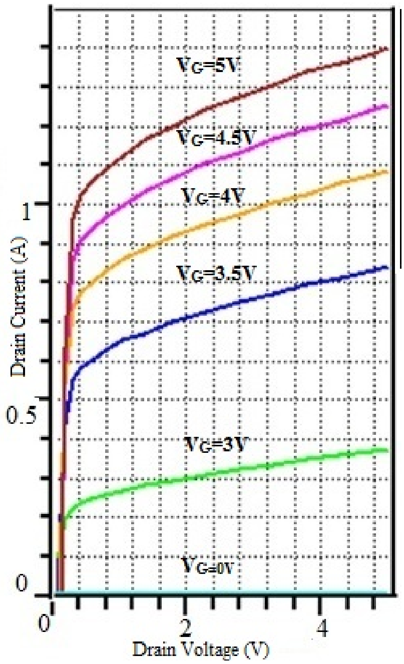


Fig. 3. Id-Vd characteristic of 6H-SiC VJFET

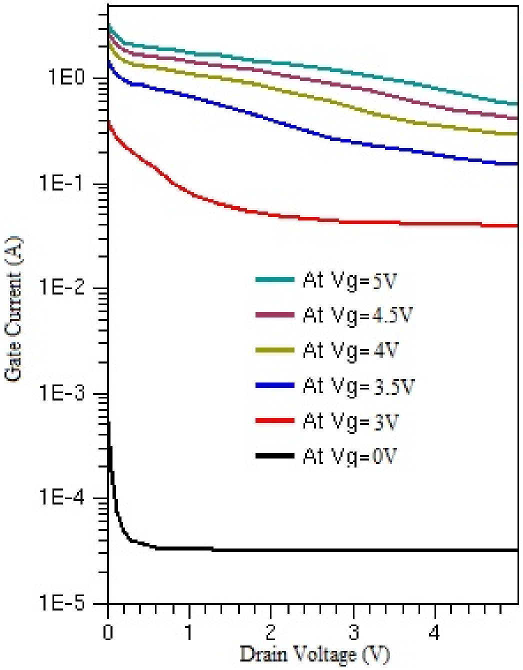


Fig. 4. Gate current at different gate voltages

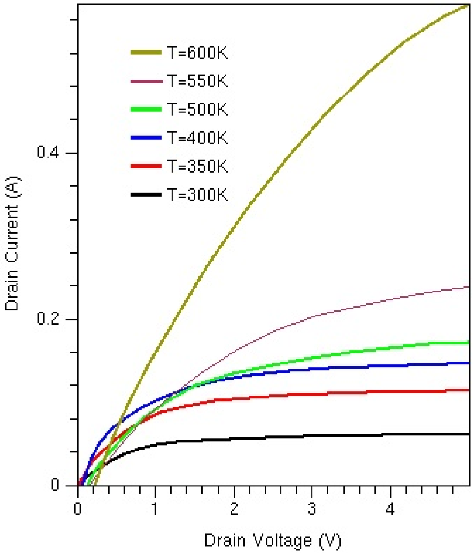


Fig. 5. Id-Vd characteristic variation with temperature

Figure 6 shows the variation of gate current with temperature at a gate voltage of 2.5V. Increase in the temperature results in the increase of ionized carrier and reduction of depletion region which results in the increase in gate current with increase in temperature.

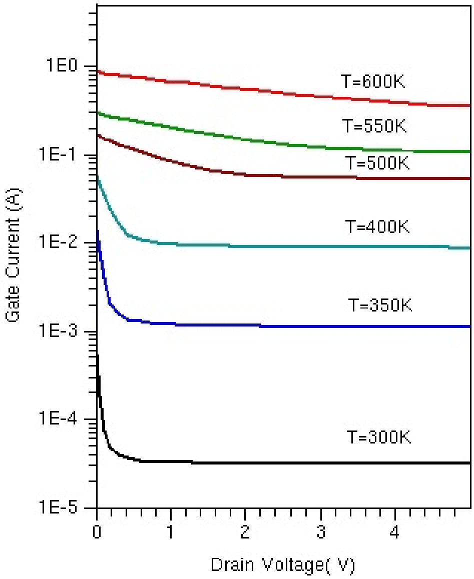


Fig. 6. Gate current at different temperature

4 Conclusions

6H-SiC VJFET is created with the Sentaurus TCAD. Breakdown voltage is found to be 1150V. I_d-V_d characteristic is plotted for different gate voltage. Gate current is also plotted at different gate voltage. The optimum gate voltage is found to be 2.5V as the increase of gate voltage causes the increase of drain current and gate current both. Increase of drain current is positive for the device but there should not be increase in the gate current. Variation of drain and gate current with temperature is shown.

References

[1] Zhao, J.H., et al.: 3.6 mΩ-cm^2, 1726 V 4H-SiC normally-off trenched-and-implanted vertical JFETs and circuit applications. IEE Proceedings – Circuits, Devices and Systems 151(3), 231–237 (2004)

[2] Zhao, J.H., Li, X., Tone, K., Alexandrov, P., Pan, M., Weiner, M.: Design of a novel planar normally-off power VJFET in 4H-SiC. Solid-State Electronics 47(2), 377–384 (2003)

[3] Khalid, M., Riaz, S., Naseem, S.: Modeling and simulations of 1 KV/5 A normally-off 4H-SiC VJFET. In: 2011 IEEE 14th International Multitopic Conference (INMIC), December 22-24, pp. 233–237 (2011)

[4] Abuishmais, I., Undeland, T.: Dynamic characterization of 63 mΩ, 1.2 kV, normally-off SiC VJFET. In: 2011 IEEE 8th International Conference on Power Electronics and ECCE Asia (ICPE & ECCE), May 30-June 3, pp. 1206–1210 (2011)

[5] Shui, Q., Gu, X., Myles, C.W., Mazzola, M.S., Gundersen, M.A.: Simulations of a high power 4H-SiC VJFET and its GaAs counterpart. In: 14th IEEE International Pulsed Power Conference, Digest of Technical Papers, PPC 2003, June 15-18, vol. 1, pp. 123–126 (2003)

[6] Zhao, J.H., Li, X., Tone, K., Alexandrov, P., Pan, M., Weiner, M.: Design and fabrication of a novel power VJFET in 4H-SiC. In: 2001 International Semiconductor Device Research Symposium, pp. 564–567 (2001)

[7] Sannuti, P., Li, X., Yan, F., Sheng, K., Zhao, J.H.: Channel electron mobility in 4H–SiC lateral junction field effect transistors. Solid-State Electronics 49(12), 1900–1904 (2005)

[8] Bhatnagar, M., Baliga, B.J.: Analysis of silicon carbide power device performance. In: Proc. 3rd ISPSD, April 22-24, pp. 176–180 (1991)

[9] Malhan, R.K., Takeuchi, Y., Kataoka, M., Mihaila, A.-P., Rashid, S.J., Udrea, F., Amaratunga, G.A.J.: Normally-off trench JFET technologyin 4H silicon carbide. Microelectron. Eng. 83(1), 107–111 (2006)

[10] Veliadis, V., Hearne, H., Stewart, E.J., Snook, M., McNutt, T., Howell, R., Lelis, A., Scozzie, C.: Investigation of the suitability of 1200-V normally-off recessed-implanted-gate SiC VJFETs for efficient power switching applications. IEEE Electron Device Lett. 30(7), 736–738 (2009)

[11] Jayant Baliga, B.: Fundamentals of Power Semiconductor Devices, pp. 25–100. Springer, New York (2008)

[12] Sankin, I., Sheridan, D.C., Draper, W., Bondarenko, V., Kelley, R., Mazzola, M.S., Casady, J.B.: Normally-Off SiC VJFETs for 800 V and 1200 V Power Switching Applications. In: 20th International Symposium on Power Semiconductor Devices and IC's, ISPSD 2008, May 18-22, pp. 260–262 (2008)

Change Detection from Satellite Images Using PNN

Akansha Mehrotra, Krishna Kant Singh, Kirat Pal, and M.J. Nigam

Indian Institute of Technology, Roorkee, India
{akanshasing,krishnaiitr2011}@gmail.com

Abstract. This paper presents a supervised change detection technique for satellite images using a probabilistic neural network (PNN). The proposed method works in two phases. In the first phase a difference image is computed. The most commonly used techniques for computing the difference image such as ratio images or log ratio images degrade the performance of the algorithm in the presence of speckle noise. To overcome the above mentioned limitations the difference image in this work is computed using normalized neighborhood ratio based method. In the next phase the PNN is used to detect efficiently any change between the two images. An estimator is used by the PNN to estimate the probability density function. The ratio of two conditional probability density functions, called the likelihood ratio is computed. Finally, the log likelihood ratio test is used to classify the pixels of the difference image into changed and unchanged classes to create a change map. The change map highlights the changes that have occurred between the two input images. The proposed method was compared quantatively as well as qualitatively with other existing state of the art methods. The results showed that the proposed method outperforms the other methods.

Keywords: Change detection, Probability density function, Probabilistic neural network (PNN).

1 Introduction

Satellite images are widely used for monitoring urban expansion and land use/cover changes at a medium or large scale, to help better observe and understand the evolution of urbanization and advance the sustainable development process. In all these application change detection methods are widely used. Change detection aims at identifying the changes in spatial representation of any point by observing it at different times [1]. The change detection methods can be broadly classified into the following groups based on the technology they use: algebra, transformation, classification and other approaches. The simplest types of methods are based on algebra like image rationing, differencing, background subtraction, change vector analysis and vegetation index differencing. The drawback with these methods is that they do not provide complete change information. Classification based methods identify the features occurring in an image in terms of the object or type of land cover these features actually represent on the ground and assigns them a unique gray level (or color).Normally, multispectral data are used to perform the classification and,

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 333–340, 2013.

indeed, the spectral pattern present within the data for each pixel is used as the numerical basis for categorization [2].The change detection methods based on unsupervised classification perform multi date image classification and then the classified images comparison is done to discover changes that occur between the two images. A framework for the detection of multiple changes in bi-temporal and multispectral remote sensing images is proposed in [3]. They used all available spectral changes to build up a feature space and developed a compressed 2-D representation of change information which could be easily understood and visualized in polar coordinate system. To identify actual changes an automatic two step method was proposed. To separate changed and unchanged pixel the EM algorithm was used. In [4] a binary semisupervised support vector machine (S^3VM) classifier is proposed. The classifier takes as input multitemporal images. It uses CVA technique and Bayesian thresholding for deriving an initial set of seed pixels having a high probability to be correctly assigned to the classes of changed and unchanged pixels. Using these training sets the classifier develops an unsupervised map which is used to obtain the final change detection map using similarity measures.[5] presents a method based on Gaussian mixture model (GMM) of the difference image and Bayes theory to perform change detection. The difference image is first modeled using GMM .The components of the GMM are then classified into changed and unchanged pixels using Bayes theory. Another method is given in [6] the neighborhood data around each pixel form a sample and are modified by the so-called local gradual descent matrix (LGDM), values of which are descending from center toward outside. Expectation maximization (EM)-based approach [7] analyses the difference image and through automatic selection of the decision threshold it minimizes the overall change detection error. Another technique based on Markov-random-field (MRF), uses the spatial contextual information included in the neighborhood of each pixel to analyze the difference image [7].The PNN has proved to be quite efficient as it has a fast training process, an inherent parallel structure, and guaranteed optimal classification performance [8].

This paper presents a change detection method for identifying change between bi-temporal satellite images based on the Probabilistic neural network (PNN).A difference image is obtained by using neighborhood ratio based method. This difference image is used by the PNN to create a change map showing changes between the two images. The network architecture is made up of four types of units: input units, pattern units, summation units and output unit. The network has an estimator for probability density function (pdf). The pdfs are used to compute likelihood ratio. Finally, log likelihood ratio test is done to assign the pixels of the difference image into changed and unchanged classes to create a change map.

2 Network Architecture

The architecture of the network is shown in fig.1.The architecture is made up of four types of units.

1. Input units: The input units are the pattern vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ …$\boldsymbol{x}_n$ where $\boldsymbol{x}_k$ is the pattern vector of the k_{th} pixel in the image.

2. Pattern units: The pattern units are mean values m_1, m_2 and covariance matrices C_1, C_2 of class ω_c and ω_{uc} respectively. The mean value and covariance matrix of each class are computed using training vectors.
3. Summation Units: The summation units are the estimator function of both the classes. The estimator function estimates the probability density function.
4. Output Unit: The output unit is the final change map (CM).

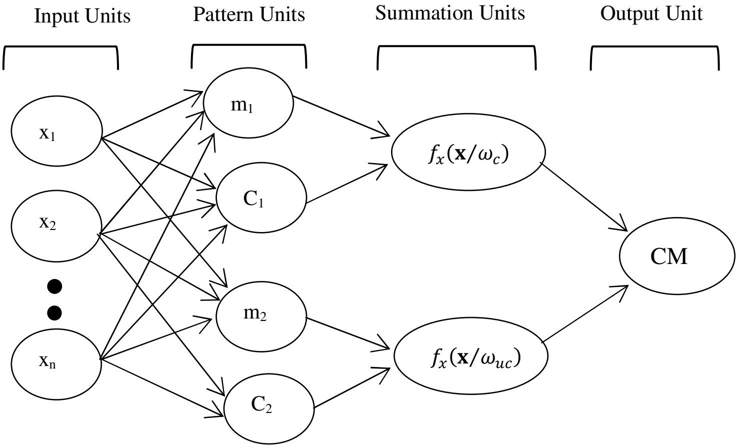


Fig. 1. Architecture of probabilistic neural network for change detection

3 Proposed Method

The change detection method proposed here takes as input two images taken at two different times over the same geographical area. The proposed method works in two phases in the first phase difference image is created and in the next phase the difference image is fed as input to the PNN to obtain change detection map. The overview of the proposed method is shown in fig.2.Input pattern vectors are created for each pixel location from the difference image I_d. The weight vectors w_c and w_{uc} are obtained from the training patterns of both the classes.

Phase I: Creation of Difference Image

Let us consider two satellite images I_1 and I_2 of size M × N taken at time t_1 and t_2 which are co-registered with respect to each other. In general, the change detection algorithms work on the difference image computed from the abovementioned satellite images. The most commonly used techniques for computing the difference image are ratio method [9] and the Log ratio method [10], but these method suffer from a serious limitation. These methods are sensitive to the presence of speckle noise and thus reduce the performance of the algorithm to a great extent. Thus, in this paper a

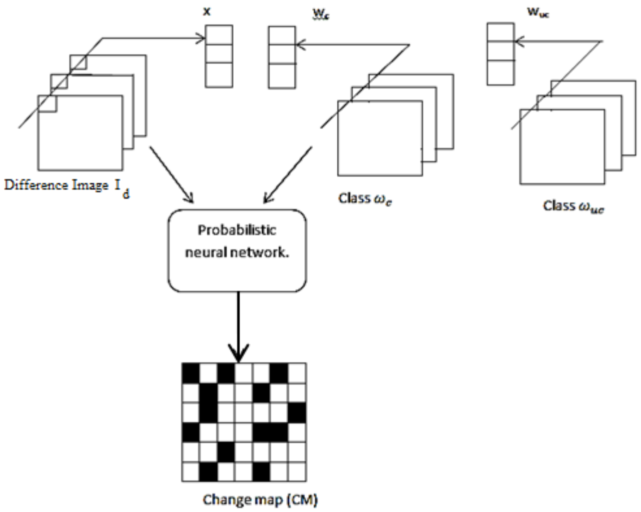


Fig. 2. Pictorial representation of the proposed method

normalized neighborhood ratio approach is used to generate the difference image which is an improved form of neighborhood ratio approach[11]. The proposed method of finding the difference image overcomes the effect of speckle noise. The difference image can be computed using equation 1,

$$I_d(x) = \delta \frac{\max\{I_1(x),I_2(x)\}-\min\{I_1(x),I_2(x)\}}{\max\{I_1(x),I_2(x)\}+\min\{I_1(x),I_2(x)\}} + (1-\delta)\frac{\sum_{i\,\epsilon\,N}\max\{I_1(i),I_2(i)\}-\min\{I_1(i),I_2(i)\}}{\sum_{i\,\epsilon\,N}\max\{I_1(i),I_2(i)\}+\min\{I_1(i),I_2(i)\}} \quad (1)$$

where $\delta = \frac{\sigma_N}{\mu_N}$.

The first term computes the normalized ratio of the pixel under consideration and the second term computes the sum of the normalized ratio of the pixels in the $w \times w$ neighborhood N of the pixel under consideration.σ_N and μ_N are the variance and mean of the gray level in the neighborhood respectively.

Phase II: Creation of Change Map Using PNN

A probabilistic neural network for change detection is presented here to classify the difference image (I_d) into two classes changed (ω_c) and unchanged (ω_{uc}). The PNN shown in fig.1 is used. pattern vector $\boldsymbol{x}$ is a three dimensional vector where each component represents the grey level value of a pixel of the difference image (I_d) in the three bands (RGB).Thus, $\boldsymbol{x}$ can be defined as

$$\boldsymbol{x} = [I_{dR} \quad I_{dG} \quad I_{dB}]^T \quad (2)$$

where I_{dR} ,I_{dG} *and* I_{dB} represent the red , green and blue band of the difference image (I_d). Thus, for the difference image with n (where n = M×N) pixels there will be $x_1, x_2, \ldots\ldots x_n$ input pattern vectors one for each pixel location. For change detection problem the number of classes is two. Each class has a mean value $m_k$and covariance matrix C_k that depends on whether it belongs to class ω_c or class ω_{uc}.The following estimator is used by the probabilistic neural net to estimate the probability density function.

$$f_x(\boldsymbol{x}/\omega_k) = \frac{1}{(2\pi)^{n/2}\,|C_k|^{1/2}}\, e^{-\frac{1}{2}[(x-m_k)^T C_k^{-1}(x-m_k)]} \tag{3}$$

Where $f_x(\boldsymbol{x}/\omega_k)$ is the probability density function of the pattern vectors $\boldsymbol{x}$ of class ω_k , n is the dimensionality of the pattern vector $\boldsymbol{x}$. $C_k\ and\ m_k$ are the covariance matrix and mean vector of the pattern population of class ω_k and $|C_k|$ is the determinant of C_k.The PNN creates a change map by applying the following rule,

If the condition

$$P_c f_x(\boldsymbol{x}/\omega_c) > P_{uc} f_x(\boldsymbol{x}/\omega_{uc}) \tag{4}$$

holds, assign the input vector to class ω_c.Otherwise assign x to ω_{uc}.

P_c and P_{uc} are the probabilities of occurence of patterns in classes ω_c and ω_{uc} respectively. To further simplify matters, the ratio of two conditional probabilities called the likelihood ratio is used. The likelihood ratio ($\Lambda(x)$) is defined as,

$$\Lambda(x) = \frac{f_x(\boldsymbol{x}/\omega_c)}{f_x(\boldsymbol{x}/\omega_{uc})} \tag{5}$$

From computational point of view , it is more convinient to work with the log of the likelihood ratio.Therefore in equivalent form eq.(5) , reduces to a linear classifier , as described by the relation

$$y = w^T\boldsymbol{x} + b \tag{6}$$

where

$$y = \log \Lambda(x) \tag{7}$$

$$w = C_1^{-1}m_1 - C_2^{-1}m_2 \tag{8}$$

$$b = \frac{1}{2}(m_2^T C_2^{-1} m_2 - m_1^T C_1^{-1} m_1) \tag{9}$$

On the basis of equation (6), the log likelihood-ratio test for the change detection problem as follows,

If the output y of the linear combiner (including the bias b) is positive, assign the input vector $\boldsymbol{x}$ to class ω_c.Otherwise assign it to class ω_{uc}.

Thus,

$$CM(k) = \begin{cases} \omega_c & if\ y_k > 0 \\ \omega_{uc} & otherwise \end{cases} \tag{10}$$

where y_k is the output for input pattern vector x_k.

4 Experimental Results

In order to study the performance of the proposed method, the data set used is taken from [12], which contains a large set of ASAR images. Two images one acquired on 26 July 2007 and another on 12 April 2007 that highlights the flooding in Bangladesh and parts of India brought on by two weeks of persistent rain were selected. A small area with 131×133 pixels as shown in fig.3(a) and (b) is used for simulation in this paper. The ground truth of the change detection mask shown in fig.3(c) is created by manual analysis. The proposed method was implemented in Matlab7. The qualitative as well as quantitative comparison of the proposed method was done with EM based approach [7], MRF based method [7] and LGDM based method [6].The qualitative results are shown in fig.3 For quantitative comparison the change detection mask obtained from each method and the ground truth image is used. The following parameters are used:

TP (True Positive), the number of changed pixels correctly identified.
TN (True Negative), the number of pixels correctly identified as unchanged.
FN (False Negative), the number of changed pixels wrongly identified as unchanged pixels;
FP (False Positive), the number of unchanged pixels identified as changed pixels;

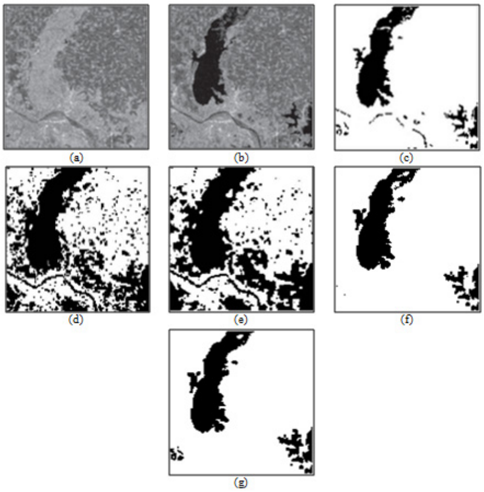


Fig. 3. (a) Small region of ESA/Envisat ASAR image acquired on April 12, 2007. (b) Small region of ESA/Envisat ASAR image acquired on July 26, 2007. (c) Ground truth of the change detection mask (d) EM-based approach (e) MRF-based approach (f) LGDM method with h = 4, $\psi = 3.5$ (g)Proposed Method(PM).

Several metrics can be derived from the above quantities to assess the performance of an method [13]. In this paper, the following metrics are adopted:

1) Omission error (OE), indicates the probability that a changed pixel is wrongly identified as unchanged pixels; OE = FN/(FN + TP)
2) Commission error (CE), indicates the probability that a unchanged pixel is wrongly identified as a unchanged pixel; CE = FP/(TN + FP)
3) Overall Accuracy (OA) is an indication of the overall accuracy of the method in identifying changed pixels as changed and unchanged pixels as unchanged; OA = (TP+TN)/(TP+TN+ FP+FN)

The proposed method achieves 0.14% omission error ensuring that the misclassification are low in the proposed method as compared to the other methods[6,7].The overall accuracy of the proposed method is 98.07 which is quite satisfactory. The other methods show high degradation in performance as compared to the proposed method. It is observed from fig.3(d) and (e) that the results obtained from EM and MRF based approaches are quite noisy. The reason for this is that the difference image is modeled incorrectly.

Table 1. Result of various Parameters used for quantitative comparison

Method	TP	TN	FP	FN	OE(%)	CE(%)	OA(%)
EM	8304	3159	63	5897	41.52	1.95	65.79
MRF	9069	3174	48	5132	36.14	1.49	70.27
LGDM	14165	2760	462	36	0.25	14.33	97.14
PM	14180	2908	314	21	0.14	9.74	98.07

5 Conclusion

In this paper a change detection method based on PNN is presented. The PNN based method has a fast training process and guaranteed optimal classification performance.

The proposed method first produces a difference image by finding the normalized ratio of the gray level information of neighborhood pixels and thus, reduces the negative effect of speckle noise. This difference image is then fed as input to the PNN. The estimator function of the PNN finds the Probability density functions. The pdfs are then used to compute likelihood ratio. The pixels of the difference image are then assigned to change and unchanged classes based on the log likelihood ratio test.

The change map shows the changes between the two images taken as input. The method was compared with some of the existing methods and the results show that the proposed method yields satisfactory results with high overall accuracy. Thus, the proposed method can be used to effectively detect changes in satellite images.

References

[1] Green, K., Kempka, D., Lackey, L.: Using remote sensing to detect and monitor land-cover and land-use change. Photogrammetric Engineering and Remote Sensing 60, 331–337 (1994)

[2] Lillesand, T.M., Kiefer, R.W.: Remote Sensing and Photo Interpretation, 3rd edn. John Wiley & Sons, New York (1994)

[3] Bovolo, F., Marchesi, S., Bruzzone, L.: A Framework for Automatic and Unsupervised Detection of Multiple Changes in Multitemporal Images. IEEE Transactions on Geoscience and Remote Sensing 50(6), 2196–2212 (2012)

[4] Bovolo, F., Bruzzone, L., Marconcini, M.: A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure. IEEE Trans. Geosci. Remote Sens. 46(7), 2070–2082 (2008)

[5] Celik, T.: Method for unsupervised change detection in satellite images. Electron. Lett. 46(9), 624–626 (2010)

[6] Yetgin, Z.: Unsupervised Change Detection of Satellite Images Using Local Gradual Descent. IEEE Transactions On Geoscience And Remote Sensing 50(5), 1919–1929 (2012)

[7] Bruzzone, L., Prieto, D.: Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. 38(3), 1171–1182 (2000)

[8] Goh, A.T.: Probabilistic neural network for evaluating seismic liquefaction potential. Can. Geotech. J. 39(1), 219–232 (2002)

[9] Oliver, C., Quegan, S.: Understanding Synthetic Aperture Radar Images. Artech House, Norwood (1998)

[10] Celik, T.: Change detection in satellite images using a genetic algorithm approach. IEEE Geosci. Remote Sens. Lett. 7(2), 386–390 (2010)

[11] Gong, M., Cao, Y., Wu, Q.: A neighborhood-based ratio approachfor change detection in SAR images. IEEE Geosci. Remote Sens. Lett. 9(2), 307–311 (2012)

[12] European Space agency, http://www.esa.int

[13] Congalton, R.G.: A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35-46 (1991)

Optimal Location and Size of Different Type of Distributed Generation with Voltage Step Constraint and Mixed Load Models

Rajendra P. Payasi[1], Asheesh K. Singh[1], and Devender Singh[2]

[1] Department of Electrical Engineering, MNNIT, Allahabad (U.P.), India, 211004
[2] Department of Electrical Engineering, IIT_BHU, Varanasi (U.P.), India, 221005
payasirp@rediffmail.com, asheesh_k_singh.com,
dsingh.eee@itbhu.ac.in

Abstract. The optimal location and size of distributed generation in distribution network are essentially affected by type of DG, constraints, and loading condition. The type of distributed generation (DG) categorized on the basis of their terminal characteristic in terms of active and reactive power delivering capability have been considered for study. The voltage step change that occurs on sudden disconnection of DG is one of constraints to limit the size of DG more than the voltage level constraint. The loads connected to network are normally voltage dependent and varies with seasonal atmospheric conditions. The voltage dependency and seasonal variation of load necessitate to represent the load by load models for analysis. In this paper, the study has been carried out for distributed generation planning (DGP) for different type of DG in 38-bus test distribution system with voltage step constraint including normally considered constraints i.e. bus voltage constraint, line power capacity constraint, and seasonal mixed load models. The analysis shows that optimal location and size are significantly affected by type of DG, voltage step constraint, and load models.

Keywords: Distributed generation, distribution system, distributed generation planning, load models.

1 Introduction

Distributed generation, also termed as embedded generation or dispersed generation or decentralized generation, is defined as small electric power generation units connected directly to the distribution network or connected to the network on the customer site of the meter [1]. The limitations of traditional generation and on the other hand immense technical, economical and environmental benefits of DG as well as technological development in DG have renewed the interest on DG [2]-[4]. The loss minimization in distribution network is one of the vital requirement to operate the system economically which could be achieved by proper distributed generation planning (DGP) i.e. by placement of distributed generations (DGs) at optimum

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 341–357, 2013.

locations with optimum size and suitable type corresponding to minimum power loss under certain constraints.

In practice, the loads in distribution network are voltage dependent and the major variations could be observed according to variations in seasonal weather conditions such as summer day, summer night, winter day, and winter night. The load at each bus may also be the composition of different kinds of voltage dependent loads such as industrial, residential, and commercial. The different types of voltage dependent loads are represented by basic load models as described in [5] and the concept of mixed load model along with seasonal variations in loads is adopted for DGP in [6]. The authors defined four major types of DGs but considered only one type of DG, which is capable of delivering both real and reactive power, for study in [7]. In [8], authors studied the effect of voltage step constraint on size of DGs connected at particular location for three operating power factors (0.95 lagging, unity, and 0.95 lagging). Thus mixed load model, types of DGs, and voltage step constraint are the important factors to be considered together for proper DGP.

In [9]-[11], different kinds of basic voltage dependent load models have been considered and compared with constant power load model. In [12]-[14], voltage independent variable loads have been adopted for DG placement. The DGP problem was also solved by adopting DG which can supply both real and reactive power but without considering load models and voltage step constraint in [15]-[21]. From literature survey [22], very few works have been found on implementation of voltage step constraint, mixed load model, and all major types of DGs in DGP. In [24], authors considered the different type of DG and mixed load models for DGP but not considered voltage step constraint.

In this paper, 38-bus system is adopted as described in [9], [10] (Fig. 7 in Appendix). The investigation regarding DGP analysis is performed considering summer day mixed, summer night mixed, winter day mixed, winter night mixed load models which include industrial, residential, and commercial load models at every bus in certain proportion (assumed as in Table II). The investigation also considered the voltage step limit of 3% along with bus voltage limits and line power capacity limit as constraints. The analysis has been performed corresponding to minimum P_L for different types of DGs under voltage step constraint.

The paper is structured as follows: Section II describes the types of DGs. In section III, phenomenon of voltage step has been illustrated with the help of a simple two-bus system. Section IV describes the load models and test cases considered for investigation. Section V describes the methodology adopted. Section VI presents the simulation results and analyses of studies. The last section VII presents the conclusion of the paper.

2 Type of Distributed Generation

There are different types of traditional and nontraditional DGs classified and described in [3] from the constructional, technological, size, and power time duration pint of view. The DGs have also been classified into four major types, based on terminal characteristics in terms of real and reactive power delivering capability, as described in [7]. In this paper, the four major types have been considered for comparative studies which are discussed as follows:

Type1. This type of DG is capable of delivering only real power. Photovoltaic, micro-turbines, and fuel cells, which are integrated to the main grid with the help of converters/inverters, are the examples of Type1. The converters/inverters-connected Type1 DG can control both real and reactive power outputs up to certain extent and may be categorized as Type2.

Type2. This type of DG is capable of delivering both real and reactive power. DG units based on synchronous machines (cogeneration, gas turbine, etc.) come under this type. In the present work the generation limits of the synchronous generators have not been considered explicitly. However, this is considered by constraining operating power factor in the range of 0.8 ld to unity.

Type3. This type of DG is capable of delivering only reactive power. Gas turbines in synchronous compensator mode and other sources of reactive power are the examples of this type.

Type4. This type of DG is capable to deliver real power but consumes reactive power. Mainly induction generators, which are used in wind farms, come under this category. The doubly fed induction generator (DFIG) systems may produce reactive power similar to synchronous generator and hence DFIG may be considered as Type2 DG.

In this paper, DGs adopted for studies are the basic DGs based on their terminal characteristic in terms of real and reactive power delivering capability.

3 Voltage Rise and Voltage Step

The voltage rise is the increase in voltage with inclusion of DG, and voltage step change is instantaneous drop in voltage with loss of DG. The phenomenon of voltage step as explained by Dent *et al* [8], is different from voltage rise and is illustrated in this section with the help of Fig. 1. Bus 2, as depicted in Fig. 1, has a load as (P_{D2}+jQ_{D2}) and DG size as (P_{DG2}+jQ_{DG2}).The power flowing from bus-2 to bus-1 through line of impedance (R+jX) , when DG is connected at bus-2, would cause steady state voltage rise at bus 2 ($V_{rise\,2}$) as given below (assuming that the voltage at A remains constant as 1 p.u.) [8].

$$V_{rise\,2} = (P_{DG2} - P_{D2})R + (Q_{DG2} - Q_{D2})X \tag{1}$$

On subtracting the voltage at B without DG ($V_{WODG\,2}$) from the voltage at bus-2 with DG ($V_{WDG\,2}$), the voltage step at bus B ($V_{step\,2}$) on loss of the DG is given as follows.

$$V_{step\,2} = V_{WDG\,2} - V_{WODG\,2} = -(P_{DG2}R + Q_{DG2}X) \tag{2}$$

The voltage step limit is taken on the basis of full output of the DG. The voltage step limit is expected to restrict the DG size more than the normal voltage limit constraints. As per the UK standards V_{step} limit is specified as 3% for planned switching outages and 6% for unplanned outages whereas 5% is common in use in USA [8].

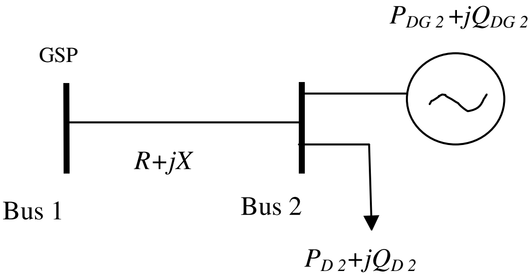


Fig. 1. Two-bus system for voltage step analysis

In this paper, study has been carried out for planned outages and V_{step} in p.u. has been calculated at i^{th} bus each step size of DG as follows:

$$V_{step\,i} = V_{WDG\,i} - V_{WODG\,i} \quad for\ i = 2\ to\ N_B \tag{3}$$

4 Load Models and Test Cases

To quantify the effect of different type of DGs, seasonal mixed load models (*SDM, SNM, WDM, WNM*), and voltage step constraints on DGP, a 38-bus distribution system [9],[10] is adopted. The line impedances, load data (balanced) and the line power limits, expressed in p.u. at the base voltage of 12.66 kV and base MVA of 1.0 MVA, are adopted [9], [10], and [24]. In conventional load flow analysis, the real and reactive power loads are assumed as constant i.e. not dependent on voltage or frequency. While in fact, the distribution loads are voltage dependent and practically these are industrial, residential, and commercial. A voltage dependent load model is a static load model that represents the power relationship to voltage as an exponential equation, which can be expressed in following form [5].

$$P_i = P_{0i}\left(|V_i| / |V_{0i}|\right)^{\alpha} \tag{4}$$

$$Q_i = Q_{0i}\left(|V_i| / |V_{0i}|\right)^{\beta} \tag{5}$$

Above equations (4) and (5) neglect the frequency dependence of distribution load, due to the fact that the frequency variation is relatively in narrow range. In practice, the load on each bus is composition of industrial, residential, and commercial which varies according to seasonal day, and night. Therefore, in this paper the load model at each bus is represented by following equations.

$$P_i = W_{1pi}\,.\,P_{0i}\left(|V_i| / |V_{0i}|\right)^{\alpha_i} + W_{2Pi}\,.\,P_{0i}\left(|V_i| / |V_{0i}|\right)^{\alpha_r} + W_{3Pi}\,.\,P_{0i}\left(|V_i| / |V_{0i}|\right)^{\alpha_c} \tag{6}$$

$$Q_i = W_{1Qi}\,.\,Q_{0i}\left(|V_i| / |V_{0i}|\right)^{\beta_i} + W_{2Qi}\,.\,Q_{0i}\left(|V_i| / |V_{0i}|\right)^{\beta_r} + W_{3Qi}\,.\,Q_{0i}\left(|V_i| / |V_{0i}|\right)^{\beta_c} \tag{7}$$

where, α_i and β_i are for industrial load model; α_r and β_r for residential load model; α_c and β_c for commercial load model. The values of α's and β's are zeros for constant power load model.

W_{1Pi} & W_{1Qi}, W_{2Pi} & W_{2Qi}, and W_{3Pi} & W_{3Qi} are the composition weights for real & reactive powers of industrial, residential and commercial loads respectively at bus i, except for unloaded buses (UB). The composition factors are assumed such that

$$W_{1Pi} + W_{2Pi} + W_{3Pi} = 1 \quad for\ i = 1\ to\ N_B\ ,\ i \neq UB \tag{8}$$

$$W_{1Qi} + W_{2Qi} + W_{3Qi} = 1 \quad for\ i = 1\ to\ N_B\ ,\ i \neq UB\ . \tag{9}$$

The values for exponents of voltage for real and reactive component of summer day, summer night, winter day, and winter night loads are given in Table 1 [6]. The assumed composition weights of each load model at each bus is as shown in Table 2. In this study it is assumed that $W_{1Pi}=W_{1Qi}$, $W_{2Pi}=W_{2Qi}$, and $W_{3Pi}=W_{3Qi}$.

The study is performed considering practical situations of load as follows: 1) each bus having mix of industrial, residential, and commercial load in certain proportion; 2) Load vary with seasonal day and night. Apart from these situations, *T1,T2,T3* and *T4* have been considered for comparative study. A 38-bus system is assumed to be supplying power to mixed of industrial, residential, and commercial load without violating voltage step constraint as well as usually adopted constraints i.e. bus voltage limits and line capacity limit. The following test cases are considered for optimal size and location for constant and seasonal mixed load models considering P_L minimization as objective function.

1) Type 1 DG with and without *VSL* constraint.
2) Type 2 DG with and without *VSL* constraint
3) Type 3 DG with and without *VSL* constraint
4) Type 4 DG with and without *VSL* constraint

Table 1. Typical load types and exponent values [6]

Load type		Exponent values					
		Industrial		Residential		Commercial	
		α_i	β_i	α_r	β_r	α_c	β_c
Summer	Day	0.18	6.00	0.72	2.96	1.25	3.50
	Night	0.18	6.00	0.92	4.04	0.99	3.95
Winter	Day	0.18	6.00	1.04	4.19	1.50	3.15
	Night	0.18	6.00	1.30	4.38	1.51	3.40

Table 2. Value of Relevant Factors of load Models at Each bus

Bus no	W_{1Pi}=W_{1Qi}	W_{2Pi}=W_{2Qi}	W_{3Pi}=W_{3Qi}
2	0.2000	0.6000	0.2000
3	0.1500	0.6500	0.2000
4	0.2000	0.5000	0.3000
5	0.1100	0.3400	0.5500
6	0.1000	0.3500	0.5500
7	0.3000	0.5000	0.2000
8	0.3000	0.5000	0.2000
9	0.0800	0.2000	0.7200
10	0.0800	0.2000	0.7200
11	0.1200	0.2000	0.6800
12	0.2500	0.3000	0.4500
13	0.2500	0.3500	0.4000
14	0.2000	0.3000	0.5000
15	0.0500	0.3000	0.6500
16	0.0800	0.2000	0.7200
17	0.0800	0.2000	0.7200
18	0.3000	0.4000	0.3000
19	0.3000	0.4000	0.3000
20	0.3000	0.4000	0.3000
21	0.3000	0.4000	0.3000
22	0.3000	0.4000	0.3000
23	0.3500	0.4500	0.2000
24	0.2000	0.6500	0.1500
25	0.2000	0.6500	0.1500
26	0.1000	0.2500	0.6500
27	0.1000	0.2500	0.6500
28	0.1000	0.3000	0.600
29	0.2500	0.3500	0.4000
30	0.5000	0.3000	0.2000
31	0.2500	0.3500	0.4000
32	0.3000	0.5000	0.2000
33	0.2500	0.3000	0.4500

5 Proposed Methodology

The methodology adopted for DG analysis uses incremental power flow and exhaustive search method to obtain the optimal location and size of DG for real power loss (P_L) minimization. The details of problem formulation, indices calculations, computational procedure for the purpose of database generation, and analysis of results, are given in the following sections. Further, the implementation procedure is also discussed.

5.1 Formulation

The optimal location and size of DG are determined by minimization of real power loss in distribution system with operating constraint of the system. The total real power loss is expressed as follows [11],[24].

$$P_L = \sum_{i,j \in N_L} \frac{P_{ij}^2 + Q_{ij}^2}{|V_i|^2} r_{ij} \tag{10}$$

Loss is function of all system bus voltage (V_i), line resistances ($r_{i,j}$), *α,* and *β*. The total losses mainly depend on voltage profile.

The apparent power intake (S_{intake}) at main substation is expressed as:

$$S_{intake} = [(P_{intake})^2 + (Q_{intake})^2]^{1/2} \tag{11}$$

where, P_{intake} at the main substation is represented as:

$$P_{intake} = P_1(V, P_0, Q_0, \alpha, \beta) = \sum_{i=1}^{N_B} P_{0i}(|V_i|/|V_{0i}|)^{\alpha} + P_L \tag{12}$$

and Q_{intake} at the main substation is represented as:

$$Q_{intake} = Q_1(V, P_0, Q_0, \alpha, \beta) = \sum_{i=1}^{N_B} Q_{0i}(|V_i|/|V_{0i}|)^{\beta} + Q_L \tag{13}$$

And total system power requirement is expressed as:

$$S_{\text{sys}} = [(P_{\text{intake}} + P_{DG})^2 + (Q_{\text{intake}} + Q_{DG})^2]^{1/2} \tag{14}$$

where, $Q_{DG} = 0.0$ for T1; $P_{DG} = 0.0$ for *T3* Q_{DG} = - ve for *T4*

It is observed that for a distribution system

$$\sum_{i=1}^{N_B} P_0(|V_i|/|V_{0i}|)^{\alpha} > P_L \tag{15}$$

$$\sum_{i=1}^{N_B} Q_0(|V_i|/|V_{0i}|)^{\beta} > Q_L \tag{16}$$

Thus the P_{intake} and $Q_{intake,}$ expressed as (12) and (13) respectively are largely decided by bus voltages (V_i) and load exponents (*α's* and *β's),* not by P_L and Q_L.

The above objectives are subject to the following set of power flow, line capacity limit, voltage limits and voltage step limit

$$P_i = \sum_{j=1}^{N_B} |V_i||V_j|[G_{ij}\cos(\delta_i - \delta_j) + B_{ij}\sin(\delta_i - \delta_j)], \ for\, i = 1\, to\, N_B \tag{17}$$

$$Q_i = \sum_{j=1}^{N_B} |V_i||V_j|[G_{ij}\sin(\delta_i - \delta_j) - B_{ij}\cos(\delta_i - \delta_j)], \ for\, i = 1\, to\, N_B \tag{18}$$

$$P_{i,j} = |V_i|^2 G_{ij} - |V_i||V_j|[G_{ij}\cos\theta_{ij} - B_{ij}\sin\theta_{ij}], \ for\ i, j \in N_L \tag{19}$$

$$Q_{i,j} = -|V_i|^2 B_{ij} - |V_i||V_j|[G_{ij} \sin\theta_{ij} + B_{ij} \cos\theta_{ij}], \quad for\ i, j \in N_L \tag{20}$$

$$V_{\min} \le |V_i| \le V_{\max}\ , \quad for\ i = 1\ to\ N_B \tag{21}$$

$$S_{i,j} \le CS_{i,j}^{\max}\ , \quad for\ i, j \in N_L \tag{22}$$

$$V_{step\,i} \le V_{step\,i}^{\max}\ , \quad for\ i = 1\ to\ N_B \tag{23}$$

$$S_{DG} \le\ S_{intake} \tag{24}$$

In this paper voltage limits and VSL are taken as follows:

$$V_{\min} = 0.95\,\text{p.u.},\ V_{\max} = 1.05\,\text{p.u., and } V_{step\,i}^{\max} = 0.03 \times V_{WDG\,i}$$

5.2 Indices to Quantify the Benefits of DG

To compare the results, the indices are defined as follows [10].

Real Power Loss Index (PLI). The real power loss index is defined as :

$$PLI = \frac{P_{LWDG}}{P_{LWODG}} \times 100 \tag{25}$$

The lower values of this index indicate better benefits in terms of real power loss reduction accrued to DG.

Reactive Power Loss Index (QLI). The reactive power loss index is defined as:

$$QLI = \frac{Q_{LWDG}}{Q_{LWODG}} \times 100 \tag{26}$$

The lower values of this index indicate better benefits in terms of reactive power loss reduction accrued to DG.

Voltage Profile Index (VPI). It is related to the maximum voltage drop between a node and root node among the voltage drops between root node and each node. The lower values of this index indicate better performance of network. The *VPI* can be defined as:

$$VPI = \max\left(\frac{|V_1| - |V_i|}{|V_1|}\right) \times 100, \quad for\ i = 2\ to\ N_B \tag{27}$$

Line Capacity Index (LCI). The power flows may diminish in some sections of the network and released more capacity with the power supplied near to the load. This index provides important information about the level of power flows/currents through the network regarding maximum capacity of distribution lines. Lower values of this index indicate more capacity available. Index values equal to 100 % indicate that line limit constraint is active. This is defined as:

$$LCI = \max\left(\frac{|\mathrm{S_{ij}}|}{|\mathrm{CS_{ij}}|}\right) \times 100,\ for\ ij\ set = 1\ to\ N_L \tag{28}$$

Apparent Power Intake (S_{intake}) Index (SII). The lower value of this index indicates more capacity release of substation. This index is defined as

$$SII = \frac{|S_{intake\ WDG}|}{|S_{intake\ WODG}|} \times 100 \tag{29}$$

5.3 Computational Procedure

For investigation, different types of DGs based on terminal characteristics in terms of power delivering capability has been adopted to perform comparative study for better DGP corresponding to minimum P_L. The data base, using Newton Raphson power flow method, for the 38-node distribution system is obtained without and with DG at different node.

In this paper, single DG placement is studied. The algorithm is based on incremental power flow. In this algorithm the value of $|S_{DG}|$ is incremented in steps (0.005 p.u. in present study) till maximum limit (S_{intake}) is reached. The process is repeated assuming DG location at each bus. The size of DGs are considered in practical range decided as equal to or less than power intake at main substation. The step size of power factor is taken as 0.01. The range of power factor is taken between 0.99 to 0.8 leading for *T2* and 0.99 to 0.8 lagging for *T4*. The relevant quantities such as P_L , Q_L, S_{intake}, P_{intake}, Q_{intake} , S_{sys}, *NVLVB, NLCLVL, NVSLVB, PLI, QLI, VPI, LCI,* V_i, *and* S_{ij} are evaluated using load flow study at every node with different type and size of DGs and power factors, and stored. Then from the data base, the required values corresponding to minimum P_L under voltage step limit (*VSL*) , bus voltage limits, and line capacity as constraints are determined using exhaustive search algorithm (given in Appendix) . The computational procedure is described in the following steps.

Step 1: Read of load data, line data, number of buses, DG power increment (ΔS_{DG}), α and β for all load models, voltage limits, voltage step limit, and DG power factor decrement (ΔPF_{DG}) for *T2* and *T4*.

Step 2: Select one of the load models (mixed and constant load models) by selecting exponent values, α and β.

Step 3: Run power flow program without DG ($P_{DG} = 0$, $Q_{DG} = 0$,) and save the data (P_L, Q_L, S_{intake}, P_{intake}, Q_{intake}, S_{sys}, *NVLVB, NLCLVL,* V_i, *and* $S_{i,j}$).

Step 4: Select one of types of DG.

Step 5: Decrement power factor by ΔPF_{DG} from 0.99 for *T2* and *T4* till 0.8 and skip this step for *T1* and *T3*

Step 6: Select one of the buses.

Step 7: Increment DG value by $\Delta S_{DG.}$

Step 8: Run power flow program and save the data. (*DG_bus*, P_{DG}, Q_{DG}, P_L, Q_L, S_{intake}, P_{intake}, Q_{intake}, S_{sys}, *NVLVB, NLCLVL, NVSLVB, PLI, QLI, VPI, LCI,* V_i, *and* $S_{i,j}$)

Step 9: Go to *step 7* if ($P_{DG} \le P_{intake}$ and $Q_{DG} \le Q_{intake}$).

*Step*10: Go to step 6 to select next bus , make DG value in previous bus zero, till all the buses considered.

*Step*11: Go to step 5 for type 2 and type 4 DG till $\mathrm{PF}_{\mathrm{DG}}$ is 0.8 and skip this step for type 1 and type 3 DG.

*Step*12: Go to step 4 to select other type of DG.

*Step*13: Go to step 2 till all the mixed load models are selected.

Step 14: The database obtained in terms of DG_bus, P_{DG}, Q_{DG}, P_L , Q_L, S_{intake}, P_{intake}, Q_{intake} , S_{sys}, *NVLVB, NLCLVB, NVSLVB, PLI, QLI, VPI, LCI,* V_i *, and* S_{ij} used to obtain value of quantities (for zero value of *NVLVB*, *NLCLVL*, and *NVSLVB*) corresponding to minimum P_L as listed in Table 3 and 4 using exhaustive search algorithm (given in Appendix).

5.4 Implementation Strategy

The placement will depend on many other factors such as availability of space and practical suitability. If all the locations are available, then one location corresponding to lowest aggregate energy loss for all the four seasonal load conditions may be implemented.

The energy loss in p.u. is expressed as:

$$Energy\ Loss = w_{SD}.P_{LSD} + w_{SN}.P_{LSN} + w_{WD}.P_{LWD} + w_{WN}.P_{LWN} \tag{30}$$

and

$$w_{SD} + w_{SN} + w_{WD} + w_{WN} = 1 \tag{31}$$

where, w_{SD}, w_{SN}, w_{WD}, w_{WN} are the normalized durations corresponding to summer day mixed (*SDM*), summer night mixed (*SNM*), winter day mixed (*WDM*), and winter night mixed (*WNM*) load conditions respectively. P_{LSD}, P_{LSN}, P_{LWD}, P_{LWN} are real power loss for *SDM, SNM, WDM,* and *WNM* load conditions respectively.

In this paper, the values for normalized weights are assumed as follows.

$$w_{SD}=0.33;\ w_{SN}=0.33;\ w_{WD}=0.17;\ w_{WN}=0.17$$

6 Simulation Results and Discussion

In this section, the summary of simulation results obtained for various test cases is presented. The *NVSLVB* for different load models with different types of DG are summarized in Table 3. The quantities P_{DG}, Q_{DG}, PF_{DG}, *DG-bus,* and S_{sys}, corresponding to minimum P_L with voltage step constraint are presented in Table 4. The indices corresponding to minimum P_L are depicted in Figs. 2 to 6. The values considered for comparison and discussion related to different kind of DGs are under *VSL* constraint. The analysis based on Table 3 and 4 is as follows.

6.1 Effect of Load Models in NVSLVB *(Table III)*

It is observed that for constant power load *NVSLVB* is more severe in case of *T1* and *T2* compared to *T3* and *T4* whereas for voltage dependent load models it is lesser incase of *T1* and *T2*, and zero in case of *T3* and *T4*. The *NVSLVB*s show that in case of load models, the *VSL* constraint is less severe for *T1* and *T2*, and not effective for *T3* and *T4*.

6.2 Load Models

Constant Power Load Model (Cons) (Table 4). The P_L with *T2* is 0.0822 p.u. which is less compare to other types of DGs and without DG. The P_{DG} is 1.368 p.u., at 0.8 leading power factor, which is less than *T1* and *T4*. The location of *T1*, *T2*, and *T4* is at bus 6 whereas bus 30 is for *T3*.

Summer Day Mix Load Model (SDM) (Table 4). The P_L for *T2* is 0.0939 p.u. whereas it is more for other types of DGs and without DG case. The value of P_{DG} for *T2* is 0.72 p.u. at 0.8 leading power factor, which is less than with *T1* and *T4*. The location of different types of DGs is different i.e. the location of *T1* and *T4* is 9, *T2* is 29, and *T3* is 30.

Summer Night Mixed Load Model (SNM) (Table 4). The P_L for *T2* is 0.1051 p.u. whereas it is more for other types of DGs. The P_{DG} of *T2* is 0.528 p.u. ,at 0.80 leading power factor, which is less than other type of DGs. The location of *T1* and *T4* is 12, *T2* is 31, and *T3* is 30 .

Winter Day Mixed Load Model (WDM) (Table 4). The P_L with *T2* is 0.1129 p.u. which is less than without DG (0.1644p.u.) and with other types of DGs cases. The P_{DG} is 0.436 p.u. ,at 0.8 leading power factor, which is less than T1 and T4. The location of *T2* and *T3* in this case is same as in case of *SNM* i.e. bus 31 and 30. The location for *T1* and *T4* is bus 14.

Winter Night Mixed Load Model (WNM) (Table 4). The P_L for *T2* is 0.1229 p.u. which is less than without DG (0.1636 p.u.) , and with other type of DGs cases. P_{DG} of *T2* is 0.3280 p.u. (at PF_{DG} = 0.80 leading) which is less than with other type of DGs cases. The location is different for each type of DG.

Discussion. From Table IV, it is observed that in order to minimize P_L , the P_L is minimum for all seasonal load models only with *T2* compared to other type of DGs. It is seen that depending on the type of DG the optimal locations varies significantly from one type to other. The optimum P_{DG} is minimum with *T2* compared to with *T1* and *T4* for all load models. The operating power factor for *T2* is 0.8 leading. The S_{sys} is lowest with *T1* when voltage dependent load models are considered. However, this was not observed for constant power loads. The S_{sys} is maximum with *T3* for all load models, because this type of DG improves the voltage profile thereby increasing the load. Also higher P_L is observed in case of *T3* for all load models.

In case of *T4* the optimal performance is observed at 0.99lag (closer to unity). This suggests that such DGs are to be used at unity power factor or with power factor constraints if possible. However in certain case it is not possible to run the DG on unity power factor (such as induction generator). Therefore, in simulation they are to be represented as having reactive power (fixed or variable).The leading power factor demand from *T2* remain around 0.8 but does not increase on expense of real power.

Table 3. NVSLVB for Various Load Models without VSL constraint

System	DG	*NVSLVB (out of 38)*				
condition	Type	*Cons*	*SDM*	*SNM*	*WDM*	*WNM*
Minimum P_L	*T1*	25	11	7	0	0
	T2	25	5	3	0	0
	T3	3	0	0	0	0
	T4	9	0	0	0	0

Table 4. DG size and location corresponding to Minimum P_L With VSL Constraint

Load model	*W/WO DG*	DG Type	P_{DG} *(p.u.)*	Q_{DG} *(p.u.)*	PF_{DG}	*DG bus*	S_{sys}	P_L*(Mini.) (p.u.)*
Cons	*WODG*	-	-	-	-	-	4.5963	0.1889
	WDG	*T1*	2.0400	0.0	1.0	6	4.4930	0.1010
		T2	1.3680	1.0260	0.8 ld	6	4.4706	0.0822
		T3	0.0	1.2500	0.0	30	4.5308	0.1342
		T4	2.2572	-0.3216	0.99 lg	6	4.5097	0.1150
SDM	*WODG*	-	-	-	-	-	4.4372	0.1667
	WDG	*T1*	1.0200	0.0	1.0	9	4.4477	0.1104
		T2	0.7200	0.5400	0.8 ld	29	4.4534	0.0939
		T3	0.0	1.1250	0.0	30	4.4678	0.1295
		T4	1.1039	-0.1573	0.99 lg	9	4.4531	0.1166
SNM	*WODG*	-	-	-	-	-	4.4304	0.1654
	WDG	*T1*	0.7300	0.0	1.0	12	4.4380	0.1169
		T2	0.5280	0.3960	0.8ld	31	4.4471	0.1051
		T3	0.0	1.1100	0.0	30	4.4660	0.1296
		T4	07920	-0.1129	0.99 lg	12	4.4415	0.1211
WDM	*WODG*	-	-	-	-	-	4.4224	0.1644
	WDG	*T1*	0.6000	0.0	1.0	14	4.4333	0.1207
		T2	0.4360	0.3270	0.8 ld	31	4.4374	0.1129
		T3	0.0	1.1150	0.0	30	4.4622	0.1286
		T4	0.6534	-0.0931	0.99 lg	14	4.4364	0.1243
WNM	*WODG*	-	-	-	-	-	4.4159	0.1636
	WDG	*T1*	0.4500	0.0	1.0	15	4.4235	0.1273
		T2	0.3280	0.2460	0.8 ld	32	4.4281	0.1229
		T3	0.0	1.1150	0.0	30	4.4620	0.1287
		T4	0.4851	-0.0691	0.99 lg	14	4.4240	0.1296

6.3 Indices for Comparison

The indices *PLI, QLI, VPI*, *LCI*, and *SII* are depicted in Figs. 2 to 6 and discussed as follows.

PLI and QLI (Fig. 2&3): These indicate that loss reduction is less in case of voltage dependent load models compared to constant power load model for all types of DGs. Further, for all load models, the loss reduction is more in case of *T2* compared to other types of DGs.

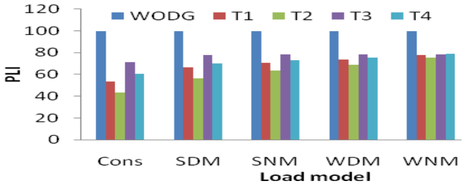


Fig. 2. *PLI* for minimum. P_L

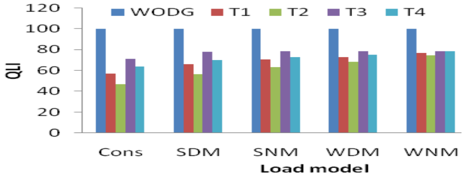


Fig. 3. *QLI* P_L Configuration

VPI (Fig. 4): This index indicates that voltage is improved when DG is connected. It is also observed that in case of constant power load model the improvement is more compared to voltage dependent load models for all types of DGs, which indicate that assumption of constant power load will not depict the real situation.

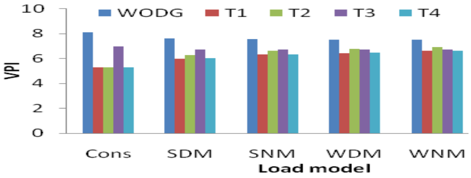


Fig. 4. VPI minimum. P_L

LCI (Fig. 5): This index indicate that assumption of constant load model shows more capacity release whereas assumption of mixed load models show almost nil capacity release for *T1,T2,* and *T4*. Further some capacity release is observed for *SDM*, *SNM*, and *WDM* in case of *T3*.

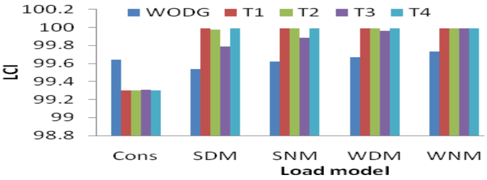


Fig. 5. *LCI* for minimum. P_L

SII (Fig.6): This index indicates that assumption of constant power load model shows more reduction in S_{intake} compared to voltage dependent load model for all type of DGs.

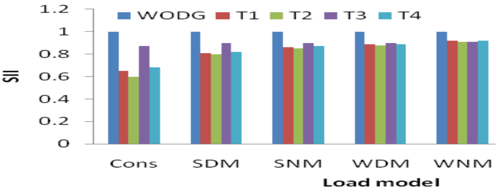


Fig. 6. *SII* for minimum. P_L

Thus, above indices reveal that assumption of constant power load model would show different value of losses, bus voltages, capacity release, and S_{intake} reduction compared to what would result in operational stages when voltage dependency comes into effect. Further, values of indices are also different of different types of DGs.

6.4 Implementation Criterion

The optimum location obtained corresponding to minimum loss are different for different seasonal mixed load models (Table 4). The single location for different types of DGs is obtained on the basis of minimum energy loss using (30) as shown in Table 5 and 6 corresponding to without and with inclusion of *VSL* constraint respectively, and observed that energy loss is minimum for *T2* in both cases (with and without *VSL* constraint) compared to other types of DGs.

Table 5. Optimum Location and Size without *VSL* Constraint

DG Type	DG bus	P_{DG} (p.u.)	Q_{DG} (p.u.)	PF_{DG}	Energy loss (p.u.)
T1	13	0.905	0	1	0.11774
T2	31	0.748	0.561	0.8 ld	**0.10424**
T3	30	0.0	1.125	0	0.12921
T4	12	1.119	-0.159	0.99 lg	0.12245

Table 6. Optimum Location and Size with *VSL* Constraint

DG Type	DG bus	P_{DG} (p.u.)	Q_{DG} (p.u.)	PF_{DG}	Energy loss (p.u.)
T1	12	0.730	0.0	1	0.11895
T2	29	0.720	0.540	0.8ld	**0.10791**
T3	30	0.0	1.125	0	0.12921
T4	12	0.950	-0.135	0.99lg	0.12286

7 Conclusions

The different types of DGs based on their terminal characteristics in terms of power delivering capability, constant power load model as well as mixed load model, and voltage step constraint have been considered for optimum location and size of DG corresponding to minimum real power loss and minimum energy loss in 38 bus distribution system.

The investigations show that in case of Type1and Type2 the numbers of voltage step limit violated buses are reduced drastically for voltage dependent loads compared to the case when constant power load models are considered. In case of Type3 and Type4 DGs, Voltage step constraint is less effective for constant power load model and not effective for the mixed load models.

The power loss , power intake and DG size are less for Type 2DG compared to other types of DGs. The energy loss is also less for Type 2 DG in both cases (without and with *VSL* constraint). The values of indices of constant power load model are significantly different than mixed load models.

Appendix

Search Algorithm for Minimum P_L

```
Step 1 : load the database files.
Step 2 : assign $k=1$, $min\_loss = P_L$ without $DG$ and $k_{max}$ = no. of set of data.
Step 3 : read $P_L(k), NVLVB(k), NLCLVL(k), NVSLVB(k), P_{DG}(k), P_{intake}(k)$.
Step 4 : if $P_L(k) > min\_loss$, go to Step 7.
Step 5 : if $((NVLVB(k) =0)$ && $(NLCLVL(k)=0)$ && $(NVSLVB=0)$ &&
         $(P_{DG}<P_{intake},)$) continue
         else go to Step 7. % ($NVSLVB$ is not considered for without $VSL$).
Step 6 : $min\_loss = P_L(k)$
Step 7 : $k_{minpl}=k$ ,
         %($k_{minpl}$ is assigned the value of $k$ corresponding to minimum $P_L$).
Step 8 : if $k = k_{max}$ , go to Step 10.
Step 9 : $k=k+1$, go to Step 3.
Step10: print $DG\_bus(k_{minpl})$, $P_{DG}(k_{minpl})$ , $Q_{DG}(k_{minpl})$ , $PF_{DG}(k_{minpl})$ , $P_L(k_{minpl})$ ,
         $Q_L(k_{minpl})$, $S_{intake}(k_{minpl})$, $P_{intake}(k_{minpl}),Q_{intake}(k_{minpl}),S_{sys}(k_{minpl})$, $NVLVB(k_{minpl})$,
         $NLCLVL(k_{minpl})$, $NVSLVB(k_{minpl})$, $PLI(k_{minpl})$, $QLI(k_{minl})$, $VPI(k_{minpl})$, $LCI(k_{minpl})$.
Step11: go to Step 1 till all the files are selected.
```

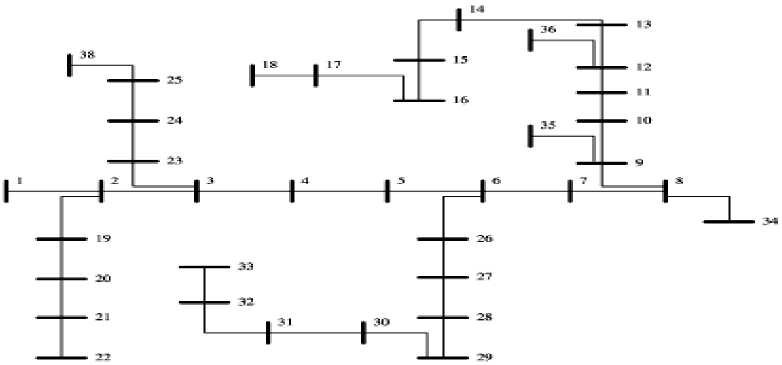


Fig. 7. 38-bus distribution system [9], and [10]

Nomenclature

α, β	Voltage exponent of real and reactive load.
Cons	Constant power load model.
SDM,SNM	Summer day and Summer night mixed load models
WDM,WNM	Winter day and Winter night mixed load models
$CS_{i,j}$	MVA capacity of line *i-j* (p.u.).
S_{intake}	Apparent power (MVA) intake at bus 1 (p.u.).
S_{sys}	Apparent power (MVA) taken by system from all sources (DG and grid) (p.u.).
NLCLVL	Number of line capacity limit violated lines.
NVLVB	Number of voltage limit violated buses.
NVSLVB	Number of voltage step limit violated buses.
P_{0i}, Q_{0i}	Real and reactive load at bus *i* at nominal voltage (p.u.).
P_D, Q_D	Total system real and reactive power demand (p.u.).
P_{DG}, Q_{DG},S_{DG}	Real, reactive, and MVA power of DG (p.u.).
P_i, Q_i	Real and reactive power injection at bus i (p.u.).
P_{intake}, Q_{intake}	Real and reactive power intake at main substation (p.u.).
P_L, Q_L	System real and reactive power loss (p.u.).
$P_{i,j}$, $Q_{i,j}$, $S_{i,j}$	Real, reactive and MVA Power flows in line *i-j* (p.u.).
T1,T2,T3,T4	Type1, Type2, Type3, and Type4 DG,
V_{0i} , V_i	Nominal voltage at *i*th bus (p.u.), Voltage of *i*th bus (p.u.).
$V_{step\ i}$, *VSL*	Voltage step at *i*th bus (p.u.), Voltage step limit (%).
WDG, *WODG*	With and without DG.
WVSL, *WOVSL*	With and without *VSL*.
N_B , N_L	Number of buses and number of lines.
ld, lg	Leading, lagging power factors
$Y_{ij}=G_{ij}+jB_{ij}$	Elements of the bus admittance matrix corresponding to buses *i* and *j*.
r_{ij}	Resistance of line *i-j* (p.u.).

References

1. Ackermann, G.T., Andersson, G., Soder, L.: Distributed generation: a definition. Electric Power Systems Research 57(3), 195–204 (2001)
2. Chiradejaand, P., Ramakumar, R.: An approach to quantify the technical benefits of distributed generation. IEEE Trans. Energy Convers. 4, 764–773 (2004)
3. El-Khattam, W., Salama, M.M.A.: Distributed Generation Technologies, Definitions and Benefits. Electric Power System Research 71(2), 119–128 (2004)
4. Pepermans, G., Driesen, J., Haeseldonckx, D., Belmans, R., D'haeseleer, W.: Distributed generation: definition, benefits and issues: Energy Policy, vol. 33, pp. 787–798 (2005)
5. IEEE Task Force on Load Representation for Dynamic Performance: Load representation for dynamic performance analysis. IEEE Trans. Power Syst. 8(2), 472–482 (1993)

6. Qian, K., Zhou, C., Allan, M., Yuan, Y.: Effect of load models on assessment of energy losses in distributed generation planning. Int J. Electrical power and Energy Systems 33, 1243–1250 (2011)
7. Hung, D.Q., Mithulananthan, N., Bansal, R.C.: Analytical expression for DG allocation in primary distribution network. IEEE Trans. Energy Convers. 25(3), 814–820 (2010)
8. Dent, C.J., Ochoa, L.F., Harrison, G.P.: Network distributed generation capacity analysis using OPF with voltage step constraints. IEEE Trans. Power Syst. 25(1), 296–304 (2010)
9. Singh, D., Mishra, R.K., Singh, D.: Effect of Load Model in Distributed Generation Planning. IEEE Trans. Power. Syst. 22(4), 2204–2212 (2007)
10. Singh, D., Singh, D., Verma, K.S.: Multiobjective Optimization for DG Planning with Load Models. IEEE Trans. Power System 24(1), 427–436 (2009)
11. Singh, D., Mishra, R.K.: Multi-objective feeder reconfiguration in different tariff structures. IET Gen. Trans. Distr. 4(8), 974–988 (2010)
12. Zhu, D., Broadwater, R.P., Tam, K.S., Seguin, R., Asgeirsson, H.: Impact of DG placement on reliability and efficiency with time-varying loads. IEEE Trans. Power Syst. 21(1), 419–427 (2006)
13. Ochoa, L.F., Padilha-Feltrin, A., Harrison, G.P.: Tme-series based maximization of distributed wind power generation integration. IEEE Trans. Energy Conv. 23(3), 968–974 (2008)
14. Atwa, Y.M., El-Saadany, E.F., Salama, M.M.A., Seethapathi, R.: Optimal renewable resources mix for distribution system energy loss minimization. IEEE Trans. Power Syst. 25(1), 360–370 (2010)
15. El-Khattam, W., Hegazy, Y.G., Salama, M.M.A.: An integrated distributed generation optimization model for distribution system planning. IEEE Trans., Power System 20(2), 1158–1165 (2005)
16. Vovos, P.N., Harrison, G.P., Wallace, A.R., Bialek, J.W.: Optimal power flow as a tool for fault level-constrained network capacity analysis. IEEE Trans. Power Syst. 20(2), 734–741 (2005)
17. Vovos, P.N., Bialek, J.W.: Direct incorporation of fault level constraints in optimal power flow as a tool for network capacity analysis. IEEE Trans. Power Syst. 20(4), 2125–2134 (2005)
18. Harrison, G.P., Piccolo, A., Siano, P., Wallace, A.R.: Hybrid GA and OPF evaluation of network capacity for distributed generation connections. Elect. Power Syst. Res. 78(3), 392–398 (2008)
19. Ayed, A., Algarni, S., Bhattacharya, K.: Disco operation considering DG units and their goodness factors. IEEE Trans. Power System 24(4), 1831–1840 (2009)
20. Kumar, V., Rohith Kumar, H.C., Gupta, I., Gupta, H.O.: DG Integrated approach for service restoration under cold load pickup. IEEE Trans. Power Del. 25(1), 398–406 (2010)
21. Kumar, A., Gao, W.: Optimal distributed generation location using mixed integer non-linear programming in electricity markets. IET Gen. Trans. & Distrib. 4(2), 281–298 (2010)
22. Payasi, R.P., Singh, A.K., Singh, D.: Review of Distributed Generation Planning: Objectives, Constraints and Algorithms. Int. j. of Sc., Engg, and Technology 3(3), 133–153 (2011)
23. Payasi, R.P., Singh, A.K., Singh, D.: Planning of different types of distributed generation with seasonal mixed load models. Int. j. of Sc., Engg., and Technology 4(1), 112–124 (2012)
24. Baran, M.E., Wu, F.F.: Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans. Power Del. 4(2), 1401–1407 (1989)

Microstrip Patch Antenna Miniaturization Using Planar Metamaterial Unit Cell

Indrasen Singh[1], Sanjeev Jain[1], Vijay Shanker Tripathi[1], and Sudarshan Tiwari[2]

[1] Department of Electronics and Communication Engineering
Motilal Nehru National Institute of Technology
Allahabad, 211004, India
[2] National Institute of Technology, Raipur 492010, India
{erindrasen,snjece}@gmail.com, {vst,stiwari}@mnnit.ac.in

Abstract. A Microstrip patch antenna using planar metamaterial unit cell is designed, simulated and analyzed. The metamaterial unit cell is consisting of an interdigital capacitor and a complementary split-ring resonator (CSRR) slot. The antenna is tuned to work efficiently in the frequency range from 3 GHz – 5GHz depending on the geometric specifications of antenna and interdigital finger length. Proposed antenna provides good return loss behavior. The VSWR obtained in this band is very much near to 1. It covers many applications including mobile communication. This Antenna is compared with the conventional patch antenna, which shows the significant miniaturization as compared to conventional patch antenna.

Keywords: Metamaterial, Return loss, CSRR, VSWR.

1 Introduction

Electromagnetic metamaterials have been a field of intense research activity over the past decade. Due to the demand of small, compact, low cost antennas in the various military and commercial wireless communications it has increased tremendously over the past years. Metamaterials are periodic metallic structures exhibiting unique properties not existing in natural materials. Such materials exhibits permittivity and permeability both negative and hence they are known as the Double Negative (DNG) materials, backward wave materials, or left-handed materials (LHM). Since they have Negative Refractive index (NRI) and hence they are also called as the Negative Refractive Index materials or Left handed material LHM (as they follow left hand rule). This idea is proposed by Veselago 1968 [1]. In which the extraordinary electromagnetic features of the negative index medium were predicted. In this paper a new idea is proposed for miniaturization of patch antenna using interdigital capacitor and complementary split ring resonator. The transmission line (TL) approach of metamaterials was established in 2002 [2]–[5]. Metamaterial TLs are called composite right/left-handed (CRLH) TLs because they have both right- and left-handed properties. In other words, a CRLH TL supports not only a positive phase constant, but also a negative phase constant in a specific frequency region and a zero

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 358–364, 2013.

phase constant at a nonzero frequency. Because their unusual properties they offer some interesting changes in radiation characteristics of an antenna. The idea of MTMs has been quickly adopted in research, due to rapidly-developing nanofabrication and sub-wavelength imaging techniques. [7].

In this paper, a small and wideband microstrip patch antenna loaded with a planar CRLH unit cell is presented. In order to impose CRLH properties on a patch antenna, the antenna includes an interdigital capacitor for series capacitance and a complementary split-ring resonator (CSRR) slot for shunt inductance. CSRR slots can be coupled with a TL or a waveguide in order to achieve CRLH characteristics. Owing to the CSRR and the interdigital capacitor, a CRLH unit cell is implemented in fully planar technology, and its dispersion characteristics are analyzed for small antenna application. In addition, the current distributions circulating around the CSRR slot induce a unique radiation mode that is orthogonal to the normal radiation mode. Moreover, combining two radiation modes provides a wideband property and a unique radiation pattern with high antenna efficiency, which is verified.

2 Antenna Design

The proposed antenna configuration made up of planar metamaterial unit cell is shown in fig. 1. In order to make a single planar CRLH unit cell, an interdigital capacitor is inserted into the patch which acts as series capacitance and CSRR slot is cut on the ground plane for shunt admittance. The equivalent circuit model of the CRLH unit cell is shown in Fig. 2. Interdigital capacitor ensures negative permittivity and CSRR ensures the negative permeability [8]. When both the structures are combined together then they simultaneously offer negative permittivity and negative permeability and hence it becomes double negative metamaterial.

The proposed patch antenna is designed on Teflon substrate ($\mu_r = 2.1$ and Tanδ = 0.001) with thickness of 1.57 mm and fed by a microstrip transmission line. Patch and

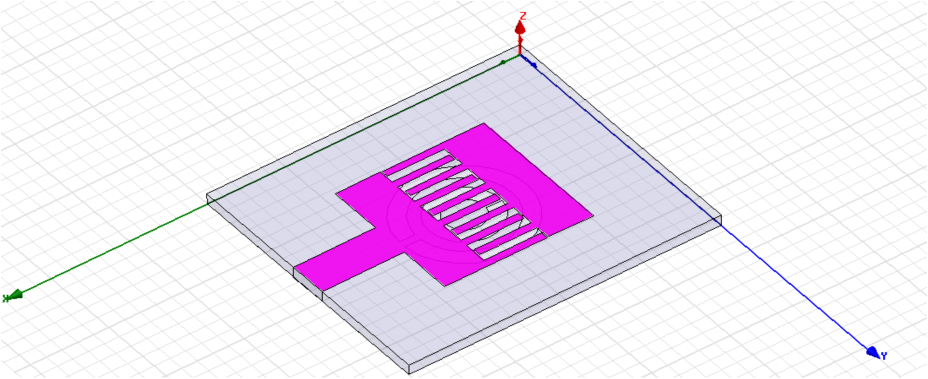


Fig. 1. Proposed antenna configuration in HFSSv12

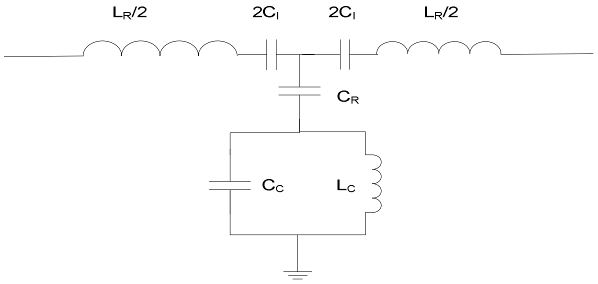


Fig. 2. Equivalent circuit model of the CRLH unit cell of the proposed antenna

ground plane are made of copper having relative permittivity as $\mu_r = 1$. Convergence was tested for each case separately in terms of evaluating S11 (dB) at a single frequency for a number of times. Once convergence was obtained, simulations were conducted in order to obtain swept frequency response extending from 4 to 4 GHz. The swept response gave us the S11, which was used to calculate the VSWR.

$$Z = j\omega L_R + \frac{1}{j\omega C_I}$$

$$Y = j\omega C_R // \left(j\omega C_C + \frac{1}{j\omega L_C} \right)$$

Simulations were performed using HFSS™ [12]. HFSS (High Frequency Structure Simulator) is the industry-standard simulation tool for 3D full-wave electromagnetic field simulation. HFSS provides E- and H-fields, currents, S-parameters and near and far radiated field results. It integrates simulation, visualization, solid modeling, and automation. Ansoft HFSS employs the Finite Element Method (FEM) for EM simulation by developing/ implementing technologies such as tangential vector finite elements and adaptive meshing.

3 Results Analysis

The proposed antenna contains a single planar CRLH unit cell composed of a CSRR slot and an interdigital capacitor. By increasing the interdigital finger length, the electrical size of the antenna was decreased due to the increased series capacitance. The proposed antenna achieves a 45% reduction in patch size compared to a conventional patch antenna. Additionally, the increased interdigital finger length along with the CSRR slot generates the mode radiation, which can be combined with the normal mode. The combination of these two modes provides a wideband property (6.8%) and unique radiation pattern that are near-isotropic for the horizontal polarization and dipolar for the vertical polarization. Regardless of the small size of the proposed antenna, very high efficiency (96%) and moderate gain (3.85 dBi) are attained. The gain of the proposed antenna is only 2.2 dB lower than that of a conventional one operating at the same frequency band with the same ground size. Based on the antenna performances mentioned in the previous sections such as small size, high efficiency, and near-isotropic radiation pattern, one can conclude that the proposed antenna is applicable for a mobile RFID reader system requiring isotropic coverage

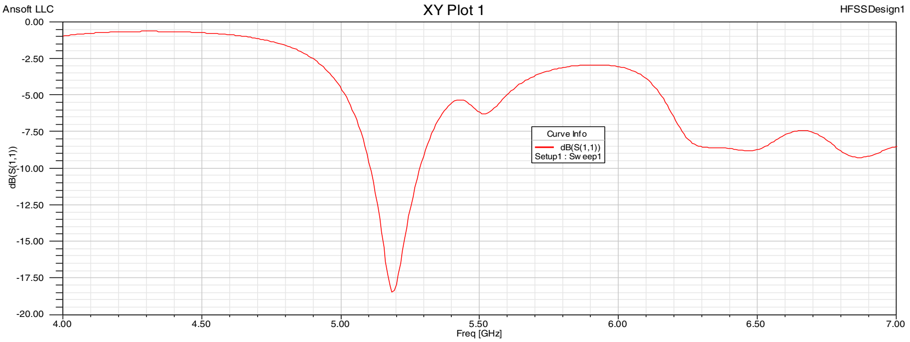


Fig. 3. Simulated Return loss of proposed patch antenna

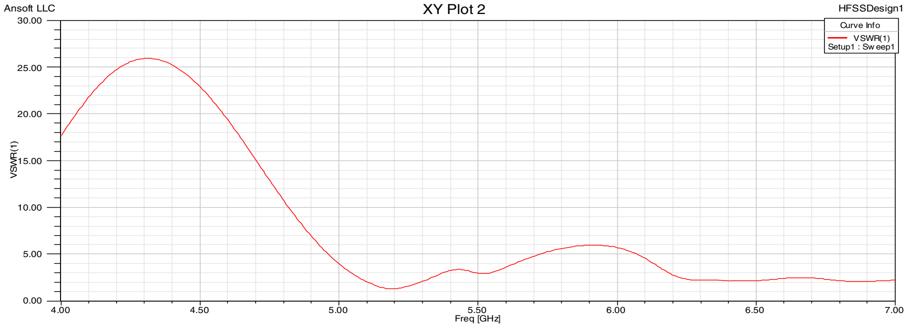


Fig. 4. Simulated VSWR of proposed patch antenna

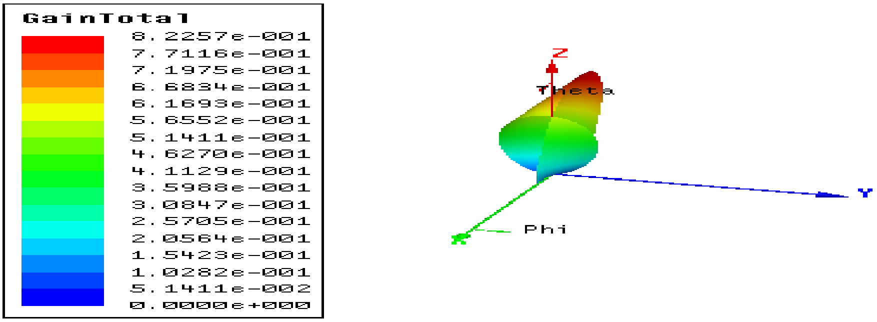


Fig. 5. 3D Polar plot of proposed patch antenna

4 Conclusion

Return loss characteristics of the proposed patch antenna with CSRR and inter digital capacitor is observed and plotted in Fig. 3. From the Figure it is concluded that antenna gives better return loss in between 3GHz to 5GHz. The percentage Bandwidth observed at -10 dB return loss is 2.4 and the gain of antenna is around 8. The corresponding VSWR plot for the same antenna is shown in the Fig. 4 VSWR values obtained as 1.3572, which is very much nearer to ideal value 1 for each case. The proposed antenna is suitable for RFID and mobile communication Application.

References

1. Veselago, V.G.: The electrodynamics of materials with simultane-ously negative values of and. Soviet Phys. 10(4), 509–514 (1968)
2. Antoniades, M., Eleftheriades, G.V.: A folded-monopole model for electrically small NRI-TL metamaterial antennas. IEEE Antennas Wireless Propag. Lett. 7, 425–428 (2008)
3. Caloz, C., Itoh, T.: Application of the transmission line theory of left-handed (LH) materials to the realization of a microstrip "LH line". In: Proc. IEEE Antennas Propag. Society Int. Symp. (AP-S), San Antonio, TX, June 16-21, pp. 412–415 (2002)
4. Dong, Y., Itoh, T.: Metamaterial-inspired broadband mushroom antenna. In: Int. Symp. AP-S Digest, Toronto, ON, pp. 1–4 (July 2010)
5. Ziolkowski, R.W., Erentok, A.: Metamaterial-based efficient electrically small antennas. IEEE Trans. Antennas Propag. 54(7), 2113–2130 (2006)
6. Ha, J., Kwon, K., Lee, Y., Choi, J.: Hybrid Mode Wideband Patch Antenna Loaded With a Planar Metamaterial Unit Cell. IEEE Trans. Antennas Propag. 60(2), 1143–1147 (2012)
7. Dong, Y., Toyao, H., Itoh, T.: Compact Circularly-Polarized Patch Antenna Loaded With Metamaterial Structures. IEEE Trans. Antennas Propag. 59(11), 4329–4333 (2011)
8. Mosallaei, H., Sarabandi, K.: Antenna Miniaturization and Bandwidth Enhancement Using a Reactive Impedance Substrate. IEEE Trans. Antennas Propag. 2, 2403–2414 (2004)
9. Ziolkowski, R.W.: Design, fabrication, and testing of double negative metamaterials. IEEE Trans. Antennas Propagat. 51(7), 1516–1529 (2003)
10. Caloz, C., Itoh, T.: Electromagnetic metamaterials: Transmission line theory and microwave applications. John Wiley & Sons (2009)
11. Singh, I., Tripathi, V.S.: Microstrip Patch Antenna and Its Applications: A Survey. Published in International Journal of Computer Technology and Applications 2(5), 1595–1599 (2011)
12. Marques, R., Martin, F., Sorolla, M.: Metamaterials With Negative Parameters: Theory, Design and Microwave Applications. Wiley, New York (2008)
13. Stuart, H.R., Pidwerbetsky, A.: Electrically small antenna ele-ments using negative permittivity resonators. IEEE Trans. Antennas Propag. 54(6), 1644–1653 (2006)

14. Ziolkowski, R.W., Erentok, A.: Metamaterial-based efficient electrically small antennas. IEEE Trans. Antennas Propag. 54(7), 2113–2130 (2006)
15. Lee, C., Leong, K.M., Itoh, T.: Composite right/left-handed trans-mission line based compact resonant antennas for RF module integration. IEEE Trans. Antennas Propag. 54(8), 2283–2291 (2006)
16. Alu, A., Bilotti, F., Engheta, N., Vegni, L.: Subwavelength, compact, resonant patch antennas loaded with metamaterials. IEEE Trans. An-tennas Propag. 55(1), 13–25 (2007)
17. Park, J.H., Ryu, Y.H., Lee, J.G., Lee, J.H.: Epsilon negative zeroth-order resonator antenna. IEEE Trans. Antennas Propag. 55(12), 3710–3712 (2007)
18. Antoniades, M., Eleftheriades, G.V.: A folded-monopole model for electrically small NRI-TL metamaterial antennas. IEEE Antennas Wireless Propag. Lett. 7, 425–428 (2008)
19. Dong, Y., Itoh, T.: Miniaturizedsubstrate integrated waveguide slot antennas based on negative order resonance. IEEE Trans. Antennas Propag. 58(12) (2010)
20. Kokkinos, T., Sarris, C.D., Eleftheriades, G.V.: Periodic FDTD analysis of leaky-wave structuresand applications to the analysis of negative-refractive-index leaky-wave antennas. IEEE Trans. Microw. Theory Tech. 54(4), 1619–1630 (2006)
21. Ueda, T., Michishita, N., Akiyama, M., Itoh, T.: Dielectric resonator based composite right/left-handed transmission lines and their applica-tion to leaky wave antenna. IEEE Trans. Microw. Theory Tech. 56(10), 2259–2268 (2008)
22. Paulotto, S., Baccarelli, P., Frezza, F., Jackson, D.: Full-wave modaldispersion analysis and broadside optimization for a class of microstrip CRLH leaky-wave antennas. IEEE Trans. Microw. Theory Tech. 56(12), 2826–2837 (2008)
23. Kodera, T., Caloz, C.: Uniform ferrite-loaded open waveguide structure with CRLH response and its application to a novel back-fire-to-endfire leaky-wave antenna. IEEE Trans. Microw. Theory Tech. 57(4), 784–795 (2009)
24. Dong, Y., Itoh, T.: Composite right/left-handed substrate integrated waveguide and half mode substrate integrated waveguide leaky-wave structures. IEEE Trans. Antennas Propag. 59(3), 767–775 (2011)
25. Pendry, J.B., Holden, A.J., Robbins, D.J., Stewart, W.J.: Magnetism from conductors and enhanced nonlinear phenomena. IEEE Trans. Microw. Theory Tech. 47(11), 2075–2084 (1999)
26. Falcone, F., Lopetegi, T., Baena, J.D., Marques, R., Martin, F., Sorolla, M.: Effective negative-epsilon stopband microstrip lines based on complementary split ring resonators. IEEE Microw. Wireless Compon. Lett. 14(14), 280–282 (2004)
27. Bonache, J., Gil, I., Garcia, J., Martin, F.: Complementary split ring resonators for microstrip diplexer design. Electron. Lett. 41(14) (July 2005)
28. Niu, J., Zhou, X.: A novel dual-band branch line coupler based on strip-shaped complementary split ring resonators. Microw. Opt. Technol. Lett. 49(11), 2859–2862 (2007)
29. Zhang, Y., Hong, W., Yu, C., Kuai, Z., Dong, Y., Zhou, J.: Planar ultrawideband antennas with multiple notched bands based on etched slots on the patch and/or split ring resonators on the feed Line. IEEE Trans. Antennas Propag. 56(9), 3063–3068 (2008)
30. Zhang, H., Li, Y.Q., Chen, X., Fu, Y.Q., Yuan, N.C.: Design of circular polarization microstrip patch antennas with complementary split ring resonator. IET Microw. Antennas Propag. 3(8), 1186–1190 (2009)

31. Dong, Y., Yang, T., Itoh, T.: Substrate integrated waveguide loaded by complementary split-ring resonators and its applications to minia-turized waveguidefilters. IEEE Trans. Microw. Theory Tech. 57(9), 2211–2223 (2009)
32. Eggermont, S., Platteborze, R., Huynen, I.: Investigation of meta-material leaky wave antenna based on complementary split ring resonators. In: Proc. Eur. Microw. Conf., Rome, Italy, pp. 209–212 (September 2009)
33. Zhang, H., Li, Y.Q., Chen, X., Fu, Y.Q., Yuan, N.C.: Design of circular/dual-frequency linear polarization antennas based on the anisotropic complementary split ring resonator. IEEE Trans. Antennas Propag. 57(10), 3352–3355 (2009)
34. Ansoft HFSS, Ansoft Corporation, http://www.ansoft.co.jp/hfss.htm

Simulation and Modeling of a Constant Voltage Controller Based Solar Powered Water Pumping System

Bhavnesh Kumar*, Yogesh K. Chauhan, and Vivek Shrivastava

School of Engineering, Gautam Buddha University
Greater Noida, Uttar Pradesh, 201310, India
kumar_bhavnesh@yahoo.co.in, chauhanyk@yahoo.com,
shvivek@gmail.com

Abstract. In this paper the performance of a squirrel cage induction motor driving a water-pump load is investigated. The induction motor is fed through a Photovoltaic (PV) array. The PV array and the drive is designed and simulated in a Simulink/MATLAB environment. For tracking of maximum power point of PV array a constant voltage controller (CVC) is also presented. The performance of directly connected and constant voltage controller fed water pumping system is compared for different operating conditions. The system with constant voltage controller is found to have better performance as compared to directly connected system.

Keywords: centrifugal pump, constant voltage controller, induction motor, PV array.

1 Introduction

Solar energy is one of the most promising renewable energy for future, particularly for the rural areas which are not connected to the national grid. In rural areas perhaps the most common practice for which they require electrical energy is the water pumping employed for the irrigation of fields. Solar energy can be directly converted into the electrical energy with the help of solar cell, which is basically a P-N junction that absorbs light, releases electrons and create holes to produce a voltage in the cell. However, the efficiency of solar cell is poor due to the involvement of number of stages for energy conversion before being available in the useful form [1].

For water pumping applications, usually low rating pumps in the range of 200-2000W are used in conjunction with the PV array. Centrifugal pumps and volumetric pumps are the two basic types of pumps used in the PV water pumping application. Generally, centrifugal pumps are preferred over other for the three important reasons: (i) in it PV energy utilization is higher, (ii) it can operate for long periods even for low insolation levels, and (iii) its load line is in close proximity to the maximum power point line [2].

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 365–376, 2013.

Although, several types of electric motors such as permanent magnet brushed DC motor, permanent magnet synchronous motor, switched reluctance motor, and induction motors have also already been tested and used for water pumping system with the PV array. But, due to the low maintenance requirement, low cost and readily availability induction motors are preferred. The induction motors are designed to give optimal performance on the rated conditions. So, variations in operating condition of motor can deteriorate the performance of the motor and hence the system [3]-[4].

The power and current characteristics of PV array are highly nonlinear and depends upon the various factors such as intensity of sunlight, temperature and cell area. Generally, these factors vary throughout day and hence the power of the PV array. So, for the full exploitation of PV array its maximum output power has to be tracked with the help of a controller under varying operating conditions [5]-[6].

Various MPP tracking schemes such perturbation and observation or hill climbing, incremental conductance and artificial intelligence based schemes had been addressed in the literature. Constant voltage controller proposed in [7] can be a good method for tracking of maximum power from PV array on different operating conditions. The method utilizes the fact that the maximum power line is almost linear in a narrow band of particular voltage. However, the work is limited to DC motor, which has higher cost and frequent maintenance problem. This work can be extended to assess the performance of system with induction motor.

In this paper the performance of the PV array fed pump coupled three phase induction motor with a constant voltage controller is investigated. The system under investigation is compared with the directly connected system for various operating conditions.

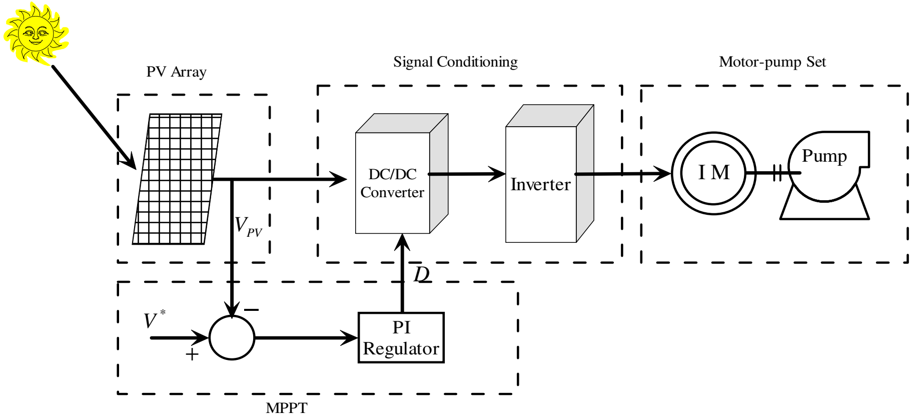


Fig. 1. Schematic diagram for constant voltage controlled system

2 System Description and Modeling

The system under investigation consists of a PV array, DC-DC converter, induction motor, and centrifugal pump. Simple but accurate model of PV array and centrifugal pump are derived in order to simulate the system. All components are modeled separately and then joined together. The Schematic diagram of the system under investigation is shown in Fig. 1.

2.1 PV Array Model

Interconnected number of solar cells in series/parallel combination is known as PV modules. A group of such modules connected in series/parallel combination to generate the required power is known as PV Array. For describing the electrical behavior of a solar cell, different mathematical models have been introduced. The most commonly used equivalent model is one diode or the two diode equivalent model [1, 3, 5]. As shown in Fig. 2 one diode equivalent model of solar cell is used because of its simple structure.

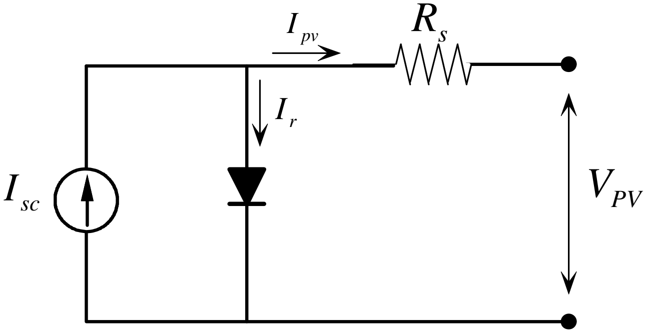


Fig. 2. Equivalent circuit of solar cell

The PV array has a non-linear output which varies with the variation in level of solar insolation and temperature. So, the effects of the changes in temperature and solar irradiation levels are also included in the final PV array model. The cell output voltage is given by eq. 1 [5].

$$V_{PV} = \frac{nkT_C}{q}\ln\left(\frac{I_{sc} + I_r - I_{pv}}{I_r}\right) - I_{pv}R_s \tag{1}$$

Where V_{PV} and I_{pv} are output voltage and currents of the cell respectively, R_s is cell resistance, I_{sc} is photocurrent or short circuit current, I_r is reverse saturation current of diode, q is electron charge, k is Boltzmann constant, T_r is reference operating temperature of cell and n is the ideality factor.

The solar cell operating temperature varies as a function of solar insolation level and ambient temperature. The effects of change in temperature on output voltage and current of solar cell are incorporated with the help of eq. (2)-(3).

$$T_V = 1 + \beta(T_c - T_a) \tag{2}$$

$$T_I = \frac{\gamma}{G_c}(T_a - T_C) \tag{3}$$

where, T_V & T_I are the temperature coefficients of solar cell output voltage and current resp., T_a & T_c are the ambient and operating temperatures resp., and β & γ are the constants.

The effect of change in temperature due to the change in solar insolation level is incorporated with the help of eq. (4)-(5)

$$C_V = 1 + \beta\alpha(G_X - G_C) \tag{4}$$

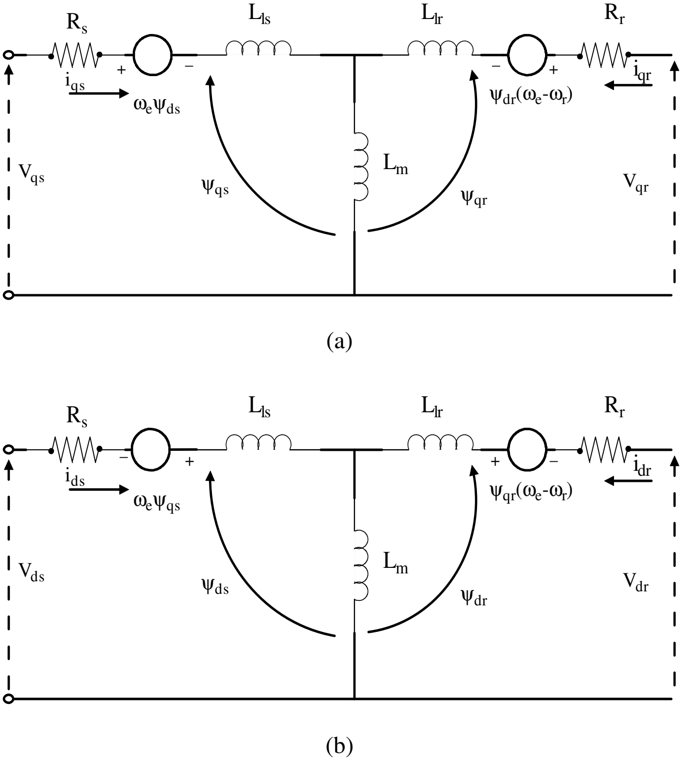
$$C_I = 1 + \frac{1}{G_C}(G_X - G_C) \tag{5}$$

Where, C_V & C_I are the correction factors for output voltage and current of solar cell respectively, G_X & G_C are the standard and operating insolation levels respectively, and α is a constant. The change in temperature ΔT_c due to change in insolation level is obtained by eq. (6)

$$\Delta T_C = \alpha(G_X - G_C) \tag{6}$$

2.2 Induction Motor Model

The dynamic equivalent circuit of a three phase induction motor expressed in *d-q* synchronously rotating reference frame is shown in Fig. 3 [8].



(a)

(b)

Fig. 3. Dynamic (a) *q- axis* (b) *d-axis* equivalent circuits of machine

where, R_s & R_r are the stator and rotor resistances resp., L_{ls} & L_{lr} are the stator and rotor leakage inductances resp., L_m is the mutual inductance, V_{ds} & V_{dr} are the *d*-axis stator and rotor voltages resp., V_{qs} & V_{qr} are *q*-axis stator and rotor voltages resp., ψ_{qs} & ψ_{qr} are the *q*-axis stator and rotor flux linkages resp., ψ_{ds} & ψ_{dr} are the *d*-axis stator and rotor flux linkages respectively.

The electromagnetic torque developed by an induction motor is given by:

$$T_e = \frac{3}{2}\frac{P}{2} L_m (I_{qs} I_{dr} - I_{ds} I_{qr}) \tag{7}$$

The mechanical part modeling of an electric motor is given by:

$$T_e = Jp\omega_m + B\omega_m + T_L \tag{8}$$

where J is the total inertia of motor shaft, B is the friction coefficient, and T_L is the load torque.

2.3 Centrifugal Pump Model

The centrifugal pump is identified by its head-flow rate (H-Q) performance curve at the nominal speed. Affinity laws are generally used to estimate the performance curve of the pump. Affinity Laws describes that the flow rate (Q), head (H) and power (P) is directly proportional to the speed, square of speed and cube of the speed respectively [9].

A centrifugal pump load is generally modeled in the form of a load torque applied to motor shaft. This load torque depends on the process requirements of head to be overcome, flow rate requirement, and the operating speed. The torque speed characteristic of the motor for a pump load is given by eq. (9) and shown in fig. 4[2].

$$T_L = T_P = K\omega_r^{\ 2} \tag{9}$$

where K is defined in terms of nominal power P_n and nominal speed ω_n of the centrifugal pump as:

$$K = \frac{P_n}{\omega_n^{\ 3}}$$

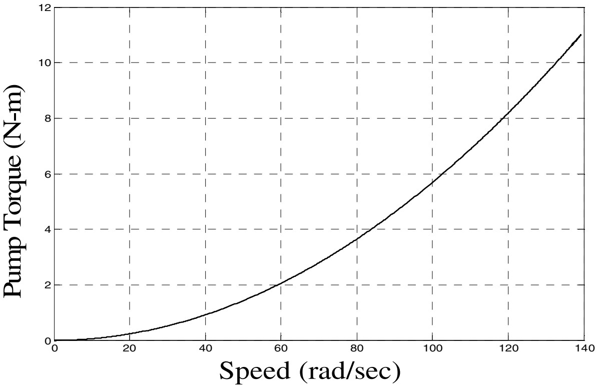


Fig. 4. Pump-torque characteristic of centrifugal pump

3 Constant Voltage Controller

In this work, a constant voltage controller (CVC) is designed to track the point of maximum power of PV array. The CVC in this work is basically a fixed gain Proportional–Integral (PI) controller used to control the DC bus voltage. The block diagram of the designed CVC scheme is shown in Figure 5.

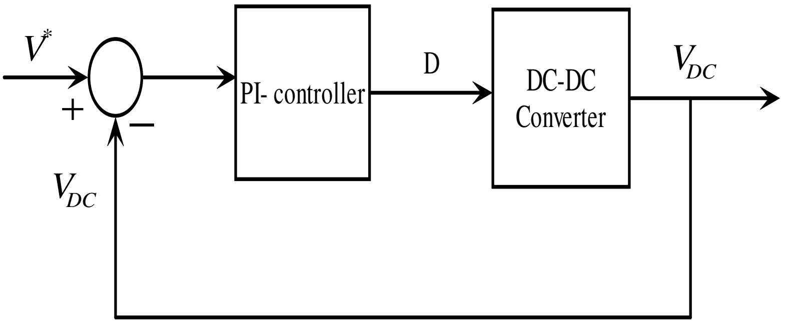


Fig. 5. Block diagram of constant voltage control

The voltage of a DC link is measured and feed backed to the comparator. In comparator, the measured value is compared with reference value of voltage and accordingly an error signal for the PI regulator is generated. PI regulator is used to generate the duty ratio of the controlled DC- DC converter circuit to maintain the constant voltage. The controller gains $K_p = 3$ and $K_i = 10$ are used after optimization by trial and error method.

4 Results and Discussion

The complete model of solar powered water pumping system in two modes i.e. directly connected and constant voltage controlled are simulated in the MATLAB with the sampling frequency of 50 KHz. As the output of the PV array is DC only so an inverter will always be an interfacing device for connecting the induction motor. A three phase, 2.2 KW, 220 V, 50 Hz induction motor is used in this work. The specifications of Induction motor are given in Table I of Appendix.

Fig. 6 shows the current-voltage (*I-V*) characteristics for the designed PV array. The *I-V* characteristics are obtained at different insolation levels of 1000 W/m^2, 800 W/m^2, 600W/m^2, 400 W/m^2, and 200 W/m^2. Fig. 7 shows the power-voltage (*P-V*) characteristics of the designed PV array at different insolation level. The PV array is designed to give short circuit current (I_{sc}) = 40 A and open circuit voltage (V_{oc}) = 230 V at an insolation of 1000 W/m^2 on an ambient temperature of 20ºC.

With an objective to investigate the performance of the system close to the real operating conditions the insolation level is randomly varied from 1000 W/m^2 to 600 W/m^2 as shown in Fig. 8. The response of PV Array voltage feeding directly a motor pump load is shown in Fig. 9. From the PV array voltage response it is evident that voltage varies about 10 % when the insolation level is varied from 1000 W/m^2 to 600 W/m^2. This variation in DC link/PV array voltage with the variation in insolation level varies the motor speed as can be depicted in Fig. 10.

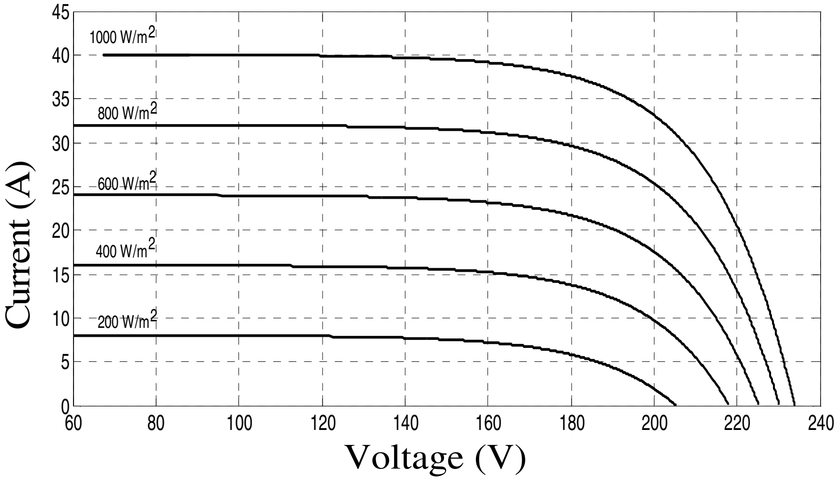


Fig. 6. Current-Voltage *(I-V)* characteristics at various insolation

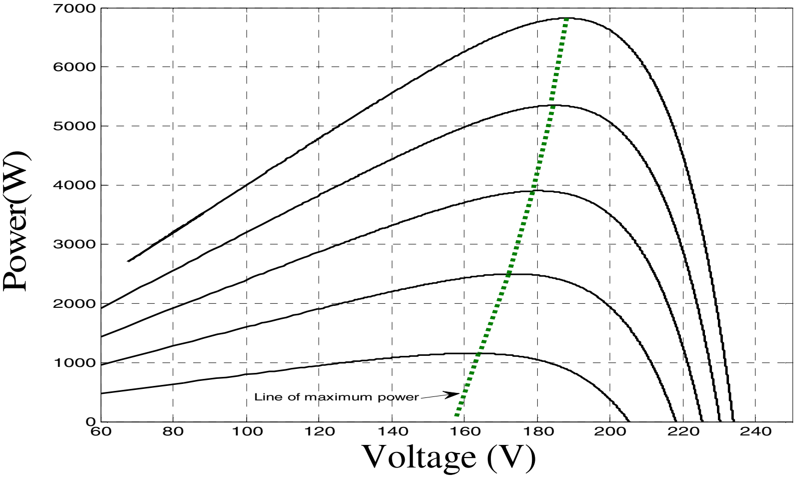


Fig. 7. Power-Voltage *(P-V)* characteristics at various insolation

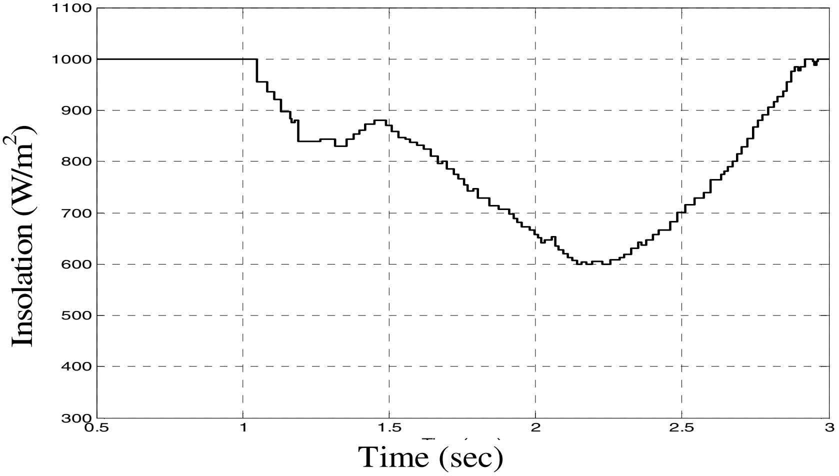


Fig. 8. Insolation level variation for directly connected system

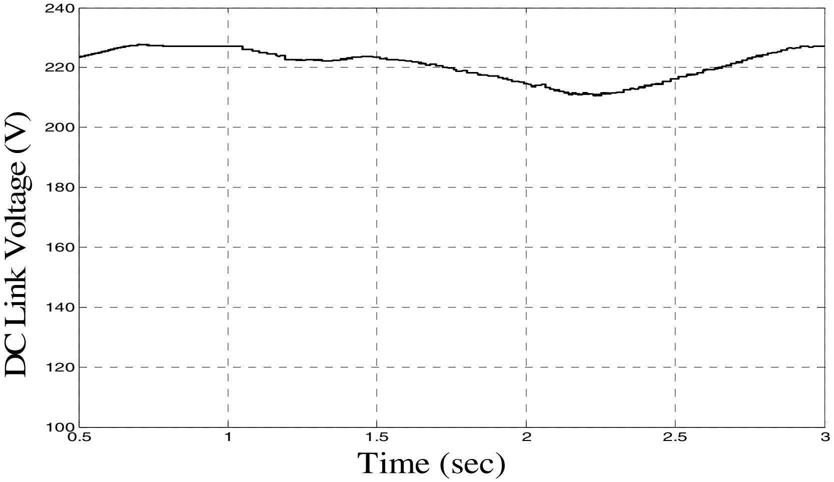


Fig. 9. Response of DC link voltage for directly connected system

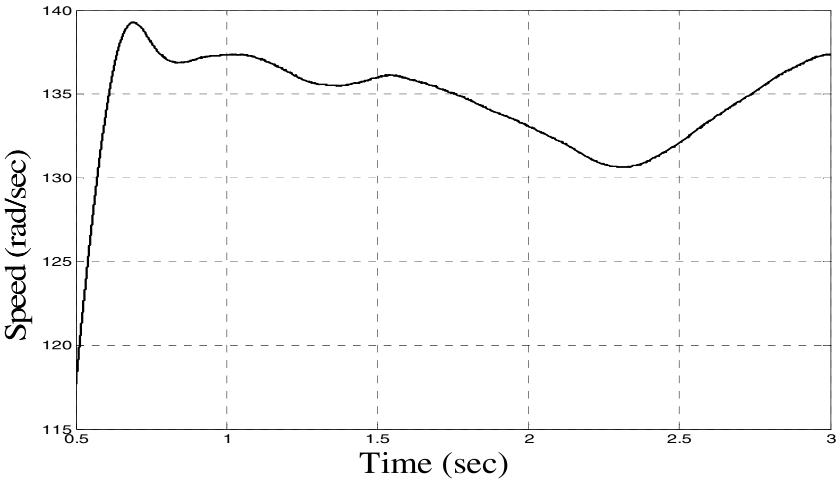


Fig. 10. Response of motor speed for directly connected system

Fig. 11 shows the variation in insolation level with maximum of 1000 W/m^2 and minimum of 600 W/m^2. Under these changing operating conditions the responses with the constant voltage controller are obtained. Fig. 12 shows the behavior of DC link/PV array voltage with the above operating conditions, from which almost constant voltage of 180 V is evident. Fig. 13 shows the speed response of the motor from which minor speed change is observed.

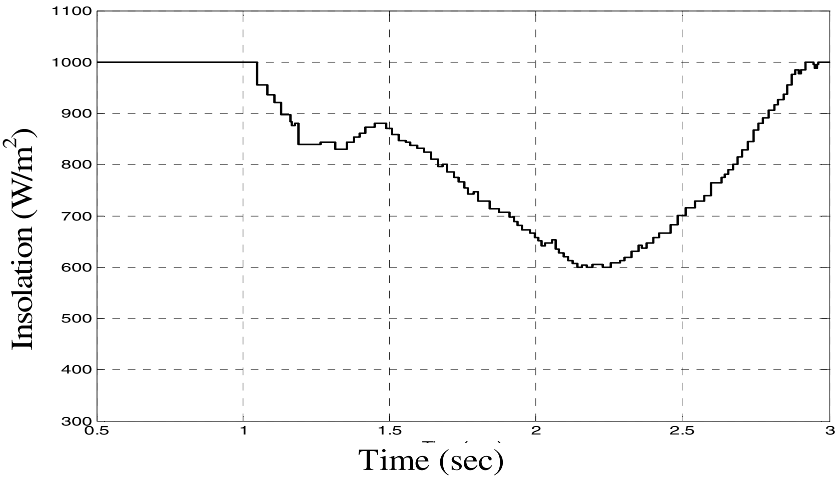


Fig. 11. Insolation level variation for constant voltage controlled system

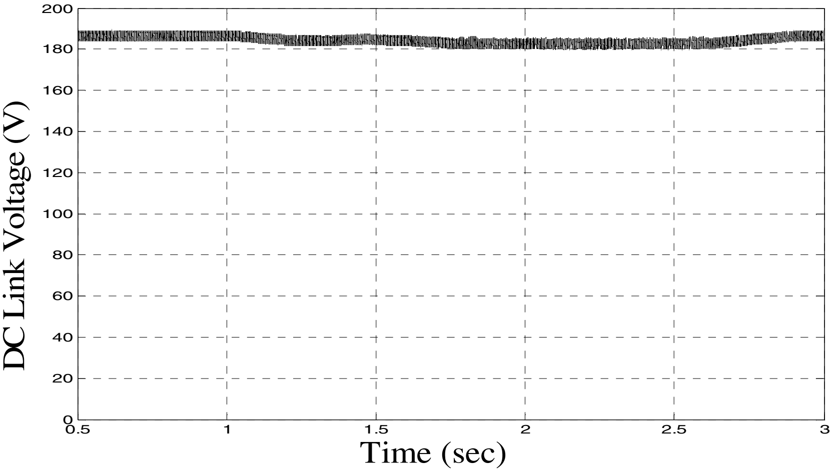


Fig. 12. Response of DC link voltage for constant voltage controlled system

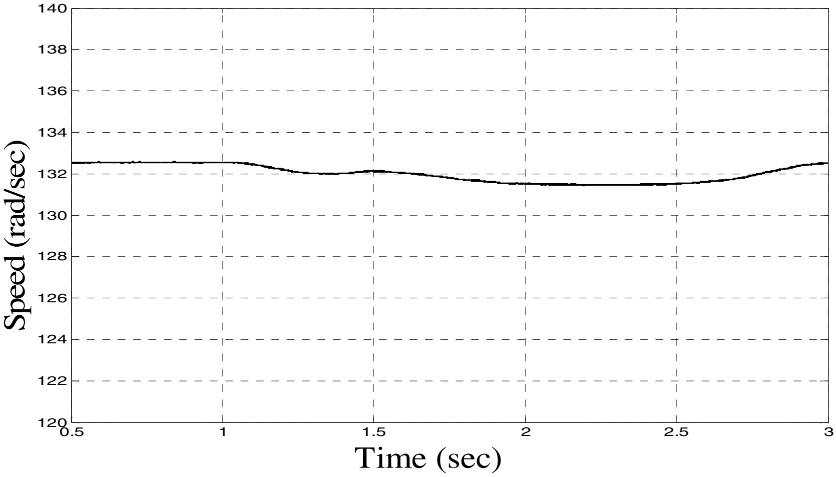


Fig. 13. Response of motor speed for constant voltage controlled system

5 Conclusion

In this paper, a one diode PV array model along with the other components of the induction motor driven water pump system are designed and simulated in the MATLAB/Simulink. The effects of solar insolation on the PV array assisted pump system have been investigated. A comparative study of performance of constant voltage controlled system with the directly connected PV system has been presented. The constant voltage controlled system offers good performance and therefore can be recommended for water pumping applications.

References

1. Singh, B.N., Singh, B., Singh, B.P., Chandra, A., Al-Haddad, K.: Optimized performance of solar powered variable speed induction motor drive. In: Proc. Int. Conf. Power Elect., Drives and Energy Systems for Ind. Growth, vol. 1, pp. 58–66 (1996)
2. Kini, P.G., Bansal, R.C., Aithal, R.S.: Performance analysis of centrifugal pumps subjected to voltage variation and unbalance. IEEE Trans. Ind. Elect. 55(2), 562–569 (2008)
3. Kolhe, M., Joshi, J.C., Kothari, D.P.: Performance analysis of a directly coupled photovoltaic water-pumping system. In: IEEE Trans. Energy Conv., vol. 19(3) (September 2004)
4. Ramya, K., Rama Reddy, S.: Design and Simulation of a photovoltaic induction motor coupled water pumping system. In: Proc. Int. Conf. Comp., Elex. and Elec. Tech., pp. 32–39 (2012)
5. Altas, I.H., Sharaf, A.M.: A photovolatic array simulation model for matlab-simulink GUI environment. In: Proc. IEEE Int. Conf. Clean Electrical Power, pp. 341–345 (2007)
6. Nayar, C.V., Vasu, E., Phillips, S.J.: Optimized solar water pumping system based on an induction motor driven centrifugal pump. In: Proc. IEEE Int. Conf. Comp., Comm., Control and Power Engg., vol. 5, pp. 383–393 (1993)
7. Elgendy, M.A., Zahawi, B., Atkinson, D.J.: Comparison of directly connected and constant voltage controlled photovoltaic pumping systems. IEEE Trans. Sustainable Energy 1(3), 184–192 (2010)
8. Bose, B.K.: Power Electronics and Variable Frequency Drives. IEEE Press (1997)
9. Biji, G.: Modelling and simulation of PV based pumping system for maximum efficiency. In: Proc. Int. Conf. Power, Signals, Controls and Computation, pp. 1–6 (2012)

Appendix

Table 1. Induction Motor Parameters

Parameters	Values
Stator Resistance (R_s)	0.435 Ω
Stator Inductance (L_{ls})	2.0 mH
Rotor Resistance (R_r)	0.816 Ω
Rotor Inductance (L_{lr})	2.0 mH
Mutual inductance (L_m)	69.3 mH
Inertia Constant (J)	0.02 Kg-m^2
Friction Factor (F)	0.002 N-m-sec

Design of Low Power FSM Using Verilog in VLSI

Himani Mittal, Dinesh Chandra, and Arvind Tiwari

J.S.S. Academy of Technical Education,
Noida, India
himanimit@yahoo.co.in

Abstract. With day by day increase in integration density of CMOS technology the concern for area usage for VLSI circuits is increased,giving more importance to timing and power dissipation constraints. Controllers are running continuosly so critical for power (while parts of the data path may be shut down), and for timing because the delay through the controller may constrain the delay through the data path.

In this work we propose a procedure for the decomposition of a network of interacting FSMs starting from a single state-table specification. The straightforward singlemachine implementation is called *undecomposed FSM.* We call *decomposed FSM* the interacting FSM implementation. The sub-machines in the decomposed FSM communicate through a set of l interface signals. The decomposed implementation has low power dissipation because *one single sub-machine is clocked* at any given time and it controls the outputs values while all other sub-machines are idle: they do not receive the clock signal and dissipate little power. When a sub-machine terminates its execution, it sends an *activation signal* to another sub-machine which takes control of the computation, then it de-activates itself.

Keywords: Heuristic Approach, Mealy and Moore Machines, STM (State Transition Matrix), ULP (Ultra Low Power), Power saving, Entropy, State Codes.

1 Introduction

Power consumption has become a major design parameter in the project of integrated circuits. Two independent factors have contributed for this. On one hand, low power consumption is essential to achieve longer autonomy for portable devices. On the other hand, increasingly higher circuit density and higher clock frequencies are creating heat dissipation problems, which in tum raise reliability concerns and lead to more expensive packaging.In static CMOS circuits, the probabilistic switching activity of nodes in the circuit is a good measure of the average power dissipation of the circuit. Methods that can efficiently compute the average switching activity, and thus power dissipation, in CMOS combinational [lo] and sequential [9] circuits have been developed. In this work, we are concemed with the problem of optimizing logic-level sequential circuits for low power. This problem has received some attention recently. Several techniques for state assignment have been presented which aim at

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 377–386, 2013.

reducing the average switching activity of the present state lines, and consequently of the intemal nodes in the combinational logic block (see for example **[7]).** Retiming has also been tailored so that the distribution of the registers within the logic block minimizes the total amount of glitching in the **sequential circuit** [5].Techniques based on disabling the inputlstate registers when some input conditions are met have been proposed .The most effective in reducing the overall switching activity in sequential circuits [l], [2], [3]. The disabling of the inputfstate registers is decided on a clock-cycle basis and can be done either by using ***B***register load-enable signal or by gating the clock . **A** common feature of these methods is the addition of extra circuitry that is able to identify input conditions for which some or all of the inpuustate register:can be disabled. In this situation there will be zero switching activity in the logic driven by input signals coming from the disabled registers.This class of techniques is sometimes referred to as *dynamiL* ***power*** *management.*The method we propose in this paper falls into this class of tech.niques. We use finite state machine (FSM) decomposition to obtain the conditions for which a significant part of the registers in the circuit can be disabled. The original FSM is divided into two sub-FSM where one of them is significantly smaller than the other. Except for transitions that involve going from one state in one sub-machine to ***i***state in the other, only one of the sub-machines needs to be clocked .By selecting for the small sub-FSM a cluster of states in which the original FSM has a high probability of being in, most of the time will be disabling all the state registers in the larger sub-FSM. The overhead associated with the FSM decomposition makes this method not very effective for typical FSMs with a small number of states. However, for large machines, impressive gains up to 80% power reduction are possible

1.1 Estimation of Power

Entropy is a measure of the randomness carried by a set of discrete events observed over time. In the studies of the information theory, a method to quantify the information content Ci of an event E_i in this manner is o take logarithmic of the event probability

$$C_i = \log_2 (1/P_i)$$

Since $0 \leq P_i \leq 1$, the logarithmic term is non negative and we have $C_i > 0$.

The average information contents of the system is the weighted sum of the information content of C_i by its occurrence probability This is also called the entropy of the system.

$$H(X) = \sum_{i=1}^{m-1} p_i \log_2 \frac{1}{p_i}$$

1.2 Problems Faced

The major shortcoming of this solution, however, is that lowering the supply voltage affects circuit speed. As a consequence, both design and technological solutions must be applied in order to compensate the decrease in circuit performance introduced by reduced voltage. In other words, speed optimization is applied first, followed by supply voltage scaling, which brings the design back to its original timing, but with a lower power requirement. A similar problem, i.e., performance decrease, is encountered when power optimization is obtained through frequency scaling. Techniques that rely on reductions of the clock frequency to lower power consumption are thus usable under the constraint that some performance slack does exist. Although this may seldom occur for designs considered in their entirety, it happens quite often that some specific units in a larger architecture do not require peak performance for some clock/machine cycles. Selective frequency scaling (as well as voltage scaling) on such units may thus be applied, at no penalty in the overall system speed. Optimization approaches that have a lower impact on performance, yet allowing significant power savings, are those targeting the minimization of the *switched capacitance* (i.e., the product of the capacitive load with the switching activity). Static solutions (i.e., applicable at design time) handle switched capacitance minimization through area optimization (that corresponds to a decrease in the capacitive load) and switching activity reduction via exploitation of different kinds of signal correlations (temporal, spatial, spatial-temporal). Dynamic techniques, on the other hand, aim at eliminating power wastes that may be originated by the application of certain system workloads (i.e., the data being processed).

2 Heuristic Algorithm Approach

(1) Crossing Transition should be minimum to reduce power consumption.[6]
(2) First bit of state code is Control bit distinguish between Sub machines.
(3) To distinguish between states within each submachine inner bits are needed.

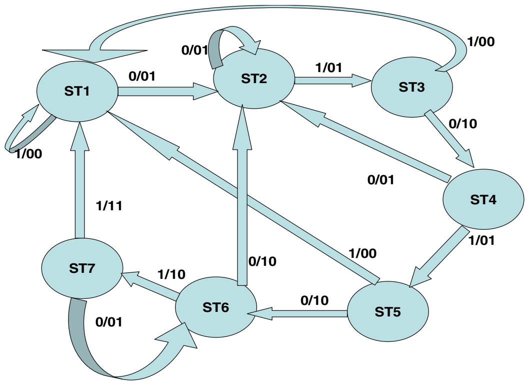


Fig. 1. Original FSM

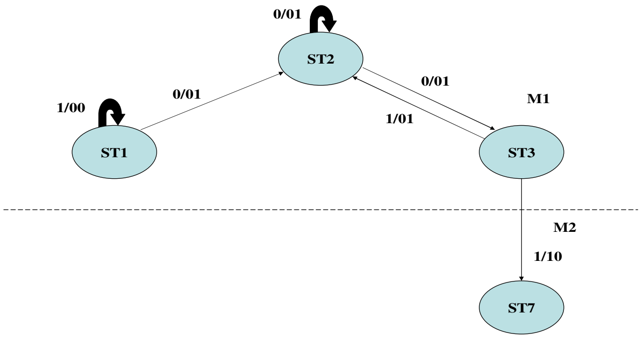


Fig. 2. Upper half FSM

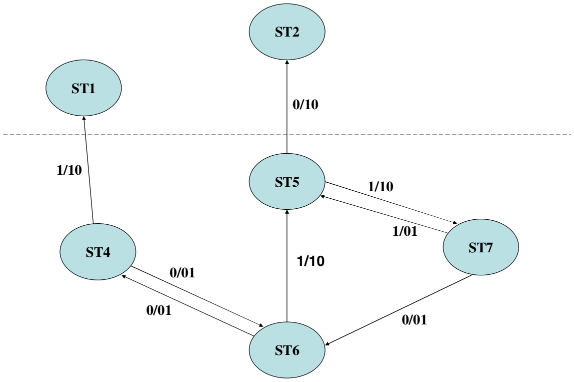


Fig. 3. Lower half FSM

2.1 Simulation of FSM

Following are the simulations over ModelSim for FSM

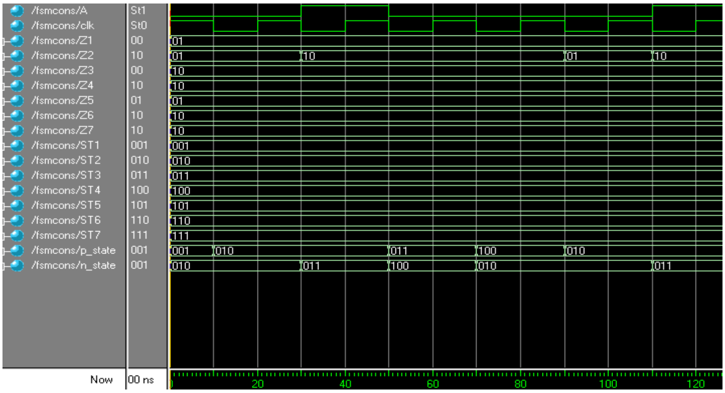
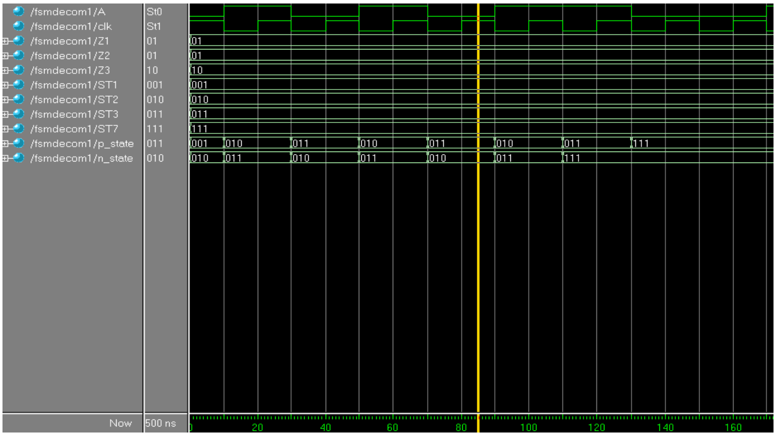
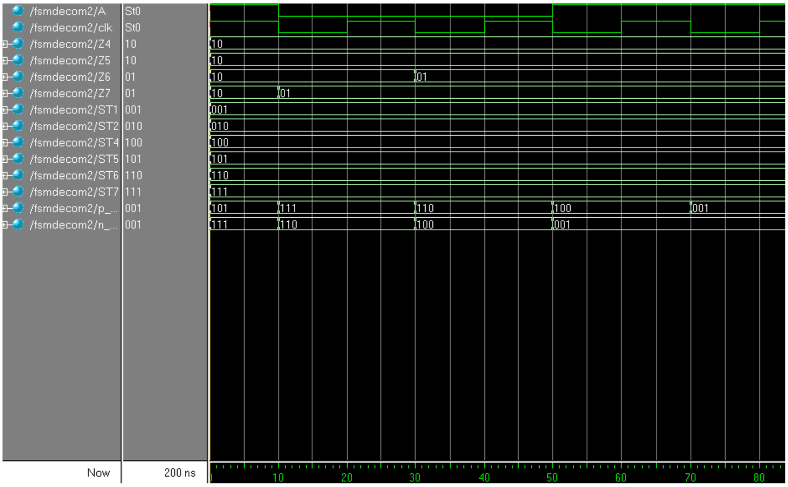


Fig. 4. State Transition in FSM for an input sequence



/fsmdecom1/A St0
/fsmdecom1/clk St1
/fsmdecom1/Z1 01
/fsmdecom1/Z2 01
/fsmdecom1/Z3 10
/fsmdecom1/ST1 001
/fsmdecom1/ST2 010
/fsmdecom1/ST3 011
/fsmdecom1/ST7 111
/fsmdecom1/p_state 011
/fsmdecom1/n_state 010
Now 500 ns

Fig. 5. State transition in Upper half FSM



/fsmdecom2/A St0
/fsmdecom2/clk St0
/fsmdecom2/Z4 10
/fsmdecom2/Z5 10
/fsmdecom2/Z6 01
/fsmdecom2/Z7 01
/fsmdecom2/ST1 001
/fsmdecom2/ST2 010
/fsmdecom2/ST4 100
/fsmdecom2/ST5 101
/fsmdecom2/ST6 110
/fsmdecom2/ST7 111
/fsmdecom2/p_... 001
/fsmdecom2/n_... 001
Now 200 ns

Fig. 6. State transition in Lower half subFSM

2.2 Observations

Power consumption in original FSM is calculated as 5.39mW.
Power consumption in decomposed sub FSM is 3.47mW & 3.64mW.
On taking the average, 3.55mW.
Percentage in power saving[3] $= \dfrac{(P_{\text{originalFSM}} - P_{\text{subFSM}})}{P_{\text{originalFSM}}}$

$= 34.13\%$

3 Architecture of FSM

There is one control signal, control_1 , which is the output of the first flip-flop; one 1x2 decoder which will generate the enable signals e1 and e2 for submachine M1 and M2 ; four AND gates A, B, C, D in front of M1 and four AND gates E,F,G,H in front of M2 which will block the state and primary input signals from propagating through M1 and M2 , respectively;three multiplexers mux1 , mux2 , mux3 which will determine whether the next state registers will be loaded from M1 or M2 ; and two multiplexers mux4 , mux5 which will determine the correct output signals[3]. Suppose submachine M1 is active and is in state s6 . Since the first bit of the state code (the control signal control_1)of s6 , which is the input to the 1x2 decoder, is 0, the output signal from the decoder to Com1 , e1 , is 1 which will turn on all the AND gates A, B, C, D in front of submachine M1 . Thus, all signals fed to the circuit corresponding to submachine M1 will propagate through. However, the output signal from the decoder to Com2 , e2 ,is 0 which will turn off all the AND gates D, E, F, G in front of submachine M2 . Thus, all signals fed to the circuit corresponding to submachine M2 will be blocked. Similarly, with the control signal control_1 being 0, signals for the next state flip-flops and output will come from Com1 through the, F, G, H in front of M2 which will block the state and 2x1 multiplexers. Now, if the input is 0, submachine M1 will transit to state START. Since the first bit of the state code of START is also 0, submachine M1 will stay active in the next clock cycle. On the other hand, if the input is 1, submachine M1 will transit to state s2 . Since the first bit of the state code (the control signal control_1)ofs2 is 1, the output signal of the decoder e2 will be 1 which will allow inputs to propagate through the AND gates E, F, G, H and Com2 , and thus turn on the submachine M2 .In the meantime, the value

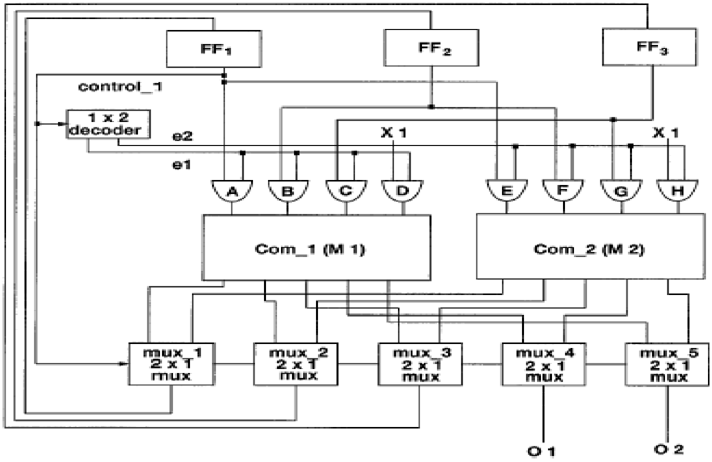


Fig. 7. Block diagram of low power architecture

of e1 is 0 which will inhibit the propagation of inputs through the AND gates A, B, C, D, and thus set all the inputs to submachine M1 to 0s in the next clock cycle, rendering submachine M1 inactive. If in subsequent clock cycles state transitions are confined to within submachine M2 , the value of e1 will remain 0 and submachine M1 will remain inactive.

3.1 Simulation at Architectural Level

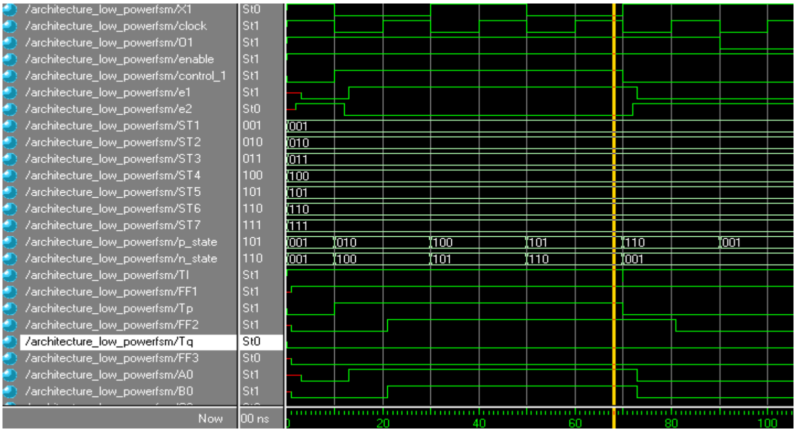

/architecture_low_powerfsm/X1 St0
/architecture_low_powerfsm/clock St1
/architecture_low_powerfsm/O1 St1
/architecture_low_powerfsm/enable St1
/architecture_low_powerfsm/control_1 St1
/architecture_low_powerfsm/e1 St1
/architecture_low_powerfsm/e2 St0
/architecture_low_powerfsm/ST1 001
/architecture_low_powerfsm/ST2 010
/architecture_low_powerfsm/ST3 011
/architecture_low_powerfsm/ST4 100
/architecture_low_powerfsm/ST5 101
/architecture_low_powerfsm/ST6 110
/architecture_low_powerfsm/ST7 111
/architecture_low_powerfsm/p_state 101
/architecture_low_powerfsm/n_state 110
/architecture_low_powerfsm/Tl St1
/architecture_low_powerfsm/FF1 St1
/architecture_low_powerfsm/Tp St1
/architecture_low_powerfsm/FF2 St1
/architecture_low_powerfsm/Tq St0
/architecture_low_powerfsm/FF3 St0
/architecture_low_powerfsm/A0 St1
/architecture_low_powerfsm/B0 St1
Now
00 ns
001
010
011
100
101
110
111
001 010 100 101 110 001
001 100 101 110 001
20
40
60
80
100

Fig. 8.

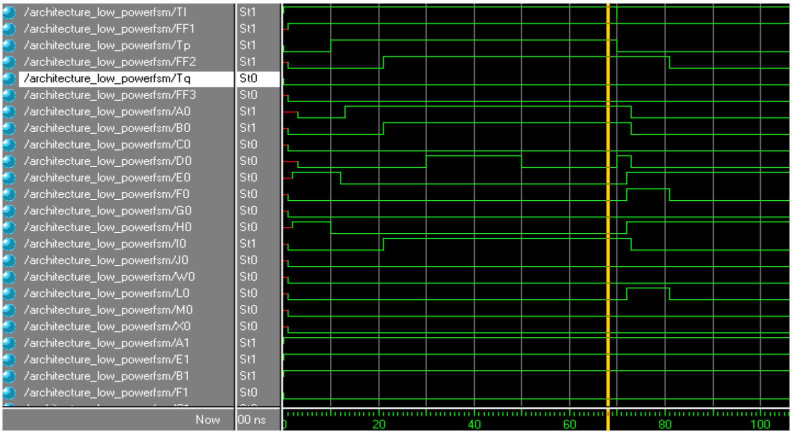
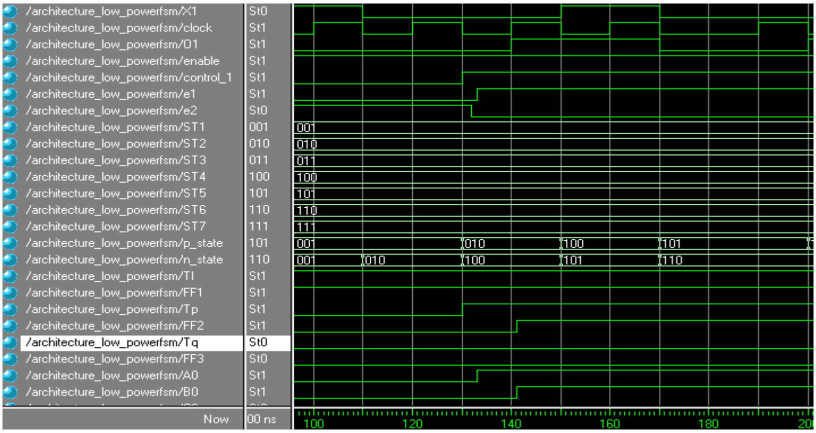
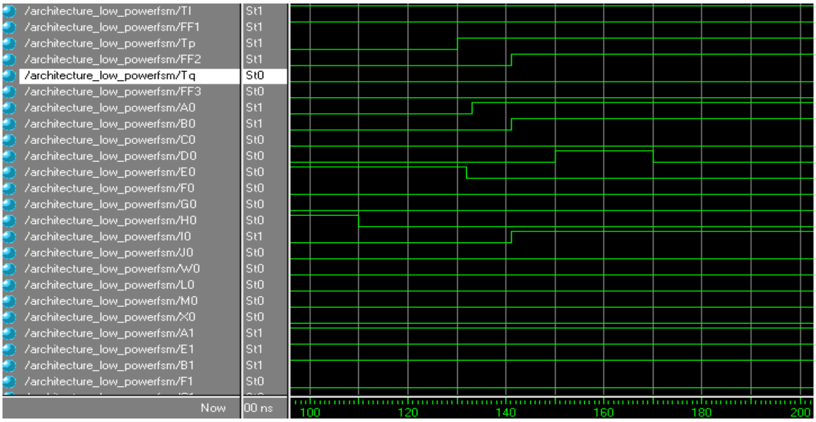

/architecture_low_powerfsm/Tl St1
/architecture_low_powerfsm/FF1 St1
/architecture_low_powerfsm/Tp St1
/architecture_low_powerfsm/FF2 St1
/architecture_low_powerfsm/Tq St0
/architecture_low_powerfsm/FF3 St0
/architecture_low_powerfsm/A0 St1
/architecture_low_powerfsm/B0 St1
/architecture_low_powerfsm/C0 St0
/architecture_low_powerfsm/D0 St0
/architecture_low_powerfsm/E0 St0
/architecture_low_powerfsm/F0 St0
/architecture_low_powerfsm/G0 St0
/architecture_low_powerfsm/H0 St0
/architecture_low_powerfsm/I0 St1
/architecture_low_powerfsm/J0 St0
/architecture_low_powerfsm/W0 St0
/architecture_low_powerfsm/L0 St0
/architecture_low_powerfsm/M0 St0
/architecture_low_powerfsm/X0 St0
/architecture_low_powerfsm/A1 St1
/architecture_low_powerfsm/E1 St1
/architecture_low_powerfsm/B1 St1
/architecture_low_powerfsm/F1 St0
Now
00 ns
20
40
60
80
100

Fig. 9.



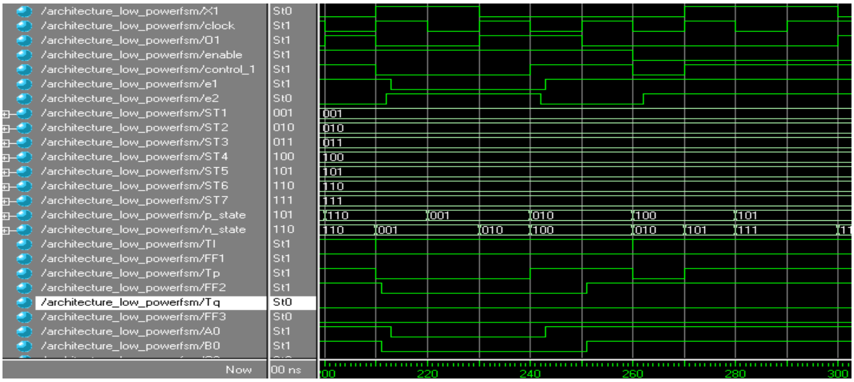
```
/architecture_low_powerfsm/X1         St0
/architecture_low_powerfsm/clock      St1
/architecture_low_powerfsm/O1         St1
/architecture_low_powerfsm/enable     St1
/architecture_low_powerfsm/control_1  St1
/architecture_low_powerfsm/e1         St1
/architecture_low_powerfsm/e2         St0
/architecture_low_powerfsm/ST1        001
/architecture_low_powerfsm/ST2        010
/architecture_low_powerfsm/ST3        011
/architecture_low_powerfsm/ST4        100
/architecture_low_powerfsm/ST5        101
/architecture_low_powerfsm/ST6        110
/architecture_low_powerfsm/ST7        111
/architecture_low_powerfsm/p_state    101
/architecture_low_powerfsm/n_state    110
/architecture_low_powerfsm/Tl         St1
/architecture_low_powerfsm/FF1        St1
/architecture_low_powerfsm/Tp         St1
/architecture_low_powerfsm/FF2        St1
/architecture_low_powerfsm/Tq         St0
/architecture_low_powerfsm/FF3        St0
/architecture_low_powerfsm/A0         St1
/architecture_low_powerfsm/B0         St1
Now                                   00 ns
```

Fig. 10.



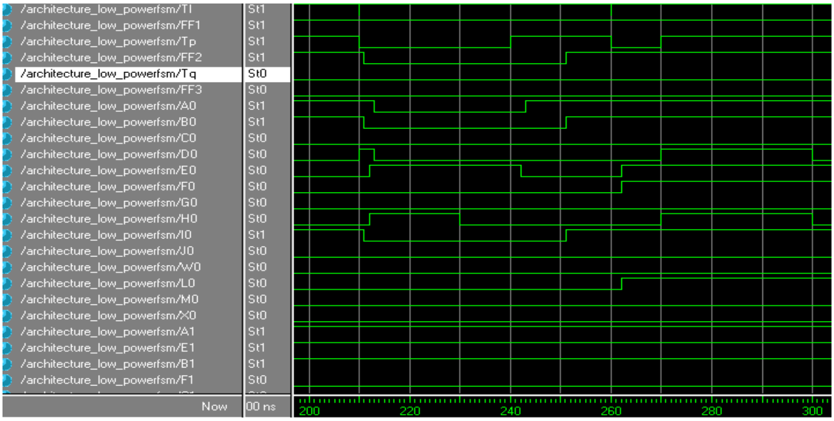
```
/architecture_low_powerfsm/Tl   St1
/architecture_low_powerfsm/FF1  St1
/architecture_low_powerfsm/Tp   St1
/architecture_low_powerfsm/FF2  St1
/architecture_low_powerfsm/Tq   St0
/architecture_low_powerfsm/FF3  St0
/architecture_low_powerfsm/A0   St1
/architecture_low_powerfsm/B0   St1
/architecture_low_powerfsm/C0   St0
/architecture_low_powerfsm/D0   St0
/architecture_low_powerfsm/E0   St0
/architecture_low_powerfsm/F0   St0
/architecture_low_powerfsm/G0   St0
/architecture_low_powerfsm/H0   St0
/architecture_low_powerfsm/I0   St1
/architecture_low_powerfsm/J0   St0
/architecture_low_powerfsm/W0   St0
/architecture_low_powerfsm/L0   St0
/architecture_low_powerfsm/M0   St0
/architecture_low_powerfsm/X0   St0
/architecture_low_powerfsm/A1   St1
/architecture_low_powerfsm/E1   St1
/architecture_low_powerfsm/B1   St1
/architecture_low_powerfsm/F1   St0
Now                             00 ns
```

Fig. 11.



```
/architecture_low_powerfsm/X1         St0
/architecture_low_powerfsm/clock      St1
/architecture_low_powerfsm/O1         St1
/architecture_low_powerfsm/enable     St1
/architecture_low_powerfsm/control_1  St1
/architecture_low_powerfsm/e1         St1
/architecture_low_powerfsm/e2         St0
/architecture_low_powerfsm/ST1        001
/architecture_low_powerfsm/ST2        010
/architecture_low_powerfsm/ST3        011
/architecture_low_powerfsm/ST4        100
/architecture_low_powerfsm/ST5        101
/architecture_low_powerfsm/ST6        110
/architecture_low_powerfsm/ST7        111
/architecture_low_powerfsm/p_state    101
/architecture_low_powerfsm/n_state    110
/architecture_low_powerfsm/Tl         St1
/architecture_low_powerfsm/FF1        St1
/architecture_low_powerfsm/Tp         St1
/architecture_low_powerfsm/FF2        St1
/architecture_low_powerfsm/Tq         St0
/architecture_low_powerfsm/FF3        St0
/architecture_low_powerfsm/A0         St1
/architecture_low_powerfsm/B0         St1
Now                                   00 ns
```

Fig. 12.



```
/architecture_low_powerfsm/TI    St1
/architecture_low_powerfsm/FF1   St1
/architecture_low_powerfsm/Tp    St1
/architecture_low_powerfsm/FF2   St1
/architecture_low_powerfsm/Tq    St0
/architecture_low_powerfsm/FF3   St0
/architecture_low_powerfsm/A0    St1
/architecture_low_powerfsm/B0    St1
/architecture_low_powerfsm/C0    St0
/architecture_low_powerfsm/D0    St0
/architecture_low_powerfsm/E0    St0
/architecture_low_powerfsm/F0    St0
/architecture_low_powerfsm/G0    St0
/architecture_low_powerfsm/H0    St0
/architecture_low_powerfsm/I0    St1
/architecture_low_powerfsm/J0    St0
/architecture_low_powerfsm/W0    St0
/architecture_low_powerfsm/L0    St0
/architecture_low_powerfsm/M0    St0
/architecture_low_powerfsm/X0    St0
/architecture_low_powerfsm/A1    St1
/architecture_low_powerfsm/E1    St1
/architecture_low_powerfsm/B1    St1
/architecture_low_powerfsm/F1    St0
                           Now   00 ns
```

Fig. 13.

Above are the simulated output waveform at an architecture level .

4 Conclusion

Methodology for the decomposition of finite state machines targeted towards low power dissipation is presented. Heuristic algorithm approach gives good decomposition techniques to obtain a state machine at, for the majority of the large examples tested, exhibits a much smaller power dissipation than the original. The results show that savings of up to 33% are possible in some of the examples.Despite the good quality of the results obtained, there are several rections for future work that may improve these results and extend e **range** of applicability of the technique.The single most important direction is probably the extension of is work to the case where the state transition graph cannot be explicitly described. In our experiments, we verified that the bottleneck sides in the extraction of the STG from the sequential circuit description.Because the decomposition algorithm works with an explicit description of the STG, it is not possible to apply the method to machines where the STG cannot even be extracted. However, it may be possible to apply the decomposition idea even in this case, if a set of transitions with high enough probability exist to justify the method. The basic idea is that, in many cases, this set of transitions can be identified using Monte Carlo methods even without doing a complete traversal of the STG. With this set of transitions available, it is possible to perform a partition of the STG, considering only the states involved in these transitions and using only an implicit representation of the STG. Once the STG is partitioned, the rest of the method is directly applicable, making it feasible to use it in cases where the machine is too large to permit the extraction of the STG. Another interesting direction for future research is on the automatic selection of the size of the ***small*** machine. Although, in some cases, significant gains can be obtained with a variety of sizes for this machine, in other cases the result depends strongly on the adequate selection of the value of this parameter. It may be possible to modify the cost function described above to automatically include a term that depends on the number of states, thereby removing the need for the user to specify the value of this parameter or to perform a search for its right value. Finally, it is clear that the results obtained in this paper can be improved if a more efficient implementation of the decomposition strategy is selected. In particular, it should be possible to reduce the overhead incurred by the addition of the extra outputs to each sub-FSM,by encoding these outputs in a different way that take into account the encoding of each of the sub-machines.

References

[1] Alidina, M., Monteiro, J., Devadas, S., Ghosh, A., Papaefthymiou, M.: Precomputation-based sequential logic optimization for low power. IEEE Trans. VLSI Syst. 2, 426–436 (1994)

[2] Benini, L., De Micheli, G., Lioy, A., Macii, E., Odasso, G., Poncino, M.: Computational kernels and their application to sequential power optimization. In: Proc. 35th Design Automation Conf., pp. 764–769 (June 1998)

[3] Benini, L., Siegel, P., De Micheli, G.: Automatic synthesis of lowpower gated-clock finite-state machines. IEEE Trans. Computer-Aided Design 15, 630–643 (1996)

[4] Benini, L., Vuillod, P., Coelho, C., De Micheli, G.: Synthesis of lowpower partially-clocked systems from high-level specifications. Presented at the 9th Int. Symp. System Synthesis (November 1996)

[5] Chow, S.-H., Ho, Y.-C., Hwang, T.: Low power realization of finite state machines—A decomposition approach. ACM Trans. Design Automat. Electron. Syst. 1(3), 315–340 (1996)

[6] Devadas, S., Newton, A.: Decomposition and factorization of sequential finite state machines. IEEE Trans. Computer-Aided Design 8, 1206–1217 (1989)

[7] Hachtel, G., Hermida, M., Pardo, A., Poncino, M., Somenzi, F.: Reencoding sequential circuits to reduce power dissipation. In: Proc. Int. Conf. Computer-Aided Design, pp. 70–73 (November 1994)

[8] Kernighan, B., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell Syst. Tech. J., 291–307 (February 1970)

[9] Monteiro, J., Devadas, S., Ghosh, A.: Retiming sequential circuits for low power. In: Proc. Int. Conf. Computer-Aided Design, pp. 398–402 (November 1993)

[10] Monteiro, J., Oliveira, A.: Finite state machine decomposition for low power. In: Proc. ACM/IEEE Design Automation Conf., pp. 758–763 (June 1998)

Pruning Search Spaces of RATA Model for the Job-Shop Scheduling

Farid Arfi[1], Jean-Michel Ilié[2], and Djamel-Eddine Saidouni[1]

[1] Computer Science Department, University of Mentouri,
MISC Laboratory, Constantine, 25000, Algeria
[2] Computer Science Department, Paris 6 University,
LIP6 Laboratory, Paris, 75005, France
arfi_f@hotmail.com, jean-michel.ilie@lip6.fr
saidouni@misc-umc.org

Abstract. In this paper, we propose a pruning method in order to reduce the search space for the job-shop scheduling problem with makespan minimization. In RATA model each trace corresponds to a feasible schedule, so we apply this method to the reachability algorithm of RATA model that explores the space of all possible schedules. We conducted an experimental study over a set of benchmarks. The results show that the proposed method is able to reduce both the space and the time in searching for optimal schedules.

Keywords: Scheduling, reachability analysis, timed model, job-shop.

1 Introduction

The job-shop scheduling problem is a paradigm of optimization and constraint satisfaction problems for distributed systems referenced in many researches over the last years. Traditionally, the optimization criterion is the so-called makespan minimization, which requires to count the time spent to perform the actions of the job-shop. Job-shop problems require the expressions of concurrent and parallel behaviors [1][2][3].

In this paper, we propose to capture job-shop scheduling problems from a very intuitive and compact description model, called Resource Allocation Timed Automata (RATA). This model inherits from the DATA model which introduces true concurrency semantics to deal with concurrent events [4]. Extensions are provided to explicitly represent the resource requirements needed for scheduling analysis. In this model, the parallelism is implicitly expressed from the starting events of actions (i.e. once started, the actions are assumed to behave in parallel until their terminations). This avoids splitting the description of running actions in start and end events, as this is proposed in the Timed Automata models (TA) dedicated to scheduling problems, e.g. [5]. As another interest, the RATA reachability graphs are generally much smaller than the TA ones, e.g. [6]. However, both suffer from the well-known combinatorial explosion problem.

A standard way to attack this explosion problem consists in restraining the execution of actions by focusing on the immediate runs, to study the scheduling

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 387–401, 2013.

problems. However in practice, other reduction techniques must be exploited. The main contribution of this paper consists in proposing a set of search space reduction techniques adapted to the RATA models and that can be combined to gain more efficiency. In the last decade, several advances were investigated, from different representations including the TA models. In particular, partial order techniques such as stubborn sets can take advantage of the independency of some executions of actions, in order to reduce the reachability graph to consider [7]. In [8], another partial order technique, namely the sleep sets, is combined with the so-called laziness reduction technique. The aim is avoiding the generation of some lazy runs, featured by configurations from which a bad exploitation of the machines can be detected. In this paper, we propose an improved version of both the stubborn set and laziness reduction techniques, and we propose to combine them since sleep sets are known to only reduce transitions but preserve useless states [9].

Moreover, in another proposition [3], the laziness techniques were replaced by the so-called domination test. This test aims at defining relations between the configurations to be explored within the search space. It is used to suppress the lazy runs and other bad configurations but also is used to replace equivalent sets of configurations by representatives. Observe that the domination test performs comparisons over the set of computed configurations, therefore the laziness reduction technique maintains its interest, since only locally applied from each considered configuration. In this paper, an improvement is proposed to better establish the dominance property between the configurations.

We also use a last reduction technique based on an estimation of the remaining time to be spent in order to achieve the runs from some configuration. Actually, several variants and improvements exist in the literature, e.g. [10].

The remaining of the paper is organized as follows. In section 2, the RATA model is presented and a job-shop use case is described using this model. Section 3 brings out our reachability algorithm dedicated to makespan minimizations. Search space reduction techniques are proposed in Section 4. In Section 5, experimental studies are presented to highlight the efficiency of our approach. This includes some comparisons against an extension of TA [2]. Section 6 presents our conclusion and perspectives.

2 RATA Model

The RATA model re-uses DATA concepts, in particular the non-atomicity of actions is captured by the fact that each transition only corresponds to a start of an action. From state to state, one or several independent actions can be launched, therefore, each state could be associated with a set of launched actions. In the model, each of these launched actions are represented by means of a distinct clock, dynamically created and initialized to 0 at the transition which starts the action. A set of temporal constraints is also associated with each state, expressing the conditions of ends concerning the launched actions in the state. As an example, consider the DATA of Figure *1.a*, modeling a system *S* where two actions, *a* and *b*, able to run concurrently. A distinct clock is assigned to these actions, *x* and *y* respectively. Starting from the initial state s_0, there are two possible transitions: $s_0 \xrightarrow{a,x} s_1$ and $s_0 \xrightarrow{b,y} s_2$. A label *(a,x)* attached to a transition indicates that the action *a* has just been launched, and that the clock *x* will give the time spent since the launching of *a*. Similarly from the reached states, the following two transitions $s_1 \xrightarrow{b,y} s_3$ and $s_2 \xrightarrow{a,x} s_3$ are possible.

In the initial state s_0, the set of temporal constraints is empty because none of the actions is running in this state. In s_1, $\{x \geq 10\}$ specifies that the action *a* finishes its execution as soon as *x* reaches *10*. Similarly, s_2 is labeled by $\{y \geq 12\}$. In s_3, the actions *a* and *b* can continue their runs in parallel, and each one can finish only if its proper clock reaches a value equal to its duration, so the associated set of temporal constraints is $\{x \geq 10, y \geq 12\}$.

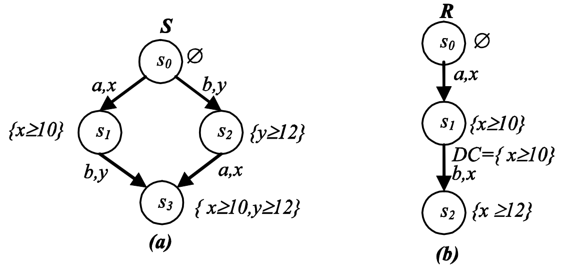


Fig. 1. Behaviors of two systems in terms of DATA

The precedence relation between actions implies to annotate each transition with some additional guard, namely Duration Condition *(DC)*. This guard on a transition expresses that the new launched action is possible, provided the preceding launched ones have been terminated. In the DATA model, this is formally expressed as a subset of the set of temporal constraints attached to the source state of the transition. Consider for instance the system *R* wherein the action *a* must be followed by the action *b*. The behavior of *R* is shown in Figure *1.b*. Since at most one action can run at a certain point, the same clock *x* can be assigned to both actions *a* and *b*. From the initial state, the unique transition expresses that *a* can be launched without any duration constraint, hence $DC=\varnothing$ for this transition (not represented in the figure since empty). From the state s_1 where the temporal constraints is $\{x \geq 10\}$ for *a*, the action *b* can obviously be run only if the action *a* finishes its execution. This condition is expressed by the set $DC=\{x \geq 10\}$ attached to the transition which launches the action *b*. So, *b* can start at any time within the enabling open interval $x \in [10,+\infty[$.

2.1 Intuition of RATA Model

The RATA model is an extension of the DATA one, assuming an execution platform of (*M*) machines. An action is executed on a predetermined machine, inducing a duration for its execution. Since a machine cannot be allocated to several actions at the same time, a mutual exclusion mechanism must hold constraining the execution of actions.

Let us consider the system *S* again, but assume that the execution platform is either P_1 or P_2. The first one contains two machines m_1 and m_2, used for executing the actions *a* and *b* respectively, whereas the second contains a single machine *m* used for executing any action. The corresponding behaviors for *S* are represented by the RATA of Figures 2.*a* and 2.*b*, respectively.

It appears that DATA and RATA have the same structure, however the duration of the launched actions are now expressed by using the function τ such that $\tau(a,m)$ yields the duration of any action *a* executed on a machine *m*. In addition to *DC*, each transition will be labeled by another guard, namely Availability Condition *(AC)*, expressing the mutual exclusion constraints on shared machines. As for *DC*, The condition *AC* for a transition is a subset of the temporal constraints of the source state, however concern the ones related to the machine of the transition. *AC* is not displayed when empty.

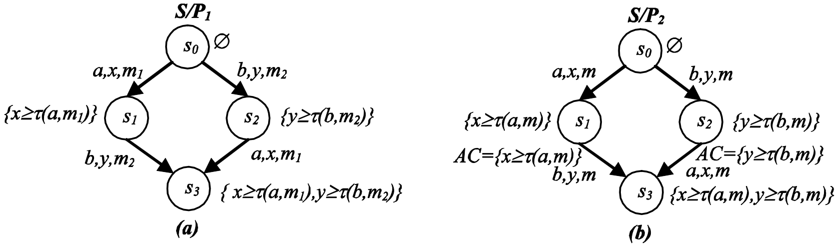


Fig. 2. System *S* executed on platforms P_1 and P_2

In Figure 2.*a*, all the transitions have an empty *AC* since there is no shared machine on P_1. In Figure 2.*b*, the transition starting from the state s_1 is labeled by $AC=\{x\geq\tau(a,m)\}$. Indeed, the temporal constraint $x\geq\tau(a,m)$ in the source state relates to the shared machine *m* of the platform P_2. So in the state s_1, the action *a* is possibly in execution on *m* and the launching of the action *b* is enabled only if the temporal constraint $x\geq\tau(a,m)$ holds (i.e. the machine *m* has finished the execution of *a* and can start *b*). Observe that similar reason implies the label of the transition starting from s_2. Observe that the state s_3 of Figure 2.*b* represents different situations of execution where at most one action can be running in s_3, with regard to the set of temporal constraints associated with s_3.

Further, *DC* and *AC* sets are removed from the figures since they can be easily deduced from the clocks and machines used in the labels of transitions, together with the information of the temporal constraints of the source and target states.

2.2 Formalization

Definition 1: Let $H=\{x,y...\}$ be a set of clocks whose values are defined in a time domain R^+ and M a set of machines. The set $\Phi(H)$ of temporal constraints γ over *H* is defined by the syntax $\gamma::=x\geq t$, where *t* is a duration value. Durations are expressed by the duration function $\tau:A\times M\rightarrow N$ s.t. $\tau(a,m)$ represents the duration of action *a* of *A* (*A* the set of actions), running on a machine *m* of *M* (*M* the set of machines). Given *F* a set of constraints, its subset F_x and F^m respectively represent the constraint to the clock *x* and the different constraints related to the machine *m*.

A valuation *v* (of the clocks) of *H* is a mapping which assigns each clock of *H* to a value in R^+. The set of all valuations for *H* is denoted $\Xi(H)$. A valuation $v\in\Xi(H)$

satisfies a temporal constraint $\gamma=(x \geq t)$ with $x \in H$, which is denoted $v|=x \geq t$, iff $v(x) \geq t$. Further, this satisfaction is linearly extended to deal with sets of temporal constraints. W.r.t. $x \in H$, $[x \rightarrow 0]v$ denotes the valuation of H which assigns the value 0 to the clock x and accords with v concerning the clocks of $H \backslash \{x\}$.

Definition 2: A RATA model RM is a tuple (S, s_0, H, M, L, T) where:

- S is a finite set of states, $s_0 \in S$ is the initial state,
- H and M are respectively the finite set of clocks and the finite set of machines,
- L: $S \rightarrow 2^{\Phi(H)}$ is a mapping that associated with each state s, a set of temporal constraints $F=L(s)$, representing the set of actions possibly in execution in s,
- $T \subseteq S \times A \times H \times M \times S$ is the set of transitions. A transition (s,a,x,m,s') also denoted $s \xrightarrow{a,x,m} s'$ represents a change from the state s to the state s', involving to start the action a on the machine m and define a clock x initialized to 0 to be associated with the action a.

Definition 3: W.r.t. a transition $(s,a,x,m,s') \in \mathrm{T}$, the sets DC and AC of constraints are defined by: $DC=L(s) \backslash (L(s') \backslash L(s')_x)$ and $AC= L(s)^m$.

The first equation is deduced from $L(s')= (L(s) \backslash DC) \cup \{x \geq \tau(a,m)\}$, where DC can be regarded as the precedence constraints defined over the actions of a job. For the first action, DC is empty, otherwise it is reduced to a singleton which relates to the preceding action of a within the same job. The cardinality of AC can be larger than one, since there may be in s several actions which share the same machine.

The launching of a transition (s, a, x, m, s') from a given valuation v associated with s, is constrained by the following two conditions:

- $v \models DC$. The specification of a system directly corresponds to the properties of precedence over the action executions.
- $v \mid = AC$. Thus, any action executed in s on a machine m must be completed to allow the firing of a transition which refers to the same machine.

Definition 4: The semantics of a RATA $RM =(S, s_0, H, M, L, T)$ is defined by associating with RM, an infinite transition system SA on the alphabet $A \cup R^+$. A state of SA, also called a configuration, is a pair $<s,v>$ where s is a state of RM and v a clock valuation for H. A configuration $<s_0,v_0>$ is initial iff s_0 is initial in RM and $\forall x \in H$, $v_0(x)=0$. The two following rules express that two types of transitions can link the SA configurations, corresponding to an elapsing of time (RA) and an execution of an action of A (RD), respectively :

$$\frac{d \in \mathrm{R}^+}{\langle s,v \rangle \overset{d}{\rightarrow} \langle s,v+d \rangle}(RA) \qquad \frac{(s,a,x,m,s') \in T \quad v \models AC \cup DC}{\langle s,v \rangle \overset{a}{\rightarrow} \langle s',[x \mapsto 0]v \rangle}(RD)$$

According to the model semantics, the label a in the RD rule implies the start of an action a and not the whole execution of a. This rule can be applied only in case both sets DC and AC are satisfied. Otherwise, the time step rule RA is applied.

By applying the above rules from the initial configurations, we are able to compute the set of reachable configurations. Further, a run is a path of reachable configurations, by application of the two former rules. A possible run denoted $(s_0,v_0)\xrightarrow{d}(s_0,v_1)\xrightarrow{a}(s_1,v_2)$, where d represents the time spent in (s_0,v_0) and a the action to be started from (s_0,v_1), first induces that $v_1=v_0+d$, moreover there are a machine m and a clock x such that (s_0,a,x,m,s_1) is a transition of the RATA and $v_2=[x\to 0]v1$.

2.3 Modeling the Job-Shop with RATA Model

In this paper, the job-shop model is obtained compositionally. First, the sequential semantics for each job is parsed, yielding a RATA model for each job. Then, a standard parallel composition is used to obtain the RATA model of the whole system. The reader can find more details in [11] about the modeling approach and the formal definition of job-shop problem. We restrict our presentation to an example of this problem and its resulting RATA model.

Consider a job-shop system R sharing a set of machines $M=\{m_1,m_2\}$, knowing that each machine performs at most one action at a time, without capacity of preemption. Further, we consider the two following jobs: $j_1=a\prec b$ and $j_2=c$, where $\prec$ represents the precedence relation concerning the execution of actions. The machine allocated to the actions are~:$\mu(a)=\mu(c)=m_1$, $\mu(b)=m_2$, and the duration of the action execution over the shared machines follows~: $\tau(a,m_1)=4$, $\tau(b,m_2)=5$ and $\tau(c,m_1)=3$. The behavior of these jobs is concisely represented by the RATAs of Figures $3.(j_1)$ and $3.(j_2)$ and the resulting composition by Figure $3.(j_1|||j_2)$.

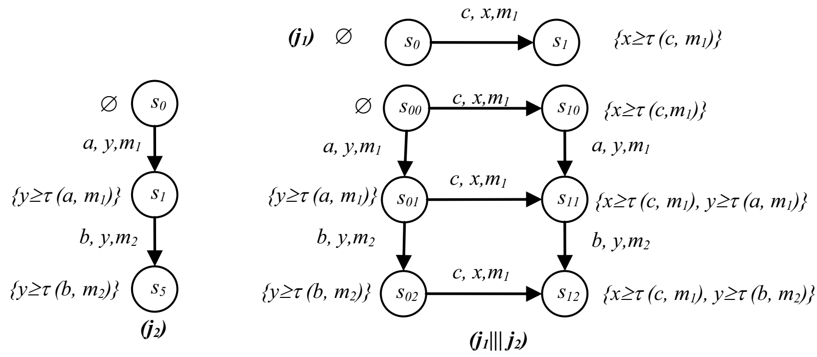


Fig. 3. RATAs of the jobs j_1, j_2 and $j_1|||j_2$.

It is worth noting that the labels of states and transitions in the model $(j_1|||j_2)$ still allow an evaluation of the constraints AC and DC. For instance, the transition from s_{10}, which can violate the mutual exclusion w.r.t. the machine m_1 is prohibited until the action c is considered as terminated in this state (The AC condition specifies that the constraint relative to m_1 in the source state must hold). Observe finally that there are only 6 nodes and 7 edges in the $(j_1|||j_2)$ model, whereas the same specification according to the standard TA approaches involves 14 nodes and 18 arcs [2].

3 Reachability Analysis Algorithm

Starting from the initial configuration of a RATA model, a run is complete if it leads to a final configuration, the potentially running actions of which are considered as terminated. From every *complete run*, say C_R, a schedule can be straightforwardly derived, associating with each action *a*, the starting time *st(a)* of the transition labeled by *a* in C_R. Actually, the length of the schedule coincides with the metric length of C_R. In order to compute the length of a run and the start times of the actions, the considered RATA is augmented by an additional clock to measure the elapsing time spent from the beginning of a run, therefore this clock is never reset to zero. Further, its valuation is denoted t_A. A configuration *(s,v)* of the RATA model is reachable within the time t_A iff (s,v,t_A) is reachable in the augmented RATA. Our objective is to perform a *makespan minimization* of job-shop problem that is to determine the minimal time schedule where all the actions are completed. The basic algorithm for this problem is presented below.

Algorithm I (A minimal-time Reachability algorithm)

```
$W \leftarrow \{(s,v0,0)\}$ ; $P \leftarrow \varnothing$; $Best \leftarrow \infty$
while ($W \neq \varnothing$) do
  $(s,v,t_A) \leftarrow selectRemove(W)$
  if ($(s,v,t_A) \not\in_d P$) then
     $P \leftarrow P \cup \{(s,v,t_A)\}$
     if ($E(s,v,t_A) < Best$) then
        $S \leftarrow \{(s,v',t_{A'}) \mid (s,v,t_A) \rightarrow (s',v',t_{A'}) \wedge \rightarrow \not\in reduce(s)\}$
        if $S = \varnothing$ then
           $Best \leftarrow E(s,v,t_A)$
        else
           $W \leftarrow W \cup_d S$
       end if
     end if
  end if
end while
Return Best
```

The reachability analysis is realized on-the-fly during the building of the RATA model. This avoids an overall construction in case the optimal solution is rapidly discovered. The algorithm operates on the configurations of the reachability graph, the information of which are mainly the location, the clock valuations and an accumulated global time. The list of configurations that must be explored is represented by an ordered set *W*, called the waiting list. The waiting list *W* initially contains the initial configuration $(s,v_0,0)$, which in our case corresponds to the state where no job are started, therefore clocks are initialized to zero. The set *P* is the passed list that means the list of already explored configurations, which is normally empty at the start of the search algorithm. The global variable *Best* holds the time of the best path found so far. (s,v,t_A) represents the configuration that is currently explored in the algorithm. From some configuration (s,v,t_A), *S* is the sub-set of immediate successors that be reached by a single transition and which does not belong to the reduced transitions of this configuration. In case the set *S* is empty, the configuration *(s,v,t_A)* is a final state, thus can be used to update the *Best* value. The function selectRemove(*W*) selects and removes from *W*, a configuration which is minimal according to the ordering $\leq_{ord}$ defined by the tree search strategy. In order to avoid visiting and exploring the same configuration repeatedly, the operations $\not\in_d$ and $\cup_d$ are performed with a respect to some specific property, called dominance property. The function reduce (*S*) is used to remove those configurations that cannot contribute to a better path or to the optimal run. The test *$E(s,v,t_A)<Best$* compares the estimation time value of the current configuration to the best time value found. This avoids the explorations of non-optimal runs if the comparison fails. In a final configuration *(s,v,t_A)*, *$E(s,v,t_A)$* yields the exact value of the global (reachability) time for the configuration, here t_A. In the next section, we discuss some decision criteria concerning the function reduce, the dominance property and the global time estimation.

4 Search-Space Reduction Techniques

We propose various search-space reduction techniques that can be used to efficiently prune useless configurations in the reachability graph. They are divided in 2 great classes, the ones that can be realized locally to the considered configuration, then before the generation of its successors, and the second ones which require comparisons between the generated successors and the existing configurations.

4.1 Reduction before the Generation of Configurations

Immediate Runs. Because the time spent on states is left unrestricted by the rule RA in Definition 4, it appears that each qualitative path in a RATA features an infinite number of runs. This can be corrected by only focusing on the restricted notion of immediate runs.

Definition 5: (*Immediate Run*) an immediate run is a run within which each transition is taken as soon as the firing conditions, AC and DC, are satisfied. A non-immediate run is defined as a run containing the fragment: (s,v)$\xrightarrow{t}$(s,v+t)$\xrightarrow{a}$ (s’,v’), where the transition taken at (s,v+t) is already enabled at (s,v+t’) with t’<t.

Every immediate run represents an immediate schedule. The two schedules S_1 and S_2 in Figure 4 respectively represent a non-immediate schedule and an immediate schedule, with regards to the example problem of the system *R* explained in Section *2.3*. In the non-immediate schedule S_1 , there are two unnecessary zones of waiting, one for the machine m_2 during the time period [4, 5] and one for m_1 during [10, 11]. In contrast, S_2 brings out a schedule where the waiting times are minimal, thus making the schedule immediate. Therefore in order to find an optimal schedule for a given path, the exploration can be restrained to the immediate runs. This restriction transforms the RATA semantics into a *discrete directed acyclic graph* of configurations, like in Figure 5.

Corollary (*Job-Shop Scheduling and RATA model*): The optimal job-shop scheduling problem can be reduced to the finding of the shortest immediate run within a RATA.

Let us consider the job-shop system in Figure $3.(j_1 ||| j_2)$. Starting from the initial configuration, the immediate runs are directly obtained from the paths of the RATA, by evaluating the satisfaction of the sets *DC* and *AC* in each reached configuration, in order to start the next actions as soon as possible (immediate execution). These evaluations require replacing each occurrence of the function τ by its corresponding value, in order to compare with the values taken by the clocks.

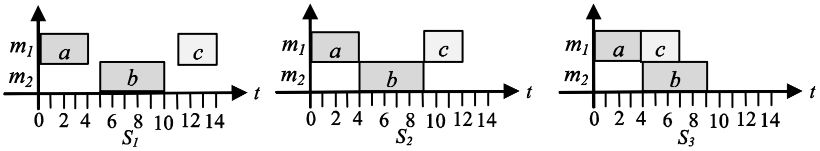


Fig. 4. Non-immediate, immediate but lazy and non-lazy schedules

Figure 5 shows the derivation tree obtained by the immediate runs of the system of Figure $3.(j_1 ||| j_2)$. The length of the optimal schedule is 9, which corresponds to the two left immediate runs represented in this figure. Moreover, each configuration is of the form $(s,v(x),v(y),t_A)$, where s represents a reachable state; $v(x)$ and $v(y)$ are respectively the valuations of the clocks x and y used in the configuration; t_A is the value of the additional clock.

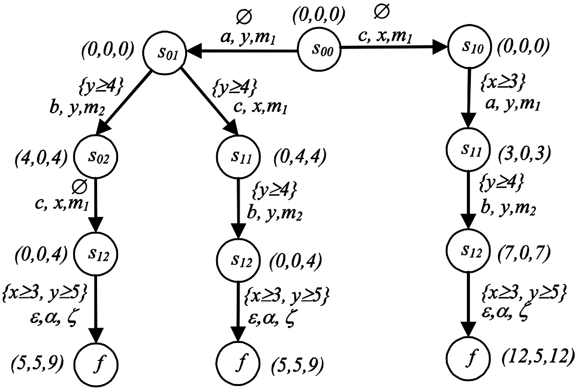


Fig. 5. The immediate runs of the RATA of figure $3.(j_1 ||| j_2)$

The evolution of the elapsing time in a configuration is not represented explicitly but is specified indirectly by the set of constraints attached the transitions issuing the state. Actually, this time depends on the transition to consider. W.r.t. some transition, it can progress from the global time value specified in the state, to a time value featuring that all the constraints attached to the transition are satisfied. For instance, with regard to the transition (s_{01}, s_{02}) and its constraints $\{y \geq 4\}$ the elapsing time in s_{01}

progresses from t_A = *0*, until the value 4 is reached. Since transitions are immediate, the global time value when reaching s_{02} is equal to 4 again.

Observe also that the last transition corresponds to a hidden action, clock and machine *(*$\varepsilon \notin A$*,* $\alpha \notin H$ and $\zeta \notin M$*)*, with an enabled condition used to terminate the execution of all of the actions not yet finished. So, the value of t_A in the final configuration of a run represents the total duration of the run, hence the length (time) of the schedule. The function reduce of Algorithm I mainly performs the generation of the finite number of immediate runs.

Stubborn Set Reduction. A valuable kind of space reduction is based on the checking of "stubborn sets" of transitions that are subset of the transitions which in some state does not influence the other transitions. Such a set can be fired whereas the others can be considered latter, in the next configuration. With regard to the job-shop scheduling problem, the stubborn set technique consists from some configuration to be explored, in selecting the transitions which not only have a minimal launching time but also correspond to the last use of a machine among the jobs. The selected transition can be immediately added to the trace, i.e. advanced w.r.t to the others. When several transitions are candidate to be selected, one is chosen arbitrarily [7].

For instance, consider the second configuration in Figure 5 (i.e. state s_{01}). Both transitions have the same (minimal) launching time *t=4* and both concern the last use of a machine (the used machines m_1 and m_2 remains unclaimed by any other action during the corresponding action processing time). As a consequence, the function reduce in Algorithm I selects one of them randomly in order to be launched.

We propose here an improvement consisting in weakening the former last use constraints~: we privilege transition, the machine of which remains unclaimed during the processing time of the considered action. As a consequence, the machine can be claimed after this time. Observe that this requires for each job to estimate the earliest starting time of the remaining actions which refers to the same machine. To privilege the transition, all the estimated values must be great than (or equal to) the ending time of the considered action.

Laziness Reduction. Although immediate runs are performed, there could remain lazy schedules. Laziness indicates suboptimal use of the resources, here the machines. Such run can produce suboptimal schedules.

A lazy run of a RATA model *RM* contains a sequence of states and transitions like $(s,v)\ldots \xrightarrow{t} \ldots (s',v') \xrightarrow{a} (s'',v'')$ wherein the transition a is enabled in *(s,v)*, but is taken after a certain delay in *(s',v')*.

In Figure 4, S_2 illustrates an immediate but lazy schedule, s.t. the machine m_1 is free at time 4 then could be used to perform the action *c*. Starting *c* after 3 time units more, introduces a "time hole" which is large enough to be filled with the action *c*. This should make the schedule suboptimal. Exploration of lazy runs can be prevented by a laziness reduction technique. In this special case, the lazy schedule S_2 of Figure 4 is already not expressed in the immediate runs of our model (see Figure 5). This is due to the used semantics, which combines parallel executions and immediate runs. An explicit interleaving of start and end events would make visible such lazy runs, as in the timed automata models [2], thus requires much effort in space and time to remove them.

In the algorithm I, the elimination of the lazy schedules is carried out by the function reduce in case the reduction given by stubborn set technique fails. From the considered configuration, it applies on every pairs of possible transitions; hence the possible successor configurations that lead to suboptimal solutions are not computed and inserted in the set *S*.

Definition 6. (*Lazyness Reduction*) Assume that two transitions labeled by the actions α and β are enabled in some configuration. Consider that τ(α,m) be the duration of the action α when running on the machine m and let enabl(α) represent the time where the action α is enabled (launched). The transition labeled by β is removed from the successors list S if the following sufficient condition holds~:

$$enabl(\alpha) + \tau(\alpha,m) \leq enabl(\beta) \qquad (L)$$

Consider the previous example again, but change the execution machine of the action *c* to m_3 .The lazy schedule *S* of this system represented in Figure 6 is obtained by the following immediate runs, also depicted in the graph (*G*) of the same figure~:

$$(s00,0,0,0) \xrightarrow{a} (s01,0,0,0) \xrightarrow{b} (s02,4,0,4)...$$

At s_{01}, the condition of the laziness reduction holds: $enabl(c)+ \tau(c,m) \leq enabl(b) \Rightarrow 0+3\leq 4$. The transition labeled by the action *b* can be pruned from s_{01}, so the only considered transition from this state is the one labeled by the action c launched at time 0 (non-lazy schedule).

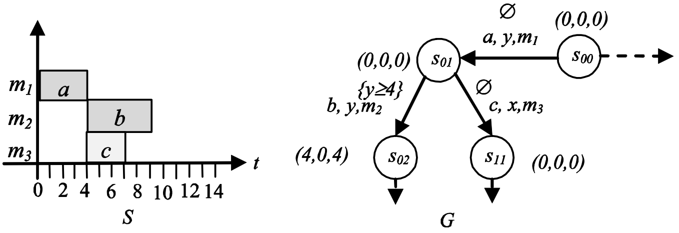


Fig. 6. Lazy schedule S and a sub-graph G of immediate runs

In case the execution time of *c* overpasses the starting of the action *b*, the former laziness technique fails. However, an improvement of Definition 6 is possible.

Laziness Reduction Improvement: Considering the problem specified in definition 6, and assumes *t* is the time interval from the moment where *β* is started to the moment where an action uses the execution machine dedicated to *α.* Then, the transition labeled by *β* is removed from S if~:

$$enabl(\alpha)\leq enabl(\beta) \;\wedge\; enabl(\alpha)+\tau(\alpha,m)\leq enabl(\beta)+t \qquad (L').$$

The first condition in *L'* is a direct consequence of the laziness characterization. The second condition suggests that it is possible to remove the action *β* if the execution machine of *α* remains unused from the configuration where *β* was considered. Therefore the execution of *β* and the other executions enabled in the future of *β* are preserved in their time after the starting of *α.*

As for the stubborn set technique, the non-use of a machine during some time period is estimated over the remaining executions of the jobs. If the machine is unclaimed for the remaining action of the jobs, then the value ∞ is assigned to *t*.

4.2 Reduction after the Generation of Configurations

As the reduction technique made before the generation of configurations are based on heuristics, it could remain some reduction to perform that we can detect after the generation of configurations. We propose another set of reduction techniques to be applied on configurations once generated.

Domination Test. This test is used to avoid exploring identical configurations or configurations that are obviously worse than already computed ones. The domination test is based on the following definition:

Definition 7 (D1): Let (s,v,t_A) and (s,v',$t_{A'}$) be any two reachable configurations. We say that (s,v,t_A) dominates (s,v',$t_{A'}$) if $t_A \le t_A' \wedge v \ge v'$.

The fact that *(v≥v')* implies that whatever the enabled transition, it will be launched in *(s,v,t_A)* before *(s,v',$t_{A'}$)* or at the same time. Moreover with *(t_A≤$t_{A'}$)* , we deduce that for every complete run reaching some final configuration *(s,v',$t_{A'}$)*, there is a run reaching *(s,v,t_A)* which leads to a better solution (i.e. with a lower execution time).

Whenever a new configuration is visited in the graph, we check whether it is dominated by an already computed one, and in this case it is discarded. Moreover, a dominated configuration in the waiting list is replaced by the dominated one. Observe that in the algorithm I, the operations $\cup_d$ and $\in_d$ respectively denote the union and membership relations between configurations, with respect of dominance property.

According to Figure 5, a dominance reduction is possible over the two configurations whose expression is *(s_{12},0,0,4)* because they have the same global time and the same clocks valuation. Therefore, consider only one, instead of both.

We now propose a finer dominance relation based on a weaker relation between the clocks used in the compared configurations, e.g. *(s,v,t_A)* and *(s,v',$t_{A'}$)*. For each clock *x*, we consider its duration denoted τ*(a,m)*.

Definition 8 (D2) : (s,v,t_A) dominates (s,v',t_A') if :

$$t_A \le t_A' \wedge \forall x \in H, (v(x) + (t_A' - t_A) \ge v'(x)) \vee (v(x) + (t_A' - t_A) \ge \tau(\mathrm{a, m})).$$

So, we admit that *v* could be less than *v'* for some clocks *x* in two cases.

- Either the clock difference *v'(x)-v(x)* is compensated by the value *t*= $t_{A'}$ -t_A which means that if the same sequence of transitions is fired from *(s,v,t_A)* and *(s',v',$t_{A'}$)*, reaching *(s_r,v_r,t_{Ar})* and *($s_{r'}$,$v_{r'}$,$t_{Ar'}$)* respectively, then the above dominance rules globally still hold for the reached configurations. In case where the reached states are final, we have $t_{Ar'} \ge t_{Ar}$.
- In the 2nd term of the or clause, the action a associated with *x* is terminated within the duration $t_{A'}$-t_A. In this case, the valuation *v(x)* must be excluded out of the dominance test since it cannot influence the future firing of transitions.

For sake of concision in this paper, the proof is not reported.

In Figure 5, we can now make a dominance reduction between the configurations *(s_{12},0,0,4)* and *(s_{12},7,0,7)*. Here, each configuration is of the form *(s,v(x),v(y),t_A)*, and the actions associated with the clocks *x* and *y* are *c* and *b*, with respective durations 3 and 5. Using previous conditions, we can deduce that the first configuration dominates the second.

Reduction Based on the Remaining Time. In the algorithm I, w.r.t. some configuration *(s,v,t_A)*, the estimation *$E(s,v,t_A)$* of the global execution time is given by the t_A value complemented by an heuristics on the remaining time to achieve the run. This estimation is compared to the Best known global time value, in order to delete bad configurations and also to search from the most promising configurations first. Clearly, this contributes to reduce the search space.

The main used heuristic in the literature simply computes a remaining time value, under the assumption that there is no conflict between the concerned machines. From some configuration, it is the maximum value between the remaining execution times of the machines, knowing that for a machine, this time corresponds to the sum of the durations of the actions that remain to be executed on the machine.

Recently, an improved heuristic was proposed based on a simple modification of the Jackson's preemptive schedule, which often yields a better estimation, that means greater remaining time values, closer from the exact values. For more details, See [10].

5 Experimental Results

To implement the makespan minimization, we have developed a C++ software tool based on the RATA model. Our tool embeds all the search space reduction techniques discussed in this paper: immediate runs, stubborn set, lazy reduction, domination test and heuristics of the remaining execution time. The searches in the graph use a combination of depth-first and best-first search strategies. The search strategies implemented in the function *selectRemove* of the algorithm I, decide which configuration will be chosen next, from the waiting list. The first criterion is to privilege the configuration which the minimum estimation value of the global execution time. The second criterion is the maximum depth of each configuration, evaluated in the number of actions that have been already executed. Thus, configurations close from a complete run can be privileged. In case of configurations having the same global time estimation and the same depth, the first one is chosen. The computational equipment for the experiments was a Pentium machine with 3 GHz and a Windows7 OS.

Three series of job-shop instances are investigated to demonstrate the performance of our algorithm. The first series (A) consists in randomly generating small instances of jobs having three operations for each job. The number of jobs varies from 2 to 6 in order to investigate the scalability of the proposed reductions. Table 1 shows a comparison between two techniques of reduction considered separately, namely the domination and the laziness techniques. This comparison is given on the derivation tree of immediate runs without considering the other reduction techniques. In Table 1, the left part indexed 1 concerns the classical use of the reduction techniques, whereas the right part indexed 2 corresponds to our proposed improvements. The columns *#laz1* and *#laz2* bring out the performances in terms of number of explored configurations. *T_laz1* and *T_laz2* highlight the processing time values. Similar notations are used for the domination reduction (dom). The number of generated configurations is limited to one million. Comparing the left and part, the improved versions appear to be better and better as the size of the job-shop problem augments.

In the second series (B), a comparison of the size of the RATA against the one of the model obtained by applying the approach of [2], proposed for TA, is given in

Table 1. Performances of the laziness/domination reductions techniques

#j	***RATA***							
	#laz1	*T_laz1*	*#dom1*	*T_dom1*	*#laz2*	*T_laz2*	*#dom2*	*T_dom2*
2	37	0	30	0	22	0	25	0
3	1520	0	305	0	156	*0*	122	*0*
4	62584	1.7	6133	0.4	2198	0.1	1043	0.1
5	/	/	233373	200	191025	*6.4*	10345	*1.2*
6	/	/	/	/	/	*/*	295084	*27*

Table 2. The column *#ds* informs on the number of discrete states for each model, where *#bf* brings out the performance in terms of number of explored configurations. We restrain our tool reductions to the domination test and best-first strategy, that are used in [2]. As the number of jobs grows, we observe a drastic size reduction by using the RATA approach. The gain rapidly reaches orders of magnitude.

Table 2. Comparison of the RATA model against timed automata

#j	***Timed automata***		***RATA***	
	#ds	*#bf*	*#ds*	*#bf*
2	77	38	25	22
3	629	384	125	105
4	4929	1561	625	306
5	37225	2810	3125	714
6	272125	32423	15625	2520

Table 3. The results for some instances of LA problems

instance	*#alg*	*#time*	*Opt*
LA01	176	0.1	666
LA03	3025	2.1	597
LA05	400	0.0	593
LA06	32460	11.2	926
LA08	17461	4.3	863
LA10	2851	0.4	958
LA11	13327	3.7	1222
LA13	3744	1.5	1150
LA15	/	/	/

In the last series (C), we consider three sets of small and medium benchmarks taken from the well-known OR-library : (1) three instances of size 10*5 (10 jobs and 5 machines), LA01, LA03 and LA05; (2) three instances of size 15*5 (15 jobs and 5 machines), LA06, LA08 and LA10 ; (3) three medium instances of size 20*5 (20 jobs and 5 machines), LA11, LA13 and LA15. All the proposed reduction techniques are used. The results of the experiments are shown in Table 3. The column *#alg* highlights the number of explored configurations and Opt shows the optimal time of the considered problem. As we can see, our algorithm is able to find the optimal for these instances except the last one in a reasonable time (within 5 minutes).

6 Conclusions and Perspectives

Exploiting the RATA model in order to solve optimal job-shop scheduling problems is a novel application of models based maximality-semantics, in addition to verification purpose [12,13]. Our experiments demonstrate that the search space developed from a RATA model could be drastically much smaller than the ones derived from standard timed automata. The proposed approach covers many reduction techniques that can be used to reduce the search space. In particular, it easily focuses on immediate transitions, moreover, the stubborn set, laziness and domination test techniques are improved and shown to be combined. To a better performance, a usual remaining time based reduction technique is introduced.

Our perspective consists in dealing with larger size problems. We refer to the ideas of [14,1], which argue that one should minimize the length of the schedules without necessarily targeting the optimal solution.

References

1. Subanatarajan, S., Thomas, T., Sebastian, P., Sebastian, E.: Multi-product Batch Scheduling with Intermediate Due Dates Using Priced Timed Automata Models. J. Computers and Chemical Engineering 33, 1661–1676 (2009)
2. Abdeddaim, Y., Asarin, E., Maler, O.: Scheduling with timed automata. J. Theoretical Computer Science 354(2), 272–300 (2006)
3. Abdeddaïm, Y., Maler, O.: Preemptive job-shop scheduling using stopwatch automata. In: Katoen, J.-P., Stevens, P. (eds.) TACAS 2002. LNCS, vol. 2280, pp. 113–126. Springer, Heidelberg (2002)
4. Belala, N., Saïdouni, D.E.: Non-Atomicity in Timed Models. In: ACIT 2005, Al-Isra Private University, Jordan (2005)
5. Alur, R., Dill, D.: A Theory of Timed Automata. J. TCS 126, 183–235 (1994)
6. Mokhdad, A., Ilié, J.M., Saidouni, D.E.: Addressing State Space Explosion Problem in Performance Evaluation Using Maximality-based Labeled Stochastic Transition Systems. In: 2nd International Conference on Computer and Software Modeling (ICCSM 2012), India, (2012)
7. Abdeddaim, Y., Niebert, P.: On the use of partial order methods in scheduling. In: Ninth International Conference on Project Management and Scheduling (PMS 2004) (2004)
8. Sebastian, P., Olaf, S., Sebastian, E.: Efficient synthesis of production schedules by optimization of timed automata. J. Control Engineering Practice 14(10), 1183–1197 (2006)
9. Godefroid, P., Wolper, P.: Using partial orders for the efficient verification of deadlock freedom and safety properties. In: Larsen, K.G., Skou, A. (eds.) CAV 1991. LNCS, vol. 575, pp. 332–342. Springer, Heidelberg (1992)
10. Sierra, M.R., Varela, R.: Pruning by dominance in best-first search for the job shop Scheduling problem with total flow time. J. Intelligent Manufacturing 21(1), 111–119 (2010)
11. Arfi, F., Ilié, J.M., Saïdouni, D.E.: Scheduling with RATA model. J. International Journal of Computer Science and Telecommunications (IJCST) 3(10), 14–20 (2012) ISSN:2047-3338
12. Saïdouni, D.E., Benamira, A., Belala, N., Arfi, F.: FOCOVE: Formal Concurrency Verification Environment for Complex Systems. In: Intelligent Systems and Automation (CISA 2008), Annaba, Algeria, vol. 1019 (1), pp. 375–380 (2008)
13. Saïdouni, D.E., Ghenaï, A. : Intégration des Refus Temporaires dans les Graphes de Refus. In : NOTERE 2006, Hermes, Toulouse, France, (2006)
14. Yang, S., Wang, D., Chai, T., Kendall, G.: An improved constraint satisfaction adaptive neural network for job-shop scheduling. J. Journal of Scheduling 13(1), 17–38 (2010)

Fuzzy Approach for Image Near-Duplicate Detection Using Gray Level Vertex Matching in Attribute Relational Bipartite Graphs

Goutam Datta and Bushan L. Raina

School of Computer Science, Lingaya's University, Faridabad, India
gdatta1@yahoo.com, rbushan@rediffmail.com

Abstract. Shape of different regions of an image depends on the detection of the corners. .However vague information such as blurring, noise etc in an image is normally due to missing curvatures in the regions of image and therefore for the reliable decision for the detection of images in the corresponding domains based on corners in gray level images of source image and its duplicate and for their equivalence we give an algorithm linking it to the set theoretic fuzzy technique. We propose Attribute Relational Bipartite Graph (ARBG) for image near-duplicate (IND) as an important model. The application of the proposed algorithm works well as vertex detectors in their respective two domains. The performance is tested on a number of test images to show the efficiency of the fuzzy based algorithm.

Keywords: ARBG, IND, Fuzzy ARBG, BVT.

1 Introduction

Image near- Duplicate(IND) refers to a pair of images G1 as source (original) image and G2 as target (Duplicate) image.Detection and retrieval of IND is very useful in a variety of real world applications. For example, in the context of copyright infringement detection. one can identify likely copyright violation by searching over the internet for the unauthorized use of images [1].Some prior work in this direction of Image Exact Duplicate (IED) detection has been exploited in various contexts. Extensive research has been performed on the Content Based Image Retrieval (CBIR) over several decades. We have yet to achieve the desired accuracy from fully automated CBIR systems [2]. In CBIR systems low as well as high level feature extractions are one of the very important tasks and as its application lots of related work has been carried out in face detection. In view of traditional image similarity model not being able to capture scene composition of IND, Zang and Chang [1]used part based image similarity measure for stochastic matching of **attribute relational graph**(ARG) related to crisp data. In many real life situations the data are not always crisp since these are usually uncertain in nature. The uncertainty is mainly of two types: Stochastic uncertainty and fuzzy uncertainty. The traditional two valued logical system based on 0 and 1 termed as crisp set theory and crisp probability theory are inadequate for detecting precise uncertainty.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 402–414, 2013.

This paper highlights the part based image similarity measure of edges with the help of Attribute Relational Bipartite Graph matching using Fuzzy membership for different grades of edges (strong, moderate and weak edges).

Some of the possible causes of IND are scene changes, movement and occlusion of foreground objects, absence or presence of foreground objects and background changes. Also camera view point change, camera tilting, exposure of lighting conditions etc. are also the causes of the variations of two near duplicate images.

Fig1 shows some examples of image near duplicate. This illustrates some of the IND due to camera position changes, a change in illuminations and position change of the objects in the image.

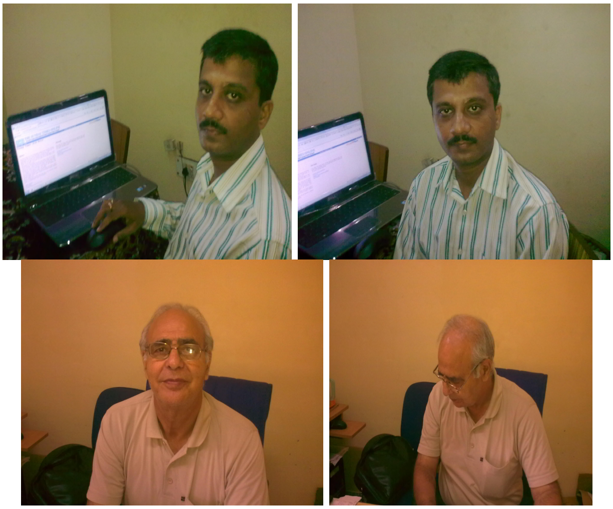


Fig. 1. Examples of some Image near Duplicate

Colour histogram is the most commonly used color representation.However; it does not include any spatial information [3]. Barrow [4] introduced ARG matching technique in his seminar work, as a result of which much research work has been done in this direction. He used energy minimizing technique (EMT) for possible modeling of the vertices of two attributive relational graphs (ARG). The computation of the similarity of the two images in their respective domain is done with the computation using heuristic rule or random algorithm like genetic algorithm. Spectral methods have assumed prominence in the machine learning for the graph partitioning [5]. This method for ARG matching [5] is a computational technique for Expectation Minimization (EM) formulation.. Zeng and Chang [1] used a Stochastic process to model the transformation from one ARG to the other. They have used the similarity of the two models by the likelihood ratio, for which STP(Stochastic Process) is decomposed in two steps. Step first refers to VCP (Vertex Copy Process) and step second as ATP (Attribute Transformation process). The Transformation process requires random as intermediate variableX= {0, 1} to specify correspondence between the vertices in G1 and G2. In order to certify the constraints Xiu≤1, they represent VCP by Markov Random Field (MRF) with two node potential that ensures 1-1 constraint being satisfied. The stochastic ARG model is based on log partition function Z(h). However, in view of h taking the values. 0 and 1 ' Z(h) can not be obtained corresponding to various values of h..as a result of which neither stochastic nor EM model can be applied. The approximate computation of likelihood ratio normally can not as well be realized by Loopy belief propagation (LBP) algorithm. Learning of similarity model under supervised fashion can neither be carried in vertex level nor unsupervised fashion in the graph level easily. Wilson and Hancock [6] used Bayesian frame work formulation for crisp related matching problem in relation to ARG's.

Vertex points in an image are generally formed at the junction of different edge segments which may be the meeting(or crisscrossing) of two images in each of the domains G1 or G2. Vertex point of an edge segment depends solely on the curvature formed at the meeting point of the two line segments in each of the images G1 and G2. We further note that firstly the vertex detection on gray level image can be classified by the gray level image being converted into its binary version for extraction of boundaries using some threshold techniques. And after the extraction of boundaries the vertex of high curvatures are detected using directed code and other approximation techniques. Secondly, we take a gray level image directly as vertex detection .We have used topology or auto correlation based approach [8-13] for purpose of gray level vertex detection of the image.

In this paper we have proposed the above analysis of STP and , introduce Attributive Relational Bipartite Graph (ARBG) to generalize ordinary graph by associating discrete or continuous feature of edges joining the vertices of corresponding intensities from source domain of G1 to target domain of G2. It is a bijective function in view of range of G1 being equal to the co-domain G2. ARBG based modeling is to compare the similarity of two images G1 and G2.

Here we use fuzzy based statistical framework that takes care of all types of variations in IND.Our frame work based on ARBG, to our best knowledge is the first of its kind regarding the detection of IND. The edge joining the node (vertex) in each of the two domains (source image) G1 and (target image) G2 does depend solely on the curvature formed at the meeting points at two lines in their respective domains. In this case maximum likelihood of similarity of the two ARBG's i.e. G1 and G2 is based on learning of forward and backward edge detection at vertex pairs. Non-detection of image usually arises due to the missing of significant curvature functions corresponding to the vertices. However Fuzzy set theory based modeling is well known for efficient handling of impreciseness. It is therefore reasonable for the sake of reliable decision making that we model image properties in fuzzy frame work to handle any incompleteness arising due to the imperfect data since it has been shown that vertices having high curvature are one of the dominant classes of patterns giving us significant amount of shape information[7].

The paper is organized as the following: Section 2 briefly describes the mathematical model used in this work. Section 3, as Fuzzy ARBG matching .Section 4 ,as Algorithm of Extraction of vertices. section 5 describes the experimental results and section 6 gives the conclusion.

Frame work for ARBG similarity Attribute Relational Bipartite Graph(ARBG) is an extension of the ordinary graph by associating discrete or real-valued attributes to its vertex and edges. The use of attributes allows ARBG not only be able to model the topological structure of an entity but also its non-structural properties, which often can be represented as feature vectors ARBG as defined by the following section.

2 Bipartite Graphs and Fuzzy Framework for Attribute Relational Bipartite Graph(ARG) Similarity in Two Domains Corresponding to G1XG2

Definition1: Bipartite Graphs: If the vertex set V of a simple graph G=(V,E) can be partitioned into two disjoints non empty sets V1 and V2 so that every edge in G is incident with vertex in V1 and vertex V2 then G is Bipartite.

Definition2:An attribute relational bipartite graph G=(V(i,j)xV(u,v), E,AxA), where V(i,j)xV (u,v) is Cartesian product of two vertex sets in two domains as subset of G1xG2 ,E is the edge setV(ij,uv)=(V(i,j),V (u,v))join of two vertices V(i,j) in G1 and V(u,v) inG2 and AxA is the attribute set that contains a binary attribute a(ij,uv) attaching to each edge $e_{ij,uv}$=V(ij,uv) in E.

2.1 Some Mathematical Model of Gray Level Corners in Bipartite Graphs G1 and G2

ARBG measures the similarity of the two images G1 and G2 . We apply the technique to visual scene matching and establish a part based similarity measure for detecting Near -Duplicate image using fuzzy technique. Their framework has helped us to reduce the computational time to detect the near duplicate image database.

Now we define a transform T1 as bipartite vertex transformation (BVT) associating the vertex from its (source) imageG1 to the vertex in target (image) G2 and the transform T2 as attribute transformation process(ATP) similarly defined for the attributes of the original and copied vertices and given by:

$$T1: V^{s}{}_{ij} \rightarrow V^{t}{}_{uv} \qquad (1)$$

where,

$$V^{s}{}_{ij,uv}=\{V^{s}(ij,uv)\}\ , (i.j)x(uv) \in G1xG2 \quad \text{and}$$

$$T2: F^{t}{}_{ij,uv}=\{F^{t}(ij,uv)\}, (ij,uv) \in G1xG2.$$

The transformation process as defined above requires an intermediate variable to specify the correspondence between G1 and G2.

Now we introduce a fuzzy random number X= X(h),h belonging to 0<=h<=1.In view of correspondence of vertices of the original and target image we denote Xij,uv as the sequence bearing the correspondence of ijth vertex in the original image G1 to the uvth vertex in the target image G2 with the membership function μA where xij belongs to A.And therefore 0≤μA(xij)≤1, i= 1,2,….M ; j=1,2,….N where the size of source of G1and G2 each is NXM. We draw graphs for TP and ARBG as shown:

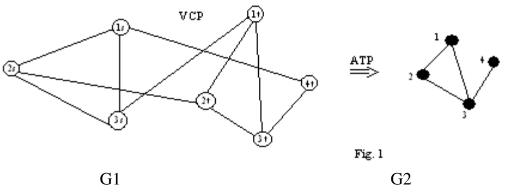


Fig. 2. Transformation Process(TP) from G1 to G2

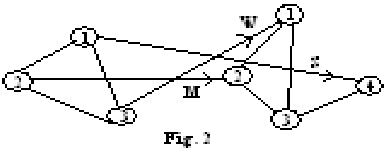


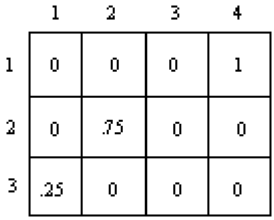
Fig. 3. Correspondence of W=WNDI,M=MNDI,S=SNDI correspondence of fuzzy ARBG,

Where WNDI=weak,MNDI=medium, SNDI=strong intensity of edges connecting V(i,j) to V(u,v).

Vertex correspondence of two ARBGsthe transformation as given in fig.1 requires an intermediate variable to specify the correspondence between the vertices in Gs and Gt. We denote it by xijuv represent it by X and referred to as a correspondence matrix which is random taking the values from 0 to 1 (inclusive) such that X=[0,1] of order NXM. Here xijuv=X having values greater than 0.75 means that the (i,j)th vertex in Gs =G1 correspondence strongly to the (u.v)th vertex in Gt=G2.Similarly the values lying between greater than 0.50 and 0.75 indicate the moderate correspondence and the values less than 0.5 indicates the weak correspondence between the source image and target image.We further note here that in view of injective function that is one to one correspondence of the vertices in Gs and Gt we need to have following constraints that is

$$\sum_{i,j} xijuv \le 1; \qquad \sum_{u,v} xijuv \le 1;$$

The matrix given by fig.3 as being represented by Fig.2 according to the above definition with regard to the strength of the edge corresponding pixels being strong, moderate and weak in each of the domains of G1 and G2 is given by the following:



	1	2	3	4
1	0	0	0	1
2	0	.75	0	0
3	.25	0	0	0

Fig.3
Matrix Representation of Fig.2

(Intensity of the edge corresponding to the equivalence or nearly equivalence of pixels in their respective domains)

One of the most widely used mapping function to fuzzify for converting each of the digital images G1 and G2 to corresponding subset A denoted by $\mu A(x)$ and otherwise defined by standard gamma function ¥ is given as follows:

$$
\begin{aligned}
¥(x{:}\alpha,\ \beta)&=0, && x\leq \alpha\\
&=(x-\alpha)\ /\ (\ \beta-\ \alpha), && \alpha<x\leq\beta \ldots\ldots\ldots\ldots \qquad \ldots\ldots\ldots\ldots..(2)\\
&=1 && x>\beta
\end{aligned}
$$

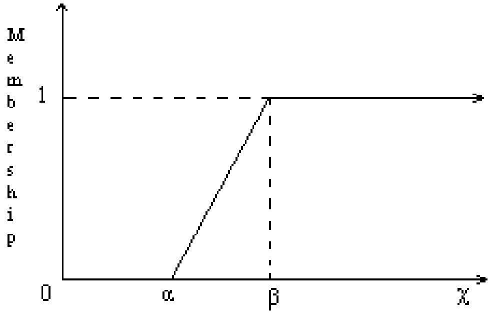


Fig. 4
S (χ, α, β)

Fig. 4. Pictorial representation of the γ function is given in fig.4

Here fig.4 shows the graphical representation where the parameter β is the cross over point at which the image signal of G1 and G2 transforms from moderate to strong and is given by γ $(\beta;\alpha,,\beta)\geq0.75$ and the initial point α on x axis at which the image signal transforms from weak to moderate signal and is given by γ $(\alpha;\alpha,,\beta)\geq0.25$

Fuzzy Alpha Cut: A fuzzy subset can be divided by a suitable thresholding membership values around the range of interest.

The fuzzy alpha cut γ-alpha comprises all elements of X whose degree of membership in γ is greater than or equal to alpha where

$$\gamma\text{- alpha}=\{x \in X{:}\ \mu A(x)>\alpha\} \text{ where } 0\leq \alpha\leq1$$

2.2 Fuzzy Modeling of Detection of Image and Its Duplicate Based on Fuzzy Connectivity Strength of the Pixels in Their Two Domains

Here Vertex V(i,j)=V(ij) in G1 corresponds to its equivalent strength of intensity to the node V(u,v)=(V(uv) in G2 in reference to the bipartite graphs as given by equation (1) such that their equivalent strength determining the intensity in their respective images based on the connecting edge eijuv=V(ij,uv) In this case fuzzy member controls the probability that the vertices in the source graph are copied to the target graph. For stochastic ARBG we do not assign the crisp values 0 and 1 but fuzzy based values lying between 0 and 1 (inclusive). Thus under the hypothesis H=h, we have$0\leq h\leq 1$.

The features associated to the vertices is given by Eq: (1). The generative model for ARBG matching is given by graph in Fig.5:

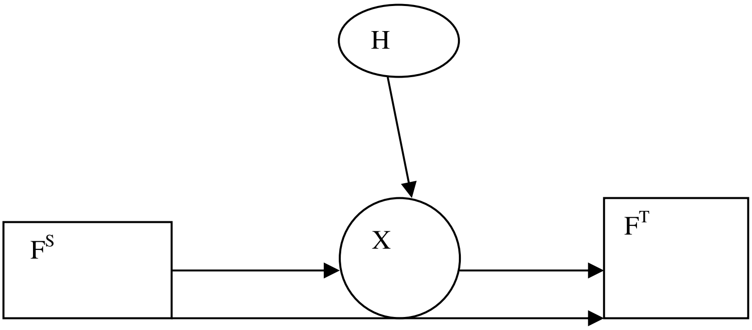


Fig. 5. Fuzzy based similarity related ARBG,H=h, $0\leq h\leq 1$

3 Algorithm of Forward and Backward Edge of Gray Matching of the Vertex Pixels in G1 and G2 for IND

Computation of features is based on two Phases. The first being that the general (i,j)th pixel P=Pij of source image G1 as in Fig.6 is associated to the general (u,v)th pixel with the nearly same intensity in the target image G2 and given by the edge $e_{ij,uv}$=V(ij,uv) which is extracted from the respective regions with their respective near intensities of the two domains.

(*i*-1, *j*-1)	(*i*,*j*-1)	(*i*+1,*j*-1)
(*i*-1,*j*)	(*i*,*j*)	(*i*+1,*j*)
(*i*-1,*j*+1)	(*i*,*j*+1)	(*i*+1,*j*+1)

Fig. 6.

The edge V(ij,uv)is referred to by the membership function $\mu[P]$ based on its respective gradient strength. This gives rise to the fuzzy edge set:

$$E(d)=\{ V(ij,uv),F_{ijuv}\} =\{P,\mu P\}, \qquad (3),$$

and secondly two membership functions associated to fuzzy connectivity strength through forward and backward direction with respect to the pixel be given by $\mu f(P)$ and $\mu b(P)$.

Each of the original images G1=I(i,j) and target image G2(u,v)=I(u,v) is convolved with the Gaussian function $G(m_i,n_i)$ to obtain the Gaussian smoothened image matrix.

$$I_{uv}(m_i,n_i)=I_{ij}(m_i,n_i))*G(m_i,n_i) \qquad (4a)$$

$$\text{where } G(m_i,n_i)=1/\sigma i\sqrt{2}\prod*\exp(-(m_i*m_i+n_i*n_i))/2\sigma_i*\sigma_i \text{ ,i=1,2} \qquad (4b)$$

Here σ_i is the degree of smoothness to the small distortion of noise and blurred boundaries. We choose σ=1.2 to smoothen those noise and blurred boundaries corresponding to 3X3 pixels. The membership of edge e_{ijuv}=V(ij,uv) given otherwise by $\mu(P)$ belongs to the cross product G1*G2image . Thus for every edge- pixel P(ij,uv) where gray value of the (i,j)th vertex in G1 is associated to equal or nearly equal gray value of (u,v)th vertex in G2.We consider as usual 3x3 windows in each of source G1 and Target image G2.In the adjoining figure no.7 , the symbols representing gray values in G2 are on parallel lines as the gray values in G1 as given below :

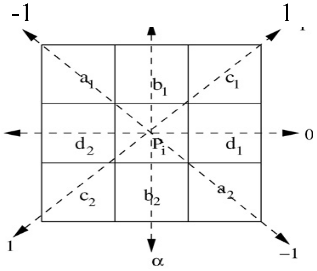


Fig. 7. Gray values in each Gi, i=1,2

If d(Pijkl,uvkl)=Ds's=1,2,3,4 denote the differences between the gray values at P(ij+kl,uv+kl) and P(ij,uv) for k,l= -1,0,1 in four different directions given by d(Pijkl,uvkl) =!P(ij+kl,uv+kl)-P(ij,uv)!,k,l=0,1,-1, so that the difference of these pixels are taken as gray level differences. The ratio of gray level changes as given by (Gr) are computed from two mutually perpendicular set of pixel pairs in the vicinity of respective central pixels P(ij,uv) in G1*G2.Following ,the contrast ratios Grij in G1 of the four values of the pixels obtained from the neighborhood central pixel P(i,j) is given by

$$Gr=[(1+D1)/(1+D2),(1+D2)/(1+D1),(1+D3)/(1+D4),(1+D4)/(1+D3) \qquad (5)$$

where Ds,s=1 to 4 correspond to the distances of the pixels in the opposite corners with respect to the two mutually perpendicular directions through P(i,j) given by

D1=d(Pi-1,j+1)-d(Pi+1,j-1)=a1-a2,
D2==d(Pi+1,j+1)-d(Pi-1,j-1)=c1-c2l,
D3==d(Pi+1,j)-d(Pi-1,j)=d1-d2 and
D4= d(Pi,j+1)-d(Pi,j-1)=b1-b2l.

The contrast ratios Gruv in G2 of the corresponding four values of pixels obtained from the respective neighbourhood central pixel f(P(u,v)=Q(u,v) are defined on the same line.

Referring [6], since in a window of eight neighborhood, an edge pixel has a maximum gray level difference in a direction perpendicular due to edge direction Ψ and this direction points along minimum difference direction as a result of which minimum pixel contrast ratio(Gm(r)) is given by

$$((Gm(r))=\min\{Gm\} \qquad (6)$$

where the parameter m (slope) is used for computing the gradient membership μ(Pij,uv)=μ(P). Here μ(P) is used to represent the uncertainties of edge strength and location of true edge point. It is more appropriate to represent it by γ-function derived by the computation of pixel contrast ratios. Having known the value of the contrast ratio of (Gm(r)), it is easy to obtain its maximum and minimum values which help us to obtain histogram representation of the pixel contrast ratio between the two images of G1 and G2.

The value of μ(P) determines the edge fuzzy membership strength corresponding to the join of the vertices(pixel)of equal or nearly equal strength intensity in G1 and G2 . Let μSij, μMij, and μ Wij represent the membership values of strong, moderate and weak of(i,j)th pixel of G1 such that (u,v)th target images corresponds to source image with (i,j)th pixel of equivalent of nearly equivalent intentisity. These characteristics can be represented in a tabular form and are called the fuzzy matrices with its elements as fuzzy values given in Fig.3.

We now select a subset of fuzzy gradient map E[ed].Here we first define two membership functions μf(P) and μb(P) and these are computed to estimate the strength of the fuzzy connectivity along the paths in forward and backward directions with regard to the pixel(P) . If P(ijuv,f,b) represent the set of points having μf(P)high and μb(P)low on the left side of the curvature junction of the pixel P then the edge corresponding to P(ij,uv) designates these set of points(edge detecting equivalences of vertex intensities in the two images) on the forward arm assigning the membership μf(P)-μb(P) and this difference varies with the sharpness of curvature and we represent these points of P(ij,uv,f,b) by fuzzy subset μfb as forward arm given by

$$\mu fb(P) =\mu f(P)-\mu b(P) . \qquad (7)$$

On the other hand similarly if μf(P)is low and μb(P)is high on the left side of the curvature junction of the pixel P then P(ij,uv)designates these set of points on the backward arm and assigned the membership μb(P)-μf(P) and this difference varies with the sharpness of curvature .We represent these points of P(ij,uv,b,f) by fuzzy subset μbf as backward arm and is given by a

$$\mu bf(P)=\mu b(P)-\mu f(P) \qquad (8)$$

We then define the fuzzy gradient map μ(ed) = {P,μ(P)}.Here we remark that the membership values associated to the pixel(P) in the forward and backward direction correspond to the gradient values m=1 and m=-1 respectively.

LetΨ ={ Ψ1 , Ψ2,…..,Ψn}represent the slope(edge direction)corresponding to sequence of pixels on an edge segment Here we note that the directions of the forward and backward fuzzy membership values are computed from the difference in Ψth edges between the connected pixels. For example we further note that the edges correspond to the following slopes given by the slope set M={m : m=0,+1,-1,+∞,-∞}.The slope subsets {-1},{1}represent the forward counts and backward counts denoted by n(f)and n(b) respectively and similarly the slope subset {0,- ∞,+ ∞}represent the n(f) and n(b) respectively and so on. These counts vary with the sharpness of the curvature type. The membership values of the forward and backward directions associated to the pixel(P)have the attributes of exponential decay as such these can be represented by μf(P)=β*exp(-s),where s=1/nf and μb(P) = β*exp(-t),t=1/nb corresponding to some scalar multiplier β.

Here corresponding to finite counts of n(f) and n(b) of the image the values of μf(P) or μb(P) should lie between 0 and 1.Thus from the above vertex testing we select the edge pixel P which can represent the three dimensional feature Fi: where

$$Fi = [\mu(P),\mu f(P),\mu b(P)] \qquad (9)$$

We detect the fuzzy edge associated to equal intensity vertices (pixels) with equal or nearly equal intensity in the two domains of G1 and G2 with the fuzzy edge map μ(ed)={P,μ(P)}.

Initially a suitable threshold value of the gradient membership has to be decided to select a subset of E(ed) and only those points of the two domains will be used for computation of μf(P) and μb(P) that detect fuzzy edge for detection of equal strength of the vertices in the two images to give IND.. To locate points from significant portions on the image, a contrast transformation is used as a pre- processing step as shown by the following :

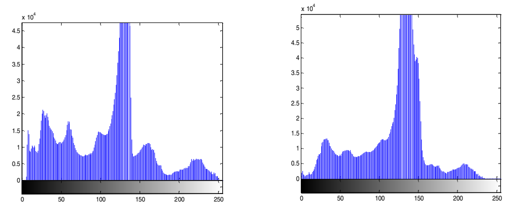


Fig. 8(a). Pixel Contrast Histogram ratio of Fig.1(a)

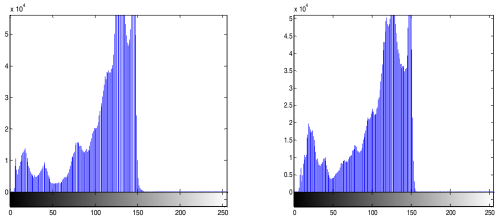


Fig. 8(b). Pixel Contrast Histogram ratio of Fig.1(b)

Then the extraction of probable edges of the images is achieved by thresholding through non- linear transformation of membership values as associated to γ function as above.

Here we set β=0.9 , α=0.7 its membership function then is the corresponding γ function given by fig. 4 which can be written as :

$$\begin{aligned} \mu d(P) &= 1, && a> 0.9, \\ &=(10\,\mu P-7)/2\,, && 0.7<b<0.9 \\ &= 0, && b<0.7 \end{aligned} \qquad (10a)$$

Thus the above pixel contrast transformation operation can alternatively be also given as:

$$\begin{aligned} \mu d(P) &=(10\,(\mu P)-7)/2, && 0.7<\mu P\leq 0.9, \\ &= 1\,, && \mu P>0.9 \end{aligned} \qquad (10b)$$

The following figures **9(a) and 9(b)** result from the membership values of μ(dp) before and after the transformation associated to equation(2a).

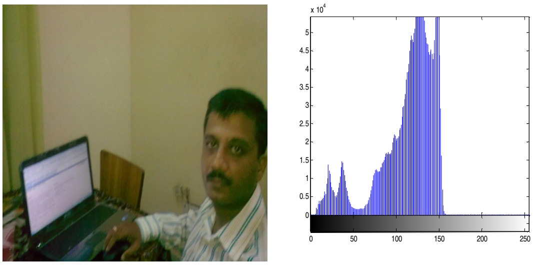


Fig. 9(a). Threshold points

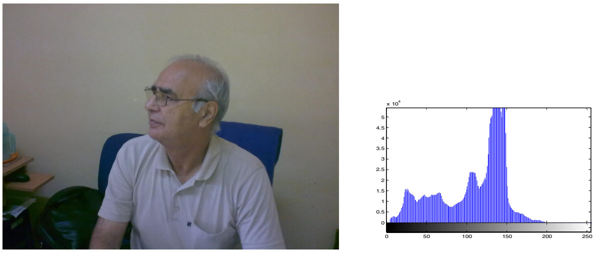


Fig. 9(b). Threshold points

We further notice that the number of insignificant vertices under discussion are reduced at the same threshold value in G1 and G2. Let Ed={P,μd(P)}denote the transformed edge map from source image to target image then the thresholding above the different membership values may be obtained by the use of proper alpha cuts[15] as discussed in section 2 .Due to this process we are able to obtain the edge of equivalent strength of pixels connecting G1 and G2 given by Ed(alpha) and denoting by Edf and given by

Edf=Ed(alpha)={P€Edf:µd(P)≥alpha}, 0≤alpha≤1.0 where Ed(alpha) is represented by the local features

Fi = [µd(P),µf(p),µb(P)].

Multilevel fuzzy edge map may be generated by the thresholding of Ed in which case the corresponding pixels of G1 and G2 will be segregated as strong , moderate and weak edge pixels based on their gradient membership values of corresponding values of µd(P).

Analysis of images related to the fuzzy membership is given by the following table:

The values of the edge points of G1 and G2 corresponding to respective µd(P) are closer to each other than if the local contrast of a region is very poor and on the other hand if the membership value of the edge joining nearly equal strength of the intensity of vertex pixels of G1 and G2 are separated above the crossover points of each µd(P) ≥0.75 then the similarity is stronger. Thus in view of edges associated to vertices with membership µd(P)≥0.75 will represent the relative comparison of the images strongly in accordance of the strength of pixels in G1 and G2.

4 Experimental Results

We have extended the performance of detectors on various images so as to extract the edge map with the help of membership function with respect to the threshold so that µd(P) is strong, moderate or weak. The region could not be extracted where contrast ratio is poor and threshold is less than or equal to0.75.Our method of detection of vertices compare very well with the methods of detectors using Harris and SUSAN methods.

We notice here that there are large variations of distinct gray values with high threshold values corresponding to µd(P) to reduce the no. of weak and noisy vertices. Though their detections are more general however, IND images based on our algorithm works reasonably well. Our proposed detector of near duplicate images is able to extract most of the significant vertices of nearly equal strength images of the two domains G1 and G2 under different imaging conditions .This is mainly because of the reason that the slope of the fuzzy property plane is obtained from the dynamic range. From the above results we therefore confirm using ARBG the near image duplicate detection working satisfactorily based on our algorithm.

5 Conclusion

We present fuzzy framework for calculating the similarity of two attributed relational bipartite graphs (ARBG) in grey level vertex matching. However, we applied fuzzy set theoretic approach for the detection of the vertices of near duplicate images. The algorithm in this approach does not require computation of chain codes or complex differential geometric operators or the complex partition function as used in the ARG matching for Stochastic frame -work since in such cases similarity is based on EM scheme which is based on two learning phases firstly on the application of supervised vertex learning and secondly on graph level learning which is complex[1]. Using fuzzy approach, our experiments have been performed on various types of twenty images and the algorithm performs reasonably well. However, we shall explore the possibility to improve the algorithm so that the parameters may be selected adaptively for thresholding.

References

1. Zhang, D.-Q., Chang, S.-F.: Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. In: ACM Conference of Multimedia 2004, ACM MM (2004)
2. Banerjee, M., Kundu, M.K., Maji, P.: Content based image retrieval using visually significant point features. Fuzzy Sets and Systems 160, 3323–3341 (2009)
3. Wang, X.-Y., Yu, Y.-J., Yang, H.-Y.: An efficient image retrieval scheme using color, texture and shape features. Computer Standards and Interfaces (2010)
4. Barrowand, H.G., Poppleston: Relational Description in picture processing. Machine Intelligence 6, 377–396
5. Zhang, D.-Q., Chang, S.-F.: Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. DVMM Technical report #206-2004-6
6. Wilson, R.C., Hancock, E.R.: Structual matching by discrete relaxation. IEEE Trans. Pattern Anal.Mach. Intell. 19, 1–2 (1997)
7. Banerjee, M., Kundu, M.K.: Handling Impreciseness in gray level corner detection using fuzzy set theoretic approach. Journal of Applied Soft Computing (2008)
8. Kitchen, L., Rosenfeld, A.: Gray- level corner detection. Pattern Recogn. Lett. 1, 95–102 (1982)
9. Zheng, Z., Wang, H., Teoh, E.: Analysis of gray level corner detection. Pattern Recogn. Lett. 20(2), 149–162 (1999)
10. Rattarangsi, A., Chin, R.T.: Scale-based detection of corners of planar curves. IEEE Trans. Pattern Anal. Mach. Intell. 14(4), 430–449 (1992)
11. Teh, C., Chin, R.T.: On the detection of dominant points on digital curves. IEEE Trans. Pattern Anal. Mach. Intell. 11(8), 859–872 (1989)
12. Rosenfeld, A., Johnston, E.: Angle detection on digital curves. IEEE Transaction on Computers C 22, 858–875 (1973)
13. Bae, S.C., Kweon, I.S., Yoo, C.D.: Cop: a new corner detector. Pattern Recogn. Lett. 20, 1349–1360 (2002)

Formal Modeling of Mobile Middleware for Tuple Space Coordination over Multiple Heterogeneous Networks

Suddhasil De, Diganta Goswami, and Sukumar Nandi

Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati, Assam – 781039, India
{suddhasil,dgoswami,sukumar}@iitg.ernet.in

Abstract. Tuple Space based Mobile Middleware (TSMM), with tuple space as coordination medium, exhibits multiple decoupling qualities during coordination, which enhances its robustness and flexibility and makes it an appropriate coordination platform for underlying mobile and dynamic networks. However, formal semantics of TSMM are required for reasoning TSMM as coordination platform, which also help in developing supported applications. This paper suggests an approach of formalizing TSMM that can be deployed over multiple heterogeneous mobile and dynamic networks. Formalization is carried out using Mobile UNITY.

Keywords: Mobile middleware, coordination, tuple space, robustness, formalization, Mobile UNITY.

1 Introduction

Advances in wireless communication technologies and mobile computing devices lead to the deployment of different scales of wireless networks. These networks are characterized by device mobility, network dynamics, and inherent unreliability in communication links. They are targeting different types of applications for the benefits of end users. Most of their applications require coordination support for proper functioning to achieve a common goal. Services of *mobile middleware* [1], with proper *coordination medium* incorporated within it, becomes inevitable for facilitating coordination among different active components of a supported application (called *agents*) executing in computing environments of different devices (called *hosts*). One such coordination medium, tuple space [2], attains different dimensions of uncoupling between interacting agents [3], even in underlying heterogeneous networks. Mobile middleware incorporating tuple space for coordination is referred as *Tuple Space based Mobile Middleware* (TSMM) [4].

In TSMM, *tuple* is basic unit of data exchanged during agent interactions via a shared repository (called *tuple space*), while *antituple* is basic unit of search key to identify tuples residing in tuple space. Tuple space as coordination medium uncouples interacting agents about time, space (i.e. naming), interacting data and operations on them [3,5]. Uncoupling in several dimensions

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 415–430, 2013.

enable TSMM of providing loose coupling of coordination, which enhances its robustness and flexibility and makes it an appropriate coordination platform for underlying dynamic heterogeneous networks. However, to facilitate application designers, formal specifications of semantics of TSMM are required to be clearly stated. Formalization not only enables proper analysis of robustness and flexibility of TSMM over heterogeneous networks, but also defines its precise semantics and prepares foundation for its implementation. This paper extends an earlier work [6], by providing a formal treatment to TSMM that is deployed over multiple mobile, dynamic and unreliable heterogeneous networks. Mobile UNITY [7], a general-purpose reasoning tool, is used for formalization.

In literature, formal semantics of tuple space model has been presented earlier [5,8,9]. In these works, basic tuple space operations and agent mobility of TSMM are formalized using Mobile UNITY. However, unlike [5,8,9], this paper focuses on formalizing aggregated functionalities of TSMM, including multiple dimensions of uncoupling, communication and discovery mechanisms etc. Tuple space operations are abstracted in this formalization as simple calls to respective primitives, while agent mobility is abstracted as a function to simplify its representation. Also, this paper shows building of formal representation of TSMM by combining individual specifications of its different functionalities. Compared to [6], this paper extends by including support of multiple underlying heterogeneous networks, all of which are mobile, dynamic and unreliable. In particular, TSMM, formalized in [6], considers Infrastructure Basic Service Set (iBSS) [10] as the only underlying network, whereas, TSMM in this paper supports both iBSS and Independent Basic Service Set (IBSS) [10]. Two heterogeneous networks are considered for this paper to keep the formalization readable. However, this formal treatment can be easily extended to include other underlying heterogeneous networks for TSMM. Rest of the paper is organized as follows. Section 2 gives a brief overview of TSMM, which is next formalized using Mobile UNITY in Section 3. Finally, Section 4 concludes the paper.

2 Overview of TSMM Having Multiple Decoupled Coordination

TSMM is the coordination platform to support agent interactions in mobile distributed applications, thereby providing ubiquity to user activities.

Architecture. TSMM comprise of several components, which can be organized within agent or host. Each instance of agent contains an *agent tuple space* (ATS) and its interfaces, local operation manager, remote operation manager, ATS reaction manager and acquaintance list. One instance of host runs in one device and supports execution of single/multiple agents. In each host, different components manage functionalities of communication, discovery, host server, *host tuple space* (HTS) and its interfaces, agent management, mobility etc. Architecture of TSMM with all its components is shown in figure 1.

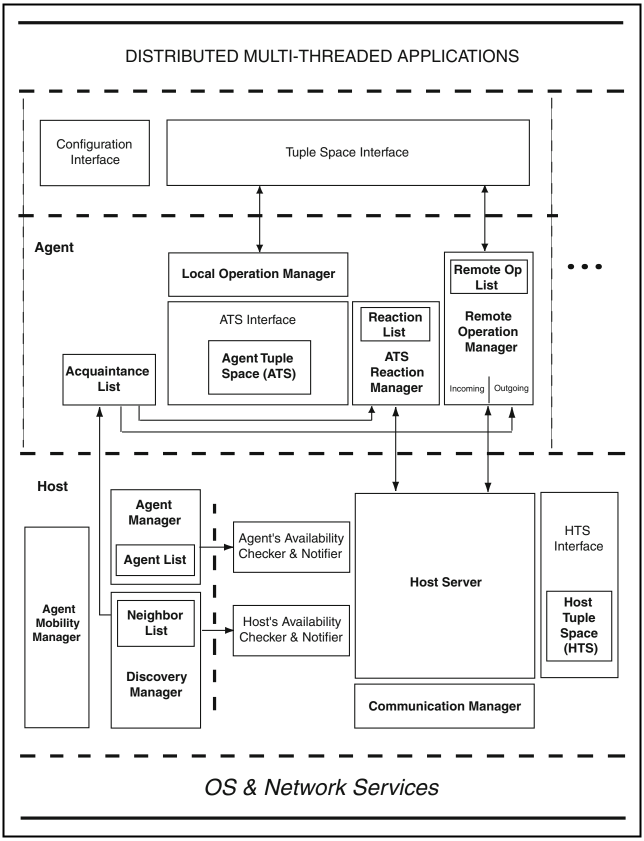


Fig. 1. Architecture of TSMM showing its significant components [6]

Tuple Space Model. In TSMM, tuples and antituples comprise of *unordered* sequence of fields [11], whereas tuple space is *indexed* in nature [12]. During interaction between any pair of agents, initiator of interaction becomes *reference agent* and destination becomes *target agent*. Different primitives are defined for writing, reading and withdrawing tuples from tuple space (like, `ATS`) using tuple-producing, tuple-reading and tuple-consuming primitives. Tuple-producing primitives cover `out` and `outg`, while tuple-reading primitives include `rd`, `rdp`, `rdg` and `rdgp`, and tuple-consuming primitives are `in`, `inp`, `ing` and `ingp`, details of which are given in [5]. Local primitives are executed in own `ATS`, whereas for executing remote primitives, invoked parameters are shipped to specified target agent(s), executed in its `ATS` and results of execution are sent back.

Reactivity Model. TSMM incorporates *reactivity* in `ATS` to monitor and respond to different events (like, presence of a particular tuple in tuple space etc.) during execution. Reactivity is implemented by generating and registering *reaction* in `ATS`. Registered reaction, with condition specified by antituple, fires if that condition gets satisfied i.e. antituple matches tuple in tuple space. Firing of reaction signifies execution of application-defined reactive codes, like notifying presence of tuples, withdrawing tuples from `ATS` etc, and responses are sent back.

Decoupled Coordination Model. In TSMM, agent interactions use decoupled reactivity [5], whereby `HTS` provides additional decoupling medium to accomplish complete decoupling of agent interactions. `HTS` stores two special tuples (viz. reaction tuple and response tuple). Reaction tuples are created from different parameters of invoked remote primitives, while response tuples are created from result of execution of these primitives. Reaction tuple is first inserted into `HTS` of reference host. On availability of target host, it is withdrawn from that `HTS`, passed through underlying infrastructure to target host, and subsequently inserted into its `HTS`. Eventually, reaction tuple is withdrawn from target host's `HTS`, once target agent is available. Parameters of invoked primitive is next extracted and execution of that primitive starts at `ATS` of target agent. In case of remote tuple-reading and -consuming primitives, target agent packs results of execution (viz. sought tuple(s) from its `ATS`) into response tuple. Following previous approach, that response tuple eventually reaches reference agent, and sought tuple(s) are extracted from it. For achieving consistency in coordination, reference agent responds back to target agent(s) with additional ACK tuple and NACK tuple for any invoked remote tuple-consuming primitive. ACK tuple positively acknowledges acceptance of responded tuple as sought tuple, whereas NACK tuple returns non-accepted responded tuple back.

Supplementary Components. For execution over multiple unreliable networks having mobility, as well as for resolving heterogeneity of such multiple underlying networks, TSMM includes its own communication and discovery mechanisms [13] that uses transport service for data transmission. This paper considers iBSS and IBSS as underlying networks for TSMM. When deployed over iBSS, three categories of hosts are earmarked, viz. *stationary host*, *mobile host* and *access point*, whereas for IBSS, deployed hosts are earmarked as *mobile host* only. Discovery mechanism furnishes an updated knowledge of available agents and hosts. This knowledge is attained by sending and receiving beacons and is preserved in `NeighborList`. Communication mechanism emphasizes on reliably transferring reaction/response tuples from one host to another. It uses additional acknowledgement mechanism to achieve this reliability.

3 Proposed Approach of Formalization of TSMM

This section proposes an approach of formalizing TSMM as a Mobile UNITY system, comprising of a set of formal programs representing different agents and hosts. Favoring Mobile UNITY over other formal tools is due to its suitability in formalizing inherently non-terminating programs (like mobile middleware) and reasoning about agents temporal behavior using its proof rules. **System** *TSMM*, shown in Figure 2, comprises of multiple instances of two Mobile UNITY programs, and their interactions are specified in **Interactions** section.

```
System TSMM

  Program $host(i)$ at $\lambda$
      ⋮  ⋮    {Program description of $host(i)$, given separately}

  Program $agent(k)$ at $\lambda$
      ⋮  ⋮    {Program description of $agent(k)$, given separately}

  Components
    $\langle [\,]\ i :: host(i)\ \rangle\ \ [\,]\ \ \langle [\,]\ k :: agent(k)\ \rangle$

  Interactions
      {Attach $\mathcal{T}_{\mathrm{W}}$ of hosts with wired network interfaces in $iBSS$ as transiently-shared variable when connected}
    $shared_{\mathrm{W_{iBSS}}}$ ::
    $\langle [\,]\ i,j :: host(i).\mathcal{T}_{\mathrm{W}} \approx host(j).\mathcal{T}_{\mathrm{W}}$
          when $(host(i).nwdeploy = iBSS) \wedge (host(j).nwdeploy = iBSS) \wedge (host(i)\Gamma' host(j))$
              $\wedge\ (\mathtt{isSH}(host(i)) \vee \mathtt{isAP}(host(i))) \wedge (\mathtt{isSH}(host(j)) \vee \mathtt{isAP}(host(j)))$
          engage  $host(i).\mathcal{T}_{\mathrm{W}}$        disengage  current $\| \bot$  $\rangle$

      {Attach $\mathcal{T}_{\mathrm{WL}}$ of mobile host and access point in $iBSS$ as transiently-shared variable, only when colocated}
  $[\,]$ $shared_{\mathrm{WL_{iBSS}}}$ ::
    $\langle [\,]\ i,j :: host(i).\mathcal{T}_{\mathrm{WL}} \approx host(j).\mathcal{T}_{\mathrm{WL}}$
          when $(host(i).nwdeploy = iBSS) \wedge (host(j).nwdeploy = iBSS) \wedge (host(i)\Gamma' host(j))$
              $\wedge\ ((\mathtt{isMH}(host(i)) \wedge \mathtt{isAP}(host(j))) \vee (\mathtt{isAP}(host(i)) \wedge \mathtt{isMH}(host(j))))$
          engage  $host(i).\mathcal{T}_{\mathrm{WL}}$        disengage  current $\| \bot$  $\rangle$

      {Attach $\mathcal{T}_{\mathrm{WL}}$ of mobile hosts in $IBSS$ as transiently-shared variable, only when colocated}
  $[\,]$ $shared_{\mathrm{WL_{IBSS}}}$ ::
    $\langle [\,]\ i,j :: host(i).\mathcal{T}_{\mathrm{WL}} \approx host(j).\mathcal{T}_{\mathrm{WL}}$
          when $(host(i).nwdeploy = IBSS) \wedge (host(j).nwdeploy = IBSS) \wedge (host(i)\Gamma' host(j))$
              $\wedge\ \mathtt{isMH}(host(i)) \wedge \mathtt{isMH}(host(j))$
          engage  $host(i).\mathcal{T}_{\mathrm{WL}}$        disengage  current $\| \bot$  $\rangle$

      {Prepare to register active agents in respective hosts}
  $[\,]$ $regAgent :: \langle [\,]\ i,k :: host(i).\mathcal{Q}_{in} := host(i).\mathcal{Q}_{in} \bullet agent(k).aid$   when $(host(i).\lambda = agent(k).\lambda)\ \rangle$

      {Prepare to deregister terminated/migrated agents from respective hosts}
  $[\,]$ $deregAgent :: \langle [\,]\ i,k :: host(i).\mathcal{Q}_{out} := host(i).\mathcal{Q}_{out} \bullet agent(k).aid$   when $\neg(host(i).\lambda = agent(k).\lambda)\ \rangle$

      {Prepare to transfer reaction/response tuple from agent to host}
  $[\,]$ $\langle [\,]\ i,k :: host(i).\mathcal{Q}_{T^S_{a_k}}, agent(k).\mathcal{Q}_{T^S_{a_k}} := host(i).\mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{head}(agent(k).\mathcal{Q}_{T^S_{a_k}}), \mathtt{tail}(agent(k).\mathcal{Q}_{T^S_{a_k}})$
          when $(host(i).\lambda = agent(k).\lambda) \wedge \neg(agent(k).\mathcal{Q}_{T^S_{a_k}} = \bot)\ \rangle$

      {Prepare to transfer reaction/response tuple from host to agent}
  $[\,]$ $\langle [\,]\ i,k :: agent(k).\mathcal{Q}_{T^R_{a_k}}, host(i).\mathcal{Q}_{T^R_{a_k}} := agent(k).\mathcal{Q}_{T^R_{a_k}} \bullet \mathtt{head}(host(i).\mathcal{Q}_{T^R_{a_k}}), \mathtt{tail}(host(i).\mathcal{Q}_{T^R_{a_k}})$
          when $(host(i).\lambda = agent(k).\lambda) \wedge \neg(host(i).\mathcal{Q}_{T^R_{a_k}} = \bot)\ \rangle$

end
```

Fig. 2. Mobile UNITY system of TSMM

i-th host is specified by **Program** $host(i)$, whereas k-th agent is represented by **Program** $agent(k)$, where i and k are assumed to be quantified over appropriate ranges. Different conditions of interactions in **Interactions** section are enforced through **when** clauses. Clauses **engage** and **disengage**, and construct **current** are used for transient sharing between different hosts. Also, first three statements in **Interactions** section, labeled as $shared_{\mathrm{W_{iBSS}}}$, $shared_{\mathrm{WL_{iBSS}}}$ and $shared_{\mathrm{WL_{IBSS}}}$, are

```
Program $agent(k)$ at $\lambda$

  declare
      $type$ : $\in\{stationary, mobile\}$
    [] $aid, taid, a$ : agentid [] $taids$ : sequence of agentid
    [] $\mathbf{T}$ : tuple space [] $t, tuple$ : tuple [] $\mathbf{t}, tuples$ : set of tuple [] $a, atuple$ : antituple
    [] $r$ : $\mathrm{RT_{tuple}}$ [] $\mathcal{Q}_{T^S_{a_k}}, \mathcal{Q}_{T^R_{a_k}}$ : queue of $\mathrm{RT_{tuple}}$
    [] $prid$ : primitiveid [] $prType$ : $\in\{local, remote\}$
    [] $ROL$ : sequence of (primitiveid, primitivename, set of agentid of target agents)
    [] $RL$ : sequence of (reactionid, primitiveid) [] $\mathbb{T}$ : set of {agentid, set of tuple}
    [] $prName$ : $\in\{$OUT, OUTG, RD, RDG, RDP, RDGP, IN, ING, INP, INGP$\}$
    [] $mode$ : $\in\{$ONCE, ONCE/TUPLE$\}$
    [] $TAs, rform$ : natural
    [] $prBulk, prRdIn, UsrRdy4Evt$ : boolean

  always
      $aid :=$ getMyAgentID($k$) [] $type :=$ getAgentType($stationary, mobile$)
    [] $\mathtt{isPresent}_{\mathrm{in}ROL}(prid, taid) \equiv \langle\ \exists e :: (e \in ROL) \wedge (e \uparrow 1 = prid) \wedge (aid \in e \uparrow 3)\ \rangle$
    [] $\mathtt{isEmpty}_{\mathrm{in}ROL}(prid) \equiv \langle\ \exists e :: (e \in ROL) \wedge (e \uparrow 1 = prid) \wedge (e \uparrow 3 = \emptyset)\ \rangle$

  initially
      $\lambda =$ Location($k$)
    [] $TAs = 0$ [] $rform = 0$ [] $\mathbf{T} = \perp$ [] $ROL = \perp$ [] $RL = \perp$ [] $\mathbb{T} = \emptyset$
    [] $t = \varepsilon$ [] $tuple = \varepsilon$ [] $\mathbf{t} = \emptyset$ [] $tuples = \emptyset$ [] $a = \varepsilon$ [] $atuple = \varepsilon$
    [] $\mathcal{Q}_{T^S_{a_k}} = \perp$ [] $\mathcal{Q}_{T^R_{a_k}} = \perp$ [] $UsrRdy4Evt = FALSE$

  assign
        {Migrate to different location}
    [] $\lambda :=$ Location(Move()) if ($type = mobile$)
        {Capture different parameters when user application is ready}
    [] $\langle$ $prType, prName, UsrRdy4Evt :=$ getPrimType(), getPrimName(), $FALSE$
         || $prRdIn, prBulk :=$ getPrimRDorIN(), getPrimBulk()
         || $tuple :=$ getTuple() if $((prRdIn = FALSE) \wedge (prBulk = FALSE))$
         || $tuples :=$ getTuples() if $((prRdIn = FALSE) \wedge (prBulk = TRUE))$
         || $atuple :=$ getAntiTuple() if $(prRdIn = TRUE)$
         || $TAs :=$ getTargetAgentCount() if $(prType = remote)$
         || $\langle \| \ a : 1 \le a \le TAs :: taids[a] :=$ getTargetAgentID($a$)$\rangle$ if $(prType = remote)$
         || $mode :=$ getMode(ONCE, ONCE/TUPLE) if $((prType = remote) \wedge (prRdIn = TRUE))$
       $\rangle$ if $(UsrRdy4Evt = TRUE)$
```

Fig. 3. Mobile UNITY **Program** *agent*(k): part 1

reactive statements as they have used "≈" notation. Agent behaviors, including functionalities of ATS, Local Operation Manager, Remote Operation Manager, ATS Reaction Manager etc. are contained in *agent*(k) as shown in Figure 3, Figure 4, and Figure 5. Similarly, functionalities of different components of host, including Transport Interface, Discovery Manager, Communication Manager, Host Server, Agent Manager etc., are contained in *host*(i) as shown in Figure 6, Figure 7, Figure 8, and Figure 9. However, in above formal system, many aspects of TSMM are not directly formalized, to keep this formal system simple.

Different variables related to hosts and agents are specified in this formal system. For instance, $\mathcal{Q}$ is used to express any queue used to define different

```
{- - - - - - - - - - Start of Local Operation Manager - - - - - - - - - -}
      {Perform different local tuple space primitives}
$[]$ $\langle$ $t, tuple, prType := tuple, \varepsilon, \varepsilon \parallel \mathtt{out}(t, \mathbf{T})$ $\rangle$ if $\big((prType = local) \wedge (prName = \mathsf{OUT}) \wedge \neg(tuple = \varepsilon)\big)$
$[]$ $\langle$ $\mathbf{t}, tuples, prType := tuples, \emptyset, \varepsilon \parallel \mathtt{outg}(\mathbf{t}, \mathbf{T})$
   $\rangle$ if $\big((prType = local) \wedge (prName = \mathsf{OUTG}) \wedge \neg(tuples = \emptyset)\big)$
$[]$ $\langle$ $a, atuple, prType := atuple, \varepsilon, \varepsilon$
     $\parallel \langle\ t := \mathtt{rdp}(a, \mathbf{T}) \parallel \mathtt{retTuple2Usr}(t)\ \rangle$     if $(prName = \mathsf{RDP})$
     $\parallel \langle\ \mathbf{t} := \mathtt{rdgp}(a, \mathbf{T}) \parallel \mathtt{retTuples2Usr}(\mathbf{t})\ \rangle$  if $(prName = \mathsf{RDGP})$
     $\parallel \langle\ t := \mathtt{inp}(a, \mathbf{T}) \parallel \mathtt{retTuple2Usr}(t)\ \rangle$     if $(prName = \mathsf{INP})$
     $\parallel \langle\ \mathbf{t} := \mathtt{ingp}(a, \mathbf{T}) \parallel \mathtt{retTuples2Usr}(\mathbf{t})\ \rangle$  if $(prName = \mathsf{INGP})$
   $\rangle$ if $\big((prType = local) \wedge \neg(atuple = \varepsilon)\big)$
{- - - - - - - - - - End of Local Operation Manager - - - - - - - - - -}
{- - - - - - - - - - Start of Remote Operation Manager - - - - - - - - - -}
      {Initiate (as reference agent) execution of different remote tuple space operations}
$[]$ $\langle$ $t, tuple, prType := tuple, \varepsilon, \varepsilon \parallel prid := \mathtt{getPrID}(prName) \parallel rform := 1$
     $\parallel \langle \parallel a : 1 \le a \le TAs :: \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r}}(rform, prid, prName, t, mode, aid, taids[a])\rangle$
   $\rangle$ if $\big((prType = remote) \wedge (prName = \mathsf{OUT}) \wedge \neg(tuple = \varepsilon)\big)$
$[]$ $\langle$ $\mathbf{t}, tuples, prType := tuples, \emptyset, \varepsilon \parallel prid := \mathtt{getPrID}(prName) \parallel rform := 1$
     $\parallel \langle \parallel a : 1 \le a \le TAs :: \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r}}(rform, prid, prName, \mathbf{t}, mode, aid, taids[a])\rangle$
   $\rangle$ if $\big((prType = remote) \wedge (prName = \mathsf{OUTG}) \wedge \neg(tuples = \emptyset)\big)$
$[]$ $\langle$ $a, atuple, prType := atuple, \varepsilon, \varepsilon \parallel prid := \mathtt{getPrID}(prName) \parallel rform := 1$
     $\parallel ROL := ROL \cup \{prid, prName, taids\}$
     $\parallel \langle \parallel a : 1 \le a \le TAs :: \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r}}(rform, prid, prName, a, mode, aid, taids[a])\rangle$
   $\rangle$ if $\big((prType = remote) \wedge (prRdIn = TRUE) \wedge \neg(atuple = \varepsilon)\big)$

$[]$ $\langle$ $r, \mathcal{Q}_{T^R_{a_k}} := \mathtt{head}(\mathcal{Q}_{T^R_{a_k}}), \mathtt{tail}(\mathcal{Q}_{T^R_{a_k}}) \parallel prid := r \uparrow \mathrm{prid}$
     $\parallel \langle\ \mathbb{T}_{prid} := \mathbb{T}_{prid} \cup \{r \uparrow \mathrm{tAid}, r \uparrow \mathrm{data}\} \parallel \langle\ \exists e : (e \in ROL) \wedge (e \uparrow 1 = prid) :: e \uparrow 3 := e \uparrow 3 \setminus r \uparrow \mathrm{tAid}\ \rangle$
       $\rangle$ if $\big((r \uparrow \mathrm{rAid} = aid) \wedge \mathtt{isPresent}_{\mathrm{in}ROL}(prid, r \uparrow \mathrm{tAid})\big)$     {Handling Response tuple}
   $\rangle$ if $\big(\neg(\mathcal{Q}_{T^R_{a_k}} = \perp) \wedge (\mathtt{head}(\mathcal{Q}_{T^R_{a_k}}) \uparrow \mathrm{rform} = 2)\big)$

      {Return result of execution of remote tuple-reading or -consuming operation to user}
$[]$ $\langle \parallel e : (e \in ROL) \wedge (e \uparrow 3 = \emptyset)$
       $::\ prid, prName := e \uparrow 1, e \uparrow 2 \parallel ROL := ROL \setminus e$
         $\parallel \langle\ \ \langle \parallel e : e \in \mathbb{T}_{prid} :: \mathbf{t} := \mathbf{t} \cup e \uparrow \mathrm{tuples}\ \rangle \parallel \mathtt{retTuples2Usr}(\mathbf{t})$
           $\parallel \langle \parallel e : e \in \mathbb{T}_{prid} \wedge \big((prName = \mathsf{ING}) \vee (prName = \mathsf{INGP})\big)$
                $:: \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r'}}(3, prid, prName, aid, e \uparrow \mathrm{tAid})\ \rangle$
         $\rangle$ if $\big((prName = \mathsf{RDG}) \vee (prName = \mathsf{RDGP}) \vee (prName = \mathsf{ING}) \vee (prName = \mathsf{INGP})\big)$
         $\parallel \langle\ \ \langle \parallel e : e = e'.(e' \in \mathbb{T}_{prid}) :: t, taid := e \uparrow \mathrm{tuple}, e \uparrow \mathrm{tAid}\ \rangle \parallel \mathtt{retTuple2Usr}(t)$
           $\parallel \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r'}}(3, prid, prName, aid, taid)$
                                                              if $\big((prName = \mathsf{IN}) \vee (prName = \mathsf{INP})\big)$
           $\parallel \langle \parallel e : e \in \mathbb{T}_{prid} \wedge \neg(e \uparrow \mathrm{tAid} = taid) \wedge \big((prName = \mathsf{IN}) \vee (prName = \mathsf{INP})\big)$
                $:: \mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{\mathsf{r'}}(4, prid, prName, e \uparrow \mathrm{tuple}, aid, e \uparrow \mathrm{tAid})\ \rangle$
         $\rangle$ if $\big((prName = \mathsf{RD}) \vee (prName = \mathsf{RDP}) \vee (prName = \mathsf{IN}) \vee (prName = \mathsf{INP})\big)$
   $\rangle$
{- - - - - - - - - - End of Remote Operation Manager - - - - - - - - - -}
```

Fig. 4. Mobile UNITY **Program** *agent*(k): part 2

activities of TSMM; its subscripts represent purpose of using it. Also, $\mathtt{head}(\mathcal{Q})$ returns front element of $\mathcal{Q}$, while $\mathtt{tail}(\mathcal{Q})$ returns all elements of $\mathcal{Q}$ except front element. Again, $\mathcal{Q} \bullet M$ inserts message M in the rear end of $\mathcal{Q}$ and returns updated $\mathcal{Q}$. M comprises of message identity mid, source host's identity src, destination

```
{- - - - - - - - - - Start of ATS Reaction Manager - - - - - - - - - -}

      {Complete execution of different remote tuple space operations}
  [] ⟨  $r, \mathcal{Q}_{T^R_{a_k}} := \mathtt{head}(\mathcal{Q}_{T^R_{a_k}}), \mathtt{tail}(\mathcal{Q}_{T^R_{a_k}})$ || $prid := r \uparrow$ prid || $prName := r \uparrow$ pName
      || $prBulk := TRUE$
                if $((prName = \mathsf{RDG}) \vee (prName = \mathsf{RDGP}) \vee (prName = \mathsf{ING}) \vee (prName = \mathsf{INGP}))$
           ~    $FALSE$
                if $((prName = \mathsf{RD}) \vee (prName = \mathsf{RDP}) \vee (prName = \mathsf{IN}) \vee (prName = \mathsf{INP}))$
      || ⟨  ⟨ $t := r \uparrow$ data || $\mathtt{out}(t, \mathbf{T})$ ⟩            if $(prName = \mathsf{OUT})$
         || ⟨ $\mathbf{t} := r \uparrow$ data || $\mathtt{outg}(\mathbf{t}, \mathbf{T})$ ⟩          if $(prName = \mathsf{OUTG})$
         || ⟨ $a := r \uparrow$ data || $t := \mathtt{rd}(a, \mathbf{T})$ ⟩        if $(prName = \mathsf{RD})$
         || ⟨ $a := r \uparrow$ data || $\mathbf{t} := \mathtt{rdg}(a, \mathbf{T})$ ⟩      if $(prName = \mathsf{RDG})$
         || ⟨ $a := r \uparrow$ data || $t := \mathtt{rdp}(a, \mathbf{T})$ ⟩      if $(prName = \mathsf{RDP})$
         || ⟨ $a := r \uparrow$ data || $\mathbf{t} := \mathtt{rdgp}(a, \mathbf{T})$ ⟩    if $(prName = \mathsf{RDGP})$
         || ⟨ $a := r \uparrow$ data || $t := \mathtt{in}(a, \mathbf{T})$ ⟩        if $(prName = \mathsf{IN})$
         || ⟨ $a := r \uparrow$ data || $\mathbf{t} := \mathtt{ing}(a, \mathbf{T})$ ⟩      if $(prName = \mathsf{ING})$
         || ⟨ $a := r \uparrow$ data || $t := \mathtt{inp}(a, \mathbf{T})$ ⟩      if $(prName = \mathsf{INP})$
         || ⟨ $a := r \uparrow$ data || $\mathbf{t} := \mathtt{ingp}(a, \mathbf{T})$ ⟩    if $(prName = \mathsf{INGP})$
         || $rform := 2$
         || $\mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{r'}(rform, prid, prName, t, aid, r \uparrow \mathrm{rAid})$   if $(prBulk = FALSE)$
         || $\mathcal{Q}_{T^S_{a_k}} := \mathcal{Q}_{T^S_{a_k}} \bullet \mathtt{createRTuple}_{r'}(rform, prid, prName, \mathbf{t}, aid, r \uparrow \mathrm{rAid})$   if $(prBulk = TRUE)$
        ⟩  if $((r \uparrow \mathrm{tAid} = aid) \wedge (r \uparrow \mathrm{rform} = 1))$       {Handling Reaction tuple}
      || ⟨ $t := r \uparrow$ data || $\mathtt{out}(t, \mathbf{T})$
        ⟩  if $((r \uparrow \mathrm{tAid} = aid) \wedge (r \uparrow \mathrm{rform} = 4))$       {Handling NACK tuple}
     ⟩  if $(\neg(\mathcal{Q}_{T^R_{a_k}} = \perp) \wedge$
              $((\mathtt{head}(\mathcal{Q}_{T^R_{a_k}}) \uparrow \mathrm{rform} = 1) \vee (\mathtt{head}(\mathcal{Q}_{T^R_{a_k}}) \uparrow \mathrm{rform} = 3) \vee (\mathtt{head}(\mathcal{Q}_{T^R_{a_k}}) \uparrow \mathrm{rform} = 4)))$

{- - - - - - - - - - End of ATS Reaction Manager - - - - - - - - - -}

      {Discard messages destined for other agents}
  [] $\mathcal{Q}_{T^R_{a_k}} := \mathtt{tail}(\mathcal{Q}_{T^R_{a_k}})$   if $(\neg(\mathcal{Q}_{T^R_{a_k}} = \perp) \wedge \neg(\mathtt{head}(\mathcal{Q}_{T^R_{a_k}}) \uparrow \mathrm{dstAg} = aid))$

end
```

Fig. 5. Mobile UNITY **Program** *agent*(*k*): part 3

host's identity *dest*, type of message *kind*, data encapsulated within it *data*, and network interface, *ni*, through which M will be transmitted. M is generated by calling `newMsg`$(src, dest, kind, data, ni)$, which inserts its *mid* to return a complete message. Possible types of messages included in these specifications are BCON, RT, ACK, Locate, and Found messages.

3.1 Formalization of *agent*(*k*)

Each agent is represented by program *agent*(*k*), which comprises of ***declare***, ***always***, ***initially*** and ***assign*** sections. Agent behavior is specified by different variables that are declared in ***declare***, like *aid* and *type* as agent identity and nature of *agent*(*k*). **T** is declared as `ATS` of *agent*(*k*), and *prid* as identity of invoked primitive of *agent*(*k*). *ROL* is declared as remote operation list of *agent*(*k*), and *RL* is declared as reactive list of *agent*(*k*). $\mathcal{Q}_{T^S_{a_k}}$ and $\mathcal{Q}_{T^R_{a_k}}$ are declared as queues to interface between agents and their supported hosts, and transfer request/response

```
Program $host(i)$ at $\lambda$

  declare
      $type$ : $\in \{stationary, mobile, accesspoint\}$
    [] $hid$ : hostid [] $assoc$ : set of hostid
    [] $nwdeploy$ : $\in \{iBSS, IBSS\}$ [] $status$ : $\in \{standalone, connected, associated\}$
    [] $\mathbf{T}'$ : tuple space
    [] $\mathcal{Q}_{T^S_{a_k}}, \mathcal{Q}_{T^R_{a_k}}$ : queue of $\mathrm{RT}_{\mathrm{tuple}}$ [] $\mathcal{Q}_{RT_S}, \mathcal{Q}_{RT_R}$ : queue of $\mathrm{RT}_{\mathrm{tuple}}$ [] $r$ : $\mathrm{RT}_{\mathrm{tuple}}$
    [] $a$ : agentid [] $\mathcal{A}$ : set of agentid [] $\mathcal{Q}_{in}, \mathcal{Q}_{out}$ : queue of agentid
    [] $\mathcal{H}$ : set of $(\mathrm{MH}_{\mathrm{hostid}}, \mathrm{AP}_{\mathrm{hostid}}, \mathrm{timestamp})$ [] $\mathcal{L}$ : set of $(\mathrm{MH}_{\mathrm{hostid}}, \mathrm{RT}_{\mathrm{tuple}}, \mathrm{timestamp})$
    [] $\mathcal{CS}$ : message [] $\mathcal{T}_{\mathrm{W}}, \mathcal{T}_{\mathrm{WL}}$ : message [] $M, m$ : message
    [] $\mathcal{L}RT$ : set of $(\mathrm{AP}_{\mathrm{hostid}}/\mathrm{MH}_{\mathrm{hostid}}, \mathrm{RT}_{\mathrm{msgid}})$
    [] $\mathcal{N}$ : set of $(\mathrm{Host}_{\mathrm{hostid}}, \text{set of agentid}, \mathrm{timestamp}, \mathrm{extant})$
    [] $\mathcal{Q}_{S_B}, \mathcal{Q}_{R_B}$ : queue of message [] $\mathcal{Q}_{S_{RT}}, \mathcal{Q}_{R_{RT}}$ : queue of message
    [] $\mathcal{Q}_{S_{\mathrm{W}}}, \mathcal{Q}_{S_{\mathrm{WL}}}$ : queue of message [] $\mathcal{Q}_S, \mathcal{Q}_R$ : queue of message
    [] $clock, lastHTSchk, lastRTsent, lastBsent, newRTGap, rtAtmpt$ : natural

  always
      $B_{iBSS_{\mathrm{W}}} = \textbf{IBSSBROADCASTADDRESS}_{\mathbf{DS}}$ [] $B_{iBSS_{\mathrm{WL}}} = \textbf{IBSSBROADCASTADDRESS}_{\mathbf{BSA}}$
    [] $B_{IBSS_{\mathrm{WL}}} = \textbf{IBSSBROADCASTADDRESS}$
    [] $\lambda := \texttt{Location}(i)$ [] $hid := \texttt{getMyHostID}(i)$
    [] $type := \texttt{getHostType}(stationary, mobile, accesspoint)$
    [] $nwdeploy := \texttt{getNetwork}(iBSS, IBSS)$
    [] $mhGap = \textsc{SystemMHValidityInterval}$ [] $HTSaccessGap = \textsc{SystemHTSAccessInterval}$
    [] $locateGap = \textsc{SystemLocateMsgInterval}$ [] $bconGap = \textsc{SystemBeaconInterval}$
    [] $mhGap = \textsc{SystemMHValidityInterval}$ [] $bLife = \textsc{SystemBeaconLifetime}$
    [] $\texttt{isPresent}_{\mathcal{H}}(mhid) \equiv \langle\ \exists e : (e \in \mathcal{H}) \wedge (e \uparrow 1 = mhid)\ \rangle$
    [] $\texttt{isPresent}_{\mathcal{L}}(mhid) \equiv \langle\ \exists e : (e \in \mathcal{L}) \wedge (e \uparrow 1 = mhid)\ \rangle$
    [] $\texttt{isPresent}_{\mathcal{N}}(hostid) \equiv \langle\ \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = hostid)\ \rangle$
    [] $\texttt{isPresent}_{\mathcal{L}RT}(hostid) \equiv \langle\ \exists e : (e \in \mathcal{L}RT) \wedge (e \uparrow 1 = hostid)\ \rangle$
    [] $\texttt{isRepeat}_{\mathcal{L}RT}(hostid, msgid) \equiv \langle\ \exists e : (e \in \mathcal{L}RT) \wedge (e \uparrow 1 = hostid) \wedge (e \uparrow 2 = msgid)\ \rangle$
    [] $\texttt{isValid}_{\mathcal{H}}(e, now) \equiv \big((e \in \mathcal{H}) \wedge ((now - e \uparrow 3) \leq mhGap)\big)$
    [] $\texttt{isValid}_{\mathcal{L}}(e, now) \equiv \big((e \in \mathcal{L}) \wedge \big((now - e \uparrow 3) \leq locateGap\big)\big)$
    [] $\texttt{isValid}_{\mathcal{N}}(e, now) \equiv \big((e \in \mathcal{N}) \wedge \big((now - e \uparrow 3) \leq e \uparrow 4\big)\big)$
    [] $\texttt{isMsgBcon}(msg) \equiv (msg \cdot kind = Beacon)$
    [] $\texttt{isMsgRT}(msg) \equiv (msg \cdot kind = RT)$ [] $\texttt{isMsgACK}(msg) \equiv (msg \cdot kind = ACK)$
    [] $\texttt{isMsgLocate}(msg) \equiv (msg \cdot kind = Locate)$ [] $\texttt{isMsgFound}(msg) \equiv (msg \cdot kind = Found)$
    [] $\texttt{isNotOwnMsg}(msg) \equiv \neg(msg \cdot src = hid)$
    [] $\texttt{isSH}(host) \equiv (host \cdot type = stationary)$ [] $\texttt{isMH}(host) \equiv (host \cdot type = mobile)$
    [] $\texttt{isAP}(host) \equiv (host \cdot type = accesspoint)$
```

Fig. 6. Mobile UNITY **Program** $host(i)$: part 1

tuples from agents to hosts and vice versa. When user application is invoking any tuple space operation, corresponding agent captures different parameters required to complete that operation.

3.2 Formalization of *host*(*i*)

Like $agent(k)$, $host(i)$ also comprises of ***declare***, ***always***, ***initially*** and ***assign*** sections. Different variables related to host behavior is declared in ***declare*** section,

```
initially
    $clock = 0$ [] $lastHTSchk = 0$ [] $lastRTsent = 0$ [] $lastBsent = 0$
  [] $status = standalone$ [] $assoc = \emptyset$ [] $\mathcal{H} = \emptyset$ [] $\mathcal{L} = \emptyset$ [] $\mathcal{LRT} = \emptyset$ [] $\mathcal{A} = \emptyset$ [] $\mathcal{N} = \emptyset$
  [] $\mathsf{T}' =\perp$ [] $\mathcal{T}_{\mathrm{W}} =\perp$ [] $\mathcal{T}_{\mathrm{WL}} =\perp$ [] $\mathcal{CS} =\perp$
  [] $\mathcal{Q}_{T^S_{a_k}} =\perp$ [] $\mathcal{Q}_{T^R_{a_k}} =\perp$ [] $\mathcal{Q}_{in} =\perp$ [] $\mathcal{Q}_{out} =\perp$ [] $\mathcal{Q}_{RT_S} =\perp$ [] $\mathcal{Q}_{RT_R} =\perp$
  [] $\mathcal{Q}_{S_B} =\perp$ [] $\mathcal{Q}_{R_B} =\perp$ [] $\mathcal{Q}_{S_{RT}} =\perp$ [] $\mathcal{Q}_{R_{RT}} =\perp$ [] $\mathcal{Q}_{S_{\mathrm{W}}} =\perp$ [] $\mathcal{Q}_{S_{\mathrm{WL}}} =\perp$ [] $\mathcal{Q}_S =\perp$ [] $\mathcal{Q}_R =\perp$

assign
      {Increment the clock}
  [] $clock := clock + 1$

  {- - - - - - - - - - Start of Transport Interface - - - - - - - - - -}

      {Organize a message for onward transmission}
  [] ⟨   $M, \mathcal{Q}_S := \mathtt{head}(\mathcal{Q}_S), \mathtt{tail}(\mathcal{Q}_S)$
       || ⟨ $\mathcal{Q}_{S_{\mathrm{W}}} := \mathcal{Q}_{S_{\mathrm{W}}} \bullet M$   if $(M{\cdot}ni = \mathrm{W})$  ||  $\mathcal{Q}_{S_{\mathrm{WL}}} := \mathcal{Q}_{S_{\mathrm{WL}}} \bullet M$   if $(M{\cdot}ni = \mathrm{WL})$ ⟩
     ⟩ if $\neg(\mathcal{Q}_S =\perp)$

      {Transfer a message from $\mathcal{Q}_{S_{\mathrm{W}}}$ to $\mathcal{T}_{\mathrm{W}}$; make $\mathcal{T}_{\mathrm{W}}$ empty after some time}
  [] $transmit\&reset_{\mathrm{W}}$ :: ⟨ $\mathcal{T}_{\mathrm{W}}, \mathcal{Q}_{S_{\mathrm{W}}} := \mathtt{head}(\mathcal{Q}_{S_{\mathrm{W}}}), \mathtt{tail}(\mathcal{Q}_{S_{\mathrm{W}}})$   if $\neg(\mathcal{Q}_{S_{\mathrm{W}}} =\perp) \wedge (\mathcal{T}_{\mathrm{W}} =\perp)$ ;
                               $\mathcal{T}_{\mathrm{W}} :=\perp$ ⟩

      {Transfer a message from $\mathcal{Q}_{S_{\mathrm{WL}}}$ to $\mathcal{T}_{\mathrm{WL}}$; make $\mathcal{T}_{\mathrm{WL}}$ empty after some time}
  [] $transmit\&reset_{\mathrm{WL}}$ :: ⟨ $\mathcal{T}_{\mathrm{WL}}, \mathcal{Q}_{S_{\mathrm{WL}}} := \mathtt{head}(\mathcal{Q}_{S_{\mathrm{WL}}}), \mathtt{tail}(\mathcal{Q}_{S_{\mathrm{WL}}})$   if $\neg(\mathcal{Q}_{S_{\mathrm{WL}}} =\perp) \wedge (\mathcal{T}_{\mathrm{WL}} =\perp)$ ;
                                $\mathcal{T}_{\mathrm{WL}} :=\perp$ ⟩

      {Transfer a message from $\mathcal{T}_{\mathrm{W}}$ to $\mathcal{Q}_R$}
  [] ⟨ $\mathcal{Q}_R := \mathcal{Q}_R \bullet \mathcal{T}_{\mathrm{W}}$   if $\mathtt{isNotOwnMsg}(\mathcal{T}_{\mathrm{W}})$ ⟩   reacts-to $\neg(\mathcal{T}_{\mathrm{W}} =\perp)$

      {Transfer a message from $\mathcal{T}_{\mathrm{WL}}$ to $\mathcal{Q}_R$}
  [] ⟨ $\mathcal{Q}_R := \mathcal{Q}_R \bullet \mathcal{T}_{\mathrm{WL}}$   if $\mathtt{isNotOwnMsg}(\mathcal{T}_{\mathrm{WL}})$ ⟩   reacts-to $\neg(\mathcal{T}_{\mathrm{WL}} =\perp)$

      {Organize a received Beacon/RT/ACK/Locate/Found message for further processing}
  [] ⟨   $M, \mathcal{Q}_R := \mathtt{head}(\mathcal{Q}_R), \mathtt{tail}(\mathcal{Q}_R)$
       || ⟨  $\mathcal{Q}_{R_B} := \mathcal{Q}_{R_B} \bullet M$   if $\mathtt{isMsgBcon}(M)$
         || $\mathcal{Q}_{R_{RT}} := \mathcal{Q}_{R_{RT}} \bullet M$   if $\mathtt{isMsgRT}(M) \vee \mathtt{isMsgACK}(M) \vee \mathtt{isMsgLocate}(M) \vee \mathtt{isMsgFound}(M)$
         ⟩
     ⟩ if $\neg(\mathcal{Q}_R =\perp)$
  {- - - - - - - - - - End of Transport Interface - - - - - - - - - -}
```

Fig. 7. Mobile UNITY **Program** *host*(i): part 2

like *hid* as host identity of *host*(i) and *type* as nature of *host*(i). T' is declared as its `HTS`. $\mathcal{H}$ and $\mathcal{L}$ are declared for `History` (that records path of successful data transfer to different mobile hosts) and location list (that lists mobile hosts with ongoing location search) respectively for *host*(i) of stationary hosts and access points in iBSS. Moreover, $\mathcal{LRT}$ and $\mathcal{CS}$ are declared for `LastRT` (that records message identity of last data messages received from different hosts) and `CommStash` (that buffers data messages) respectively of *host*(i) of mobile hosts and access points in both iBSS and IBSS. Also, $\mathcal{N}$ and $\mathcal{A}$ are declared to represent `NeighborList` and `AgentList` respectively. Different macros related to various aspects of discovery and communication mechanisms are included in appendix.

At lowest level, TSMM interacts with transport service, which is formalized as Transport Interface by a set of assignment statements. Discovery Manager and Communication Manager interchange messages with Transport Interface through $\mathcal{Q}_S$

```
{- - - - - - - - - - Start of Discovery Manager - - - - - - - - - -}
    {Prepare to send Beacon message to destination}
[] ⟨  $\mathcal{Q}_{S_B}, lastBsent := \mathcal{Q}_{S_B} \bullet \texttt{discSend}_{\mathrm{W}_{iBSS}}(), clock$   if $\big(\texttt{isSH}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\mathcal{Q}_{S_B}, lastBsent := \mathcal{Q}_{S_B} \bullet \texttt{discSend}_{\mathrm{WL}_{iBSS}}(), clock$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\mathcal{Q}_{S_B}, lastBsent := \big(\mathcal{Q}_{S_B} \bullet \texttt{discSend}_{\mathrm{W}_{iBSS}}()\big) \bullet \texttt{discSend}_{\mathrm{WL}_{iBSS}}(), clock$
                                                    if $\big(\texttt{isAP}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\mathcal{Q}_{S_B}, lastBsent := \mathcal{Q}_{S_B} \bullet \texttt{discSend}_{\mathrm{WL}_{IBSS}}(), clock$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = IBSS)\big)$
  ⟩ if $((clock - lastBsent) > bconGap)$

    {Process received Beacon message}
[] ⟨  $\texttt{discRcv}_{\mathrm{SH}_{iBSS}}(\mathcal{Q}_{R_B})$   if $\big(\texttt{isSH}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\texttt{discRcv}_{\mathrm{MH}_{iBSS}}(\mathcal{Q}_{R_B})$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\texttt{discRcv}_{\mathrm{AP}_{iBSS}}(\mathcal{Q}_{R_B})$   if $\big(\texttt{isAP}(hid) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\texttt{discRcv}_{\mathrm{MH}_{IBSS}}(\mathcal{Q}_{R_B})$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = IBSS)\big)$
  ⟩ if $\neg(\mathcal{Q}_{R_B} = \perp)$

    {Remove expired entries from $\mathcal{N}$}
[] ⟨  $\texttt{discValid}_{\mathcal{N}_{iBSS}}()$   if $\big((\texttt{isSH}(hid) \vee \texttt{isMH}(hid) \vee \texttt{isAP}(hid)) \wedge (nwdeploy = iBSS)\big)$
   ‖ $\texttt{discValid}_{\mathcal{N}_{IBSS}}()$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = IBSS)\big)$
  ⟩

    {Update $assoc$ on account of change in associated AP of MH}
[] ⟨ $\texttt{discUpdt}_{\mathrm{MH}_{iBSS}}()$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = iBSS)\big)$
  ⟩ if $(\neg\texttt{isPresent}_{\mathcal{N}}(assoc[0]) \vee \neg\texttt{isValid}_{\mathcal{N}}(\langle \exists e : e \uparrow 1 = assoc[0] :: e \rangle, clock))$

    {Update $status$ on account of change in connectivity of MH}
[] $\texttt{discUpdt}_{\mathrm{MH}_{IBSS}}()$   if $\big(\texttt{isMH}(hid) \wedge (nwdeploy = IBSS)\big)$

    {Organize a Beacon message for onward transmission}
[] ⟨ $\mathcal{Q}_S, \mathcal{Q}_{S_B} := \mathcal{Q}_S \bullet \texttt{head}(\mathcal{Q}_{S_B}), \texttt{tail}(\mathcal{Q}_{S_B})$ ⟩   if $\neg(\mathcal{Q}_{S_B} = \perp)$
{- - - - - - - - - - End of Discovery Manager - - - - - - - - - -}

{- - - - - - - - - - Start of Host Server - - - - - - - - - -}
    {Process received RT from different agents}
[] ⟨[] $k$ :: ⟨ $r, \mathcal{Q}_{T^S_{a_k}} := \texttt{head}(\mathcal{Q}_{T^S_{a_k}}), \texttt{tail}(\mathcal{Q}_{T^S_{a_k}}) \parallel \texttt{inject}(r, \mathbf{T}')$ ⟩   if $\neg(\mathcal{Q}_{T^S_{a_k}} = \perp)$ ⟩

    {Process received RT from COMMUNICATION module}
[] ⟨ $r, \mathcal{Q}_{RT_R} := \texttt{head}(\mathcal{Q}_{RT_R}), \texttt{tail}(\mathcal{Q}_{RT_R}) \parallel \texttt{inject}(r, \mathbf{T}')$ ⟩   if $\neg(\mathcal{Q}_{RT_R} = \perp)$

    {Periodically extract RT from HTS for onward transfer to target agents in same/different hosts}
[] ⟨  $\langle \parallel a : a \in \mathcal{A} :: r := \texttt{eject}(a, \mathbf{T}') \parallel \langle \mathcal{Q}_{T^R_a} := \mathcal{Q}_{T^R_a} \bullet r$   if $\neg(r = \varepsilon)\rangle\ \rangle$
   $\parallel \langle \parallel e : (e \in \mathcal{N}) \wedge (\mathrm{A} = e \uparrow 2) :: \langle \parallel a : a \in \mathrm{A} :: r := \texttt{eject}(a, \mathbf{T}') \parallel \langle \mathcal{Q}_{RT_S} := \mathcal{Q}_{RT_S} \bullet r$  if $\neg(r = \varepsilon)\rangle\ \rangle\ \rangle$
   $\parallel lastHTSchk := clock$
  ⟩ if $(clock - lastHTSchk > HTSaccessGap)$
{- - - - - - - - - - End of Host Server - - - - - - - - - -}
```

Fig. 8. Mobile UNITY **Program** *host*(i): part 3

and $\mathcal{Q}_R$. Some behaviors of Discovery Manager and Communication Manager are abstracted as macros, which are used in different assignment statements to fulfill all functionalities of Discovery Manager and Communication Manager. Host Server interchanges request/response tuples (represented as $\mathrm{RT}_{\mathrm{tuple}}$) with Communication Manager through $\mathcal{Q}_{RT_S}$ and $\mathcal{Q}_{RT_R}$, which is formalized via a set of assignment statements. Similarly, a pair of assignment statements formalizes registration/deregistration functionalities of Agent Manager.

```
{- - - - - - - - - - Start of Communication Manager - - - - - - - - - -}

      {Prepare to send RT/Locate message to destination}
  [] ⟨   $\mathtt{commSend}_{\mathrm{SH}_{iBSS}}(\mathcal{Q}_{RT_S})$  if $(\mathtt{isSH}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commSend}_{\mathrm{MH}_{iBSS}}(\mathcal{Q}_{RT_S})$  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commSend}_{\mathrm{AP}_{iBSS}}(\mathcal{Q}_{RT_S})$  if $(\mathtt{isAP}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commSend}_{\mathrm{MH}_{IBSS}}(\mathcal{Q}_{RT_S})$  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = IBSS))$
     ⟩ if $\neg(\mathcal{Q}_{RT_S} = \perp)$

      {Process received RT/Locate/Found message, and prepare to send RT/ACK/Found message}
  [] ⟨   $\mathtt{commRcv}_{\mathrm{SH}_{iBSS}}(\mathcal{Q}_{R_{RT}})$  if $(\mathtt{isSH}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commRcv}_{\mathrm{MH}_{iBSS}}(\mathcal{Q}_{R_{RT}})$  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commRcv}_{\mathrm{AP}_{iBSS}}(\mathcal{Q}_{R_{RT}})$  if $(\mathtt{isAP}(hid) \wedge (nwdeploy = iBSS))$
     || $\mathtt{commRcv}_{\mathrm{MH}_{IBSS}}(\mathcal{Q}_{R_{RT}})$  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = IBSS))$
     ⟩ if $\neg(\mathcal{Q}_{R_{RT}} = \perp)$

      {Resend RT message whose ACK fails to reach before timeout}
  [] ⟨   $\mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \mathsf{commReSend}_{\mathsf{RT}_{iBSS}}()$  if $((\mathtt{isMH}(hid) \vee \mathtt{isAP}(hid)) \wedge (nwdeploy = iBSS))$
     || $\mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \mathsf{commReSend}_{\mathsf{RT}_{IBSS}}()$  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = IBSS))$
     ⟩ if $((clock - lastRTsent) > newRTGap)$

      {Process RT message whose destination is presently not available}
  [] ⟨   ⟨ $\mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet \mathcal{CS}{\cdot}\, data$ || $\mathcal{CS} := \perp$ ⟩  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = iBSS))$
     || ⟨  $\mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \mathtt{newMsg}(hid, B_{iBSS_{\mathrm{W}}}, Locate, \mathcal{CS}{\cdot}\, dest, \mathrm{W})$
          || $\mathcal{L} := \mathcal{L} \cup \{(\mathcal{CS}{\cdot}\, dest, \mathcal{CS}{\cdot}\, data, clock)\}$ ⟩  if $(\mathtt{isAP}(hid) \wedge (nwdeploy = iBSS))$
     || ⟨ $\mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet \mathcal{CS}{\cdot}\, data$ || $\mathcal{CS} := \perp$ ⟩  if $(\mathtt{isMH}(hid) \wedge (nwdeploy = IBSS))$
     ⟩ if $(\neg(\mathcal{CS} = \perp) \wedge (rtAtmpt > 3))$

      {Remove expired entries from $\mathcal{H}$ and $\mathcal{L}$, and preserve unsent RT}
  [] $\mathtt{commValid}_{\mathcal{HL}_{iBSS}}()$  if $((\mathtt{isSH}(hid) \vee \mathtt{isAP}(hid)) \wedge (nwdeploy = iBSS))$

      {Organize RT/ACK/Locate/Found message for onward transmission}
  [] ⟨ $\mathcal{Q}_S, \mathcal{Q}_{S_{RT}} := \mathcal{Q}_S \bullet \mathtt{head}(\mathcal{Q}_{S_{RT}}), \mathtt{tail}(\mathcal{Q}_{S_{RT}})$ ⟩  if $\neg(\mathcal{Q}_{S_{RT}} = \perp)$

{- - - - - - - - - - End of Communication Manager - - - - - - - - - -}

{- - - - - - - - - - Start of Agent Manager - - - - - - - - - -}

      {Register active agents in $\mathcal{A}$}
  [] $\mathcal{A}, \mathcal{Q}_{in} := \mathcal{A} \cup \mathtt{head}(\mathcal{Q}_{in}), \mathtt{tail}(\mathcal{Q}_{in})$  if $\neg(\mathcal{Q}_{in} = \perp)$

      {Deregister terminated/migrated agents from $\mathcal{A}$}
  [] $\mathcal{A}, \mathcal{Q}_{out} := \mathcal{A} \setminus \mathtt{head}(\mathcal{Q}_{out}), \mathtt{tail}(\mathcal{Q}_{out})$  if $(\neg(\mathcal{Q}_{out} = \perp) \wedge (\mathtt{head}(\mathcal{Q}_{out}) \in \mathcal{A}))$

{- - - - - - - - - - End of Agent Manager - - - - - - - - - -}

end
```

Fig. 9. Mobile UNITY **Program** $host(i)$: part 4

4 Conclusion

This paper has proposed an approach of formalization of a TSMM, which decouples coordination among interacting agents of supported applications when deployed over multiple heterogeneous mobile, dynamic and unreliable networks. Proposed approach formally specifies TSMM as a Mobile UNITY system, comprising of components representing different behaviors of agents and hosts of TSMM. This formalization can be shown to reason TSMM as an appropriate coordination platform for multiple underlying heterogeneous networks, which facilitates development of robust and flexible mobile computing applications.

References

1. Bruneo, D., Puliafito, A., Scarpa, M.: Mobile Middleware: Definition and Motivations. In: Bellavista, P., Corradi, A. (eds.) The Handbook of Mobile Middleware, pp. 145–167. Auerbach Pub. (2007)
2. Gelernter, D.: Generative Communication in Linda. Transactions on Programming Languages and Systems 7(1), 80–112 (1985)
3. Eugster, P.T., Felber, P.A., Guerraoui, R., Kermarrec, A.M.: The many faces of Publish/Subscribe. Computing Surveys 35(2), 114–131 (2003)
4. De, S., Nandi, S., Goswami, D.: Architectures of Mobile Middleware: A Taxonomic Perspective. In: Proc. 2nd IEEE Intl. Conf. on Parallel, Distributed and Grid Computing, PDGC 2012 (December 2012)
5. De, S., Nandi, S., Goswami, D.: Modeling an Enhanced Tuple Space based Mobile Middleware in UNITY. In: Proc. 11th IEEE Intl. Conf. on Ubiquitous Computing and Communications, IUCC 2012, pp. 1684–1691 (June 2012)
6. De, S., Goswami, D., Nandi, S., Chakraborty, S.: Formalization of a Fully-Decoupled Reactive Tuple Space Model for Mobile Middleware. In: Borcea, C., Bellavista, P., Gianelli, C., Magedanz, T., Schreiner, F. (eds.) Mobilware 2012. LNICST, vol. 65, pp. 77–91. Springer, Heidelberg (2013)
7. Roman, G.C., McCann, P.J., Plun, J.Y.: Mobile UNITY: Reasoning and Specification in Mobile Computing. Transactions on Software Engineering and Methodology 6(3), 250–282 (1997)
8. Murphy, A.L., Picco, G.P., Roman, G.C.: Lime: A Coordination Model and Middleware supporting Mobility of Hosts and Agents. Transactions on Software Engineering and Methodology 15(3), 279–328 (2006)
9. Roman, G.C., Payton, J.: Mobile UNITY Schemas for Agent Coordination. In: Börger, E., Gargantini, A., Riccobene, E. (eds.) ASM 2003. LNCS, vol. 2589, pp. 126–150. Springer, Heidelberg (2003)
10. IEEE 802.11 WG Std.: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Technical Report 802.11$^{\text{TM}}$ (June 2007)
11. De, S., Nandi, S., Goswami, D.: On Performance Improvement Issues in Unordered Tuple Space based Mobile Middleware. In: Proc. 2010 Annual IEEE India Conf., INDICON 2010 (December 2010)
12. De, S., Goswami, D., Nandi, S.: A New Tuple Space Structure for Tuple Space based Mobile Middleware Platforms. In: Proc. 2012 Annual IEEE India Conf., INDICON 2012 (December 2012)
13. De, S., Chakraborty, S., Nandi, S., Goswami, D.: Supporting Tuple Space based Mobile Middleware over Unreliable Mobile Infrastructures: Design and Formal Specifications. In: Proc. 6th IEEE Intl. Conf. on Advanced Networks and Telecommunications Systems, ANTS 2012 (December 2012)

A Appendix: Macros Related to Formalization of TSMM

A.1 Macros of Discovery Manager

$$M := \texttt{discSend}_{\mathrm{W}_{iBSS}}() \triangleq \langle\ M := \texttt{newMsg}(hid,\, B_{iBSS_{\mathrm{W}}},\, BCON,\, \texttt{buildBcon}(\mathcal{A},\, bLife),\, \mathrm{W})\ \rangle$$

$$M := \texttt{discSend}_{\mathrm{WL}_{iBSS}}() \triangleq \langle\ M := \texttt{newMsg}(hid,\, B_{iBSS_{\mathrm{WL}}},\, BCON,\, \texttt{buildBcon}(\mathcal{A},\, bLife),\, \mathrm{WL})\ \rangle$$

$$M := \texttt{discSend}_{\mathrm{WL}_{IBSS}}() \triangleq \langle\ M := \texttt{newMsg}(hid,\, B_{IBSS_{\mathrm{WL}}},\, BCON,\, \texttt{buildBcon}(\mathcal{A},\, bLife),\, \mathrm{WL})\ \rangle$$

```
$\mathtt{discRcv}_{\mathrm{SH}_{iBSS}}(\mathcal{Q}_{R_B}) \triangleq$
  $\langle \;\; m, \mathcal{Q}_{R_B} := \mathtt{head}(\mathcal{Q}_{R_B}), \mathtt{tail}(\mathcal{Q}_{R_B})$
   $\| \; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
       $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
          $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
     $\rangle$   if $\big(\mathtt{isSH}(m{\cdot}\, src) \vee \mathtt{isAP}(m{\cdot}\, src)\big)$
  $\rangle$
$\mathtt{discRcv}_{\mathrm{MH}_{iBSS}}(\mathcal{Q}_{R_B}) \triangleq$
  $\langle \;\; m, \mathcal{Q}_{R_B} := \mathtt{head}(\mathcal{Q}_{R_B}), \mathtt{tail}(\mathcal{Q}_{R_B})$
   $\| \; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
      $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
         $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
      $\| \; status, assoc[0] := associated, m{\cdot}\, src$
     $\rangle$   if $\big(\mathtt{isAP}(m{\cdot}\, src) \wedge (status = standalone)\big)$
   $\| \; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
      $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
         $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
      $\| \; assoc[0] := m{\cdot}\, src$
     $\rangle$   if $\big(\mathtt{isAP}(m{\cdot}\, src) \wedge (status = associated)\big)$
  $\rangle$
$\mathtt{discRcv}_{\mathrm{AP}_{iBSS}}(\mathcal{Q}_{R_B}) \triangleq$
  $\langle \;\;\; m, \mathcal{Q}_{R_B} := \mathtt{head}(\mathcal{Q}_{R_B}), \mathtt{tail}(\mathcal{Q}_{R_B})$
    $\| \;\; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
           $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\rangle$   if $\big(\mathtt{isSH}(m{\cdot}\, src) \vee \mathtt{isAP}(m{\cdot}\, src)\big)$
    $\| \;\; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
           $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\| \; assoc := assoc \cup \{m{\cdot}\, src\}$
        $\rangle$   if $\mathtt{isMH}(m{\cdot}\, src)$
  $\rangle$
$\mathtt{discUpdt}_{\mathrm{MH}_{iBSS}}() \triangleq \langle \;\;\; assoc[0] := \perp \;\;\; \| \;\;\; \langle assoc[0] := a.(a \in \mathcal{N})$   if $\neg(\mathcal{N} = \emptyset)\rangle$
                                        $\| \; status := standalone$   if $(assoc[0] = \perp)$   $\sim$   $associated$   if $\neg(assoc[0] = \perp)$
                                       $\rangle$
$\mathtt{discValid}_{\mathcal{N}_{iBSS}}() \triangleq \langle \| \; e : e \in \mathcal{N} :: \mathcal{N} := \mathcal{N} \setminus \{e\}$   if $\neg\mathtt{isValid}_{\mathcal{N}}(e, clock) \; \rangle$
$\mathtt{discUpdt}_{\mathrm{MH}_{IBSS}}() \triangleq \langle \;\;\; status := standalone$   if $(\mathcal{N} = \emptyset)$   $\sim$   $connected$   if $\neg(\mathcal{N} = \emptyset)$   $\rangle$
$\mathtt{discValid}_{\mathcal{N}_{IBSS}}() \triangleq \langle \| \; e : e \in \mathcal{N} :: \mathcal{N} := \mathcal{N} \setminus \{e\}$   if $\neg\mathtt{isValid}_{\mathcal{N}}(e, clock) \; \rangle$
                                                                                                                        (1)
```

```
$\mathtt{discRcv}_{\mathrm{MH}_{IBSS}}(\mathcal{Q}_{R_B}) \triangleq$
  $\langle \;\;\; m, \mathcal{Q}_{R_B} := \mathtt{head}(\mathcal{Q}_{R_B}), \mathtt{tail}(\mathcal{Q}_{R_B})$
    $\| \;\; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
              $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
           $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\| \; status := connected$
        $\rangle$   if $\big(\mathtt{isMH}(m{\cdot}\, src) \wedge (status = standalone)\big)$
    $\| \;\; \langle \;\; \mathcal{N} := \mathcal{N} \cup \{m{\cdot}\, src, m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}\}$   if $\neg\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\| \; \langle \exists e : (e \in \mathcal{N}) \wedge (e \uparrow 1 = m{\cdot}\, src) :: e \uparrow 2, e \uparrow 3, e \uparrow 4 := m{\cdot}\, data \uparrow \text{agids}, clock, m{\cdot}\, data \uparrow \text{extant}$
           $\rangle$   if $\mathtt{isPresent}_{\mathcal{N}}(m{\cdot}\, src)$
        $\rangle$   if $\big(\mathtt{isMH}(m{\cdot}\, src) \wedge (status = connected)\big)$
  $\rangle$
```

A.2 Macros of Communication Manager

$$
\begin{array}{l}
\texttt{commSend}\mathrm{SH}_{iBSS}(\mathcal{Q}_{RT_S}) \triangleq \\
\quad \langle \quad r, \mathcal{Q}_{RT_S} := \texttt{head}(\mathcal{Q}_{RT_S}), \texttt{tail}(\mathcal{Q}_{RT_S}) \ \ \| \ \ dstid := r \uparrow \mathrm{dstHost} \\
\qquad \| \ \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, dstid, RT, r, \mathrm{W}) \quad \text{if } (\texttt{isSH}(dstid) \wedge (host_{hid}\,\Gamma'\,host_{dstid})) \\
\qquad \| \ \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } (\texttt{isSH}(dstid) \wedge \neg(host_{hid}\,\Gamma'\,host_{dstid})) \\
\qquad \| \ \ \langle \quad apid := \langle\ \exists\ e : (e \in \mathcal{H}) \wedge (e \uparrow 1 = dstid) :: e \uparrow 2\ \rangle \\
\qquad\qquad \| \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, apid, RT, r, \mathrm{W}) \quad \text{if } (host_{hid}\,\Gamma'\,host_{apid}) \\
\qquad\qquad \| \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } \neg(host_{hid}\,\Gamma'\,host_{apid}) \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMH}(dstid) \wedge \texttt{isPresent}_{\mathcal{H}}(dstid)) \\
\qquad \| \ \ \langle\ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, B_{iBSS_{\mathrm{W}}}, Locate, dstid, \mathrm{W}) \ \| \ \mathcal{L} := \mathcal{L} \cup \{(dstid, r, clock)\} \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMH}(dstid) \wedge \neg\texttt{isPresent}_{\mathcal{H}}(dstid)) \\
\quad \rangle
\end{array}
$$

$$
\begin{array}{l}
\texttt{commSend}\mathrm{MH}_{iBSS}(\mathcal{Q}_{RT_S}) \triangleq \\
\quad \langle \quad r, \mathcal{Q}_{RT_S} := \texttt{head}(\mathcal{Q}_{RT_S}), \texttt{tail}(\mathcal{Q}_{RT_S}) \ \ \| \ \ m := \texttt{newMsg}(hid, assoc[0], RT, r, \mathrm{WL}) \\
\qquad \| \ \ \mathcal{Q}_{S_{RT}}, \mathcal{CS}, lastRTsent, newRTGap, rtAtmpt := \mathcal{Q}_{S_{RT}} \bullet m, m, clock, rtGap, 0 \\
\hfill \text{if } (host_{hid}\,\Gamma'\,host_{assoc[0]}) \\
\qquad \| \ \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } \neg(host_{hid}\,\Gamma'\,host_{assoc[0]}) \\
\quad \rangle \quad \text{if } ((status = associated) \wedge (\mathcal{CS} = \perp))
\end{array}
$$

$$
\begin{array}{l}
\texttt{commSend}\mathrm{AP}_{iBSS}(\mathcal{Q}_{RT_S}) \triangleq \\
\quad \langle \quad r, \mathcal{Q}_{RT_S} := \texttt{head}(\mathcal{Q}_{RT_S}), \texttt{tail}(\mathcal{Q}_{RT_S}) \ \ \| \ \ dstid := r \uparrow \mathrm{dstHost} \\
\qquad \| \ \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, dstid, RT, r, \mathrm{W}) \quad \text{if } (\texttt{isSH}(dstid) \wedge (host_{hid}\,\Gamma'\,host_{dstid})) \\
\qquad \| \ \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } (\texttt{isSH}(dstid) \wedge \neg(host_{hid}\,\Gamma'\,host_{dstid})) \\
\qquad \| \ \ \langle \quad m := \texttt{newMsg}(hid, dstid, RT, r, \mathrm{WL}) \\
\qquad\qquad \| \ \mathcal{Q}_{S_{RT}}, \mathcal{CS}, lastRTsent, newRTGap, rtAtmpt := \mathcal{Q}_{S_{RT}} \bullet m, m, clock, rtGap, 0 \\
\hfill \text{if } (host_{hid}\,\Gamma'\,host_{dstid}) \\
\qquad\qquad \| \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } \neg(host_{hid}\,\Gamma'\,host_{dstid}) \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMH}(dstid) \wedge (dstid \in assoc) \wedge (\mathcal{CS} = \perp)) \\
\qquad \| \ \ \langle \quad apid := \langle\ \exists\ e : (e \in \mathcal{H}) \wedge (e \uparrow 1 = dstid) :: e \uparrow 2\ \rangle \\
\qquad\qquad \| \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, apid, RT, r, \mathrm{W}) \quad \text{if } (host_{hid}\,\Gamma'\,host_{apid}) \\
\qquad\qquad \| \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } \neg(host_{hid}\,\Gamma'\,host_{apid}) \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMH}(dstid) \wedge \neg(dstid \in assoc) \wedge \texttt{isPresent}_{\mathcal{H}}(dstid)) \\
\qquad \| \ \ \langle\ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, B_{iBSS_{\mathrm{W}}}, Locate, dstid, \mathrm{W}) \ \| \ \mathcal{L} := \mathcal{L} \cup \{(dstid, r, clock)\} \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMH}(dstid) \wedge \neg(dstid \in assoc) \wedge \neg\texttt{isPresent}_{\mathcal{H}}(dstid)) \\
\quad \rangle
\end{array}
$$

$$
\begin{array}{l}
\texttt{commSend}\mathrm{MH}_{IBSS}(\mathcal{Q}_{RT_S}) \triangleq \\
\quad \langle \quad r, \mathcal{Q}_{RT_S} := \texttt{head}(\mathcal{Q}_{RT_S}), \texttt{tail}(\mathcal{Q}_{RT_S}) \ \ \| \ \ dstid := r \uparrow \mathrm{dstHost} \ \ \| \ \ m := \texttt{newMsg}(hid, dstid, RT, r, \mathrm{WL}) \\
\qquad \| \ \ \mathcal{Q}_{S_{RT}}, \mathcal{CS}, lastRTsent, newRTGap, rtAtmpt := \mathcal{Q}_{S_{RT}} \bullet m, m, clock, rtGap, 0 \\
\hfill \text{if } (\texttt{isMH}(dstid) \wedge (host_{hid}\,\Gamma'\,host_{dstid})) \\
\qquad \| \ \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet r \quad \text{if } (\neg\texttt{isMH}(dstid) \vee \neg(host_{hid}\,\Gamma'\,host_{dstid})) \\
\quad \rangle \quad \text{if } ((status = connected) \wedge (\mathcal{CS} = \perp))
\end{array}
$$

$$
\begin{array}{l}
\texttt{commRcv}\mathrm{SH}_{iBSS}(\mathcal{Q}_{R_{RT}}) \triangleq \\
\quad \langle \quad m, \mathcal{Q}_{R_{RT}} := \texttt{head}(\mathcal{Q}_{R_{RT}}), \texttt{tail}(\mathcal{Q}_{R_{RT}}) \\
\qquad \| \ \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \quad \text{if } (\texttt{isMsgRT}(m) \wedge (\texttt{isSH}(m\cdot src) \vee \texttt{isAP}(m\cdot src)) \wedge (m\cdot dest = hid)) \\
\qquad \| \ \ \langle \quad \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, RT, \langle \exists e : (e \in \mathcal{L}) \wedge (e \uparrow 1 = m\cdot data) :: e \uparrow 2\rangle, \mathrm{W}) \\
\qquad\qquad \| \ \mathcal{H} := \mathcal{H} \cup \{(m\cdot data, m\cdot src, clock)\} \\
\qquad\quad \rangle \quad \text{if } (\texttt{isMsgFound}(m) \wedge \texttt{isAP}(m\cdot src) \wedge \texttt{isPresent}_{\mathcal{L}}(m\cdot data))
\end{array}
$$

```
  $\rangle$
$\texttt{commRcv}\mathrm{MH}_{iBSS}(\mathcal{Q}_{R_{RT}}) \triangleq$
  $\langle \quad m, \mathcal{Q}_{R_{RT}} := \texttt{head}(\mathcal{Q}_{R_{RT}}), \texttt{tail}(\mathcal{Q}_{R_{RT}})$
    $\| \ \langle \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \langle \exists e : (e \in \mathcal{L}RT) \wedge (e \uparrow 1 = m\cdot src) :: e \uparrow 2 := m\cdot mid \rangle$
          $\rangle$ if $(\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src) \wedge \neg\texttt{isRepeat}_{\mathcal{L}RT}(m\cdot src, m\cdot mid))$
        $\| \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \mathcal{L}RT := \mathcal{L}RT \cup \{(m\cdot src, m\cdot mid)\} \ \rangle$ if $\neg\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src)$
        $\| \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, ACK, m\cdot mid, \mathrm{WL})$
      $\rangle$ if $(\texttt{isMsgRT}(m) \wedge \texttt{isAP}(m\cdot src) \wedge (m\cdot src = assoc[0]) \wedge (m\cdot dest = hid))$
    $\| \ \mathcal{CS} :=\perp$ if $(\texttt{isMsgACK}(m) \wedge \texttt{isAP}(m\cdot src) \wedge (m\cdot src = assoc[0]) \wedge (m\cdot dest = hid) \wedge (rtAtmpt < 3)$
                $\wedge ((clock - lastRTsent) < newRTGap) \wedge \neg(\mathcal{CS} =\perp) \wedge (m\cdot mid = \mathcal{CS}\cdot mid))$
  $\rangle$
$\texttt{commRcv}\mathrm{AP}_{iBSS}(\mathcal{Q}_{R_{RT}}) \triangleq$
  $\langle \quad m, \mathcal{Q}_{R_{RT}} := \texttt{head}(\mathcal{Q}_{R_{RT}}), \texttt{tail}(\mathcal{Q}_{R_{RT}})$
    $\| \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data$ if $(\texttt{isMsgRT}(m) \wedge (\texttt{isSH}(m\cdot src) \vee \texttt{isAP}(m\cdot src)) \wedge (m\cdot dest = hid))$
    $\| \ \langle \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \langle \exists e : (e \in \mathcal{L}RT) \wedge (e \uparrow 1 = m\cdot src) :: e \uparrow 2 := m\cdot mid \rangle$
          $\rangle$ if $(\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src) \wedge \neg\texttt{isRepeat}_{\mathcal{L}RT}(m\cdot src, m\cdot data))$
        $\| \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \mathcal{L}RT := \mathcal{L}RT \cup \{(m\cdot src, m\cdot mid)\} \ \rangle$ if $\neg\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src)$
        $\| \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, ACK, m\cdot mid, \mathrm{WL})$
      $\rangle$ if $(\texttt{isMsgRT}(m) \wedge \texttt{isMH}(m\cdot src) \wedge (m\cdot src \in assoc) \wedge (m\cdot dest = hid))$
    $\| \ \mathcal{CS} :=\perp$ if $(\texttt{isMsgACK}(m) \wedge \texttt{isMH}(m\cdot src) \wedge (m\cdot src \in assoc) \wedge (m\cdot dest = hid) \wedge (rtAtmpt < 3)$
                $\wedge ((clock - lastRTsent) < newRTGap) \wedge \neg(\mathcal{CS} =\perp) \wedge (m\cdot mid = \mathcal{CS}\cdot mid))$
    $\| \ \langle \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, RT, \langle \exists e : (e \in \mathcal{L}) \wedge (e \uparrow 1 = m\cdot data) :: e \uparrow 2 \rangle, \mathrm{W})$
        $\| \ \mathcal{H} := \mathcal{H} \cup \{(m\cdot data, m\cdot src, clock)\}$
      $\rangle$ if $(\texttt{isMsgFound}(m) \wedge \texttt{isAP}(m\cdot src) \wedge \texttt{isPresent}_{\mathcal{L}}(m\cdot data))$
    $\| \ \langle \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, Found, m\cdot data, \mathrm{W})$ if $(m\cdot data \in assoc)$
      $\rangle$ if $(\texttt{isMsgLocate}(m) \wedge (\texttt{isSH}(m\cdot src) \vee \texttt{isAP}(m\cdot src)))$
  $\rangle$
$M := \texttt{commReSend}\mathsf{RT}_{iBSS}() \triangleq$
  $\langle \ M, lastRTsent, newRTGap, rtAtmpt := \mathcal{CS}, clock, (2 * newRTGap), (rtAtmpt + 1)$
                                        if $(\neg(\mathcal{CS} =\perp) \wedge (rtAtmpt < 3))$
  $\rangle$
$\texttt{commValid}_{\mathcal{H}\mathcal{L}_{iBSS}}() \triangleq$
  $\langle \quad \langle \| \ e : e \in \mathcal{H} :: \mathcal{H} := \mathcal{H} \setminus \{e\}$ if $\neg\texttt{isValid}_{\mathcal{H}}(e, clock) \ \rangle$
    $\| \ \langle \| \ e : e \in \mathcal{L} :: (\mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet e \uparrow 2) \ \| \ \mathcal{L} := \mathcal{L} \setminus \{e\}$ if $\neg\texttt{isValid}_{\mathcal{L}}(e, clock) \ \rangle$
  $\rangle$
$M := \texttt{commReSend}\mathsf{RT}_{IBSS}() \triangleq$
  $\langle \ M, lastRTsent, newRTGap, rtAtmpt := \mathcal{CS}, clock, (2 * newRTGap), (rtAtmpt + 1)$
                                        if $(\neg(\mathcal{CS} =\perp) \wedge (rtAtmpt < 3)) \ \rangle$
$\texttt{commRcv}\mathrm{MH}_{IBSS}(\mathcal{Q}_{R_{RT}}) \triangleq$
  $\langle \quad m, \mathcal{Q}_{R_{RT}} := \texttt{head}(\mathcal{Q}_{R_{RT}}), \texttt{tail}(\mathcal{Q}_{R_{RT}})$
    $\| \ \langle \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \langle \exists e : (e \in \mathcal{L}RT) \wedge (e \uparrow 1 = m\cdot src) :: e \uparrow 2 := m\cdot mid \rangle$
          $\rangle$ if $(\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src) \wedge \neg\texttt{isRepeat}_{\mathcal{L}RT}(m\cdot src, m\cdot mid))$
        $\| \ \langle \ \mathcal{Q}_{RT_R} := \mathcal{Q}_{RT_R} \bullet m\cdot data \ \| \ \mathcal{L}RT := \mathcal{L}RT \cup \{(m\cdot src, m\cdot mid)\} \ \rangle$ if $\neg\texttt{isPresent}_{\mathcal{L}RT}(m\cdot src)$
        $\| \ \mathcal{Q}_{S_{RT}} := \mathcal{Q}_{S_{RT}} \bullet \texttt{newMsg}(hid, m\cdot src, ACK, m\cdot mid, \mathrm{WL})$
      $\rangle$ if $(\texttt{isMsgRT}(m) \wedge \texttt{isMH}(m\cdot src) \wedge (m\cdot dest = hid))$
    $\| \ \mathcal{CS} :=\perp$ if $(\texttt{isMsgACK}(m) \wedge \texttt{isMH}(m\cdot src) \wedge (m\cdot dest = hid) \wedge (rtAtmpt < 3)$
                $\wedge ((clock - lastRTsent) < newRTGap) \wedge \neg(\mathcal{CS} =\perp) \wedge (m\cdot mid = \mathcal{CS}\cdot mid))$
  $\rangle$
```

Simplified Control Algorithm Based on IRP Theory for Three Phase Shunt Active Power Filter

Ajay Kumar Maurya and Yogesh K. Chauhan

Department of Electrical Engineering, School of Engineering, Gautam Buddha University,
Gr. Noida, India
ajaymauryaee@gmail.com

Abstract. In this paper a concept of instantaneous imaginary power (IRP) theory has been proposed for the load compensation of three phase balanced system. In this theory the total power is separated into two components, namely imaginary and real instantaneous power. The instantaneous imaginary power has been considered as zero and dc component of the real power is therefore selected as compensation power reference for compensation of harmonics and reactive power. In the design of active power filter (APF) three current, two ac voltage and one DC link voltage sensors have been used. By using this approach the overall system design becomes easier to accomplish and the implementation cost has been reduced. The model is prepared in MATLAB/SIMULINK and the system with the proposed theory has been tested for non-linear loads. The harmonic analysis of supply side quantities comply with the IEEE 519-1992 recommendations of harmonic standard limits, which validate the implementation of proposed shunt active power theory.

Keywords: Shunt Active Power Filter (APF), instantaneous reactive power theory (IRP), power quality improvement, balance three phase nonlinear system.

1 Introduction

In this present era, the use of power electronics devices has been increased in industrial level of application (i.e. switched-mode power supplies, uninterruptible power supply systems, inverters & high-frequency lighting etc.). These loads are nonlinear in nature. Due to these types of loads the current waveforms are non-sinusoidal in nature and contain high total harmonic distortion (THD). The harmonic distortion of current is representing as distortion factor of current waveform with respect to pure sine-wave [1]. Therefore a number of harmonics are injected into supply system. It creates more severe problem in power system network [2]. Traditionally, passive LC filters have been used to eliminate these harmonics in current. However, these passive LC filters are load dependent, fixed and bulky. They can also cause resonance problems to the system [3]. In order to solve these problems, active power filters (APFs) have been developed and used to mitigate these problems.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 431–440, 2013.

Several control algorithms have been proposed for the controlling of shunt active power filter [4]-[5]. In which most of control circuits are complicated and not easier to implement.

In this paper, the simplified instantaneous reactive power theory has been proposed which uses three current sensing devices, two voltage sensing devices and a DC link voltage measurement device for the controlling of APF. This proposed method is simple and total implementation cost is reduced.

1.1 Instantaneous Reactive Power Theory

Instantaneous reactive power theory deals with instantaneous voltages and currents in three-phase circuits mathematically [6]-[7]. Instantaneous voltage space-vectors v_a , v_b and v_c are transformed into 0-α-β coordinate as follows:

$$\begin{bmatrix} v_0 \\ v_\alpha \\ v_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} \qquad (1)$$

Likewise, the instantaneous current space-vectors i_a, i_b and i_c are transformed into 0-α-β coordinate.

$$\begin{bmatrix} i_0 \\ i_\alpha \\ i_\beta \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \qquad (2)$$

In case of balance system, zero sequence components are absent in the system. Hence the zero-phase sequence components are excluding in three-phase voltages and currents [8]. The simplified form of (1) and (2) may represent as for a balance three phase system

$$\begin{bmatrix} v_\alpha \\ v_\beta \end{bmatrix} = \frac{\sqrt{3}}{2} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix} \qquad (3)$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \frac{\sqrt{3}}{2} \begin{bmatrix} 1 & -1/2 & -1/2 \\ 0 & \sqrt{3}/2 & -\sqrt{3}/2 \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \qquad (4)$$

Fig.1 represents the transformation of three phase a-b-c coordinates to two phase α-β coordinates. The a-b-c axes are apart from each other by 2π/3. The instantaneous space vectors, v_a and i_a are placed on a-axis, and their amplitude and (+,-) direction vary with the passage of time. In the same way v_b and i_b are on the b-axis, and v_c and i_c are on the c-axis. Where the α and β, axes are the orthogonal coordinates [9].

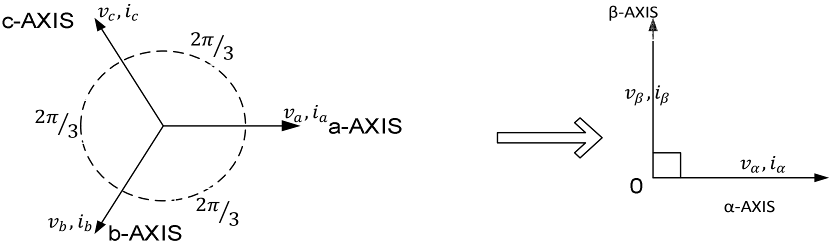


Fig. 1. α-β coordinate transformation

The conventional instantaneous real power on the three-phase circuit can be defined into α-β form as follows;

$$p = v_\alpha . i_\alpha + v_\beta . i_\beta \tag{5}$$

The instantaneous imaginary power space vector ' q ', can be defined by

$$q = v_\alpha \times i_\beta + v_\beta \times i_\alpha \tag{7}$$

This space vector is the imaginary axis vector and is perpendicular to the real plane on the α-β coordinates, In IRP theory the instantaneous real and reactive power can be expressed as:

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} v_\alpha & v_\beta \\ -v_\beta & v_\alpha \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \tag{8}$$

Equation (5) shows that the instantaneous power for three-phase system. The instantaneous real power ‘ p ’, will contribute to the instantaneous value of total energy flowing per time unit from source to load or vice versa while the instantaneous reactive power ' q ' represents energy that is being exchanged between the phases of the systems. Rearranging (8) for the calculation of α-β reference current as shown below [10]

$$i^*_{\alpha\beta} = \frac{1}{v_\alpha^2 + v_\beta^2} \begin{bmatrix} v_\alpha & v_\beta \\ v_\beta & -v_\alpha \end{bmatrix} \begin{bmatrix} p \\ q \end{bmatrix} \tag{9}$$

2 The Proposed Method

The control system block diagram of the proposed method for the three phase shunt APF is shown in Fig. 2. It is clear from Fig. 2, in this APF control circuit model only source voltage and current signals are utilized for the calculation of reference current. The hysteresis band current controller generated PWM (pulse width modulation) signal by the comparing between the reference current and the real source current. These PWM signal used to provide the gate pulses of voltage source inverter (VSI).

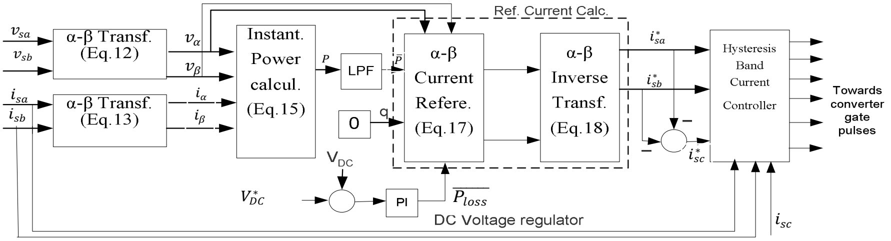


Fig. 2. Control block diagram of the proposed method

In this study, the shunt APF is connected to a three-phase three-wire system, which has sinusoidal and balanced source voltages and a three phase balanced rectifier load connected. In balance condition the sums of the instantaneous source voltages and source currents for a three phase system are zero. Measuring of two source voltages and currents are adequate for the reference current calculations. [11][12] Therefore, the (3) and (4) may represent as:

$$\begin{bmatrix} v_\alpha \\ v_\beta \end{bmatrix} = \begin{bmatrix} \sqrt{3/2} & 0 \\ 1/\sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} v_{sa} \\ v_{sb} \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \begin{bmatrix} \sqrt{3/2} & 0 \\ 1/\sqrt{2} & \sqrt{2} \end{bmatrix} \begin{bmatrix} i_{sa} \\ i_{sb} \end{bmatrix} \tag{11}$$

For making the simplified form of transformation (10) and (11) are multiplied with a factor of $\sqrt{2/3}$, and the simplified Transformations are obtained as in the form of (12) and (13).

$$\begin{bmatrix} v_\alpha \\ v_\beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/\sqrt{3} & 2/\sqrt{3} \end{bmatrix} \begin{bmatrix} v_{sa} \\ v_{sb} \end{bmatrix} \tag{12}$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/\sqrt{3} & 2/\sqrt{3} \end{bmatrix} \begin{bmatrix} i_{sa} \\ i_{sb} \end{bmatrix} \tag{13}$$

The instantaneous real ‘ p ’and reactive power ‘ q ’ calculated as:

$$\begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} v_\alpha & v_\beta \\ -v_\beta & v_\alpha \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \tag{14}$$

In this proposed method, calculation of the instantaneous imaginary power 'q' is not required which is used in conventional method, since it is no need to draw 'q' from the source. Only calculation of the instantaneous real power 'p' is adequate as shown in (15). Therefore, the proposed control method has been simplified in the calculations of reference currents.

$$p = v_{\alpha} i_{\alpha} + v_{\beta} i_{\beta} \tag{15}$$

The ac and dc values of instantaneous real power can be expressed as:

$$p = \overline{p} + \tilde{p} \tag{16}$$

Where '$\overline{p}$' is the dc component of instantaneous real power and '$\tilde{p}$' is the ac component of instantaneous real power.

In this proposed control algorithm, the instantaneous imaginary power is taken to zero (q=0) and a dc component of the real power'$\overline{p}$' is therefore selected as compensation power reference for compensation of harmonics and reactive power. The compensation current references in α-β coordinates are calculated by (15). In order to get the DC part of the instantaneous active power'$\overline{p}$' the signals need to be filtered using 1st order low pass filter with a cut-off frequency at 50 Hz. The low-pass filter will remove the high frequency component and give the fundamental part [10]. The average real power '$\overline{p_{loss}}$' is added to the dc component of the instantaneous real power '$\overline{p}$' to cover the voltage source inverter (VSI) losses of the shunt APF.

$$\begin{bmatrix} i^*_{s\alpha} \\ i^*_{s\beta} \end{bmatrix} = \frac{1}{v_{\alpha}^2 + v_{\beta}^2} \begin{bmatrix} v_{\alpha} & -v_{\beta} \\ v_{\beta} & v_{\alpha} \end{bmatrix} \begin{bmatrix} \overline{p} + \overline{p_{loss}} \\ 0 \end{bmatrix} \tag{17}$$

In this study, the compensation current references in the α-β coordinates are transformed back into the a-b-c coordinates by the use of inverse simplified α-β Transformation as given by

$$\begin{bmatrix} i^*_{sa} \\ i^*_{sb} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/2 & \sqrt{3}/2 \end{bmatrix} \begin{bmatrix} i^*_{c\alpha} \\ i^*_{c\beta} \end{bmatrix} \tag{18}$$

The current reference of phase 'c' is calculated by

$$i^*_{sc} = -(i^*_{sa} + i^*_{sb}) \tag{19}$$

These reference currents utilized for the generation of APF switching pulses by the comparing of the reference current ($i^*_{sa}, i^*_{sb}, i^*_{sc}$) and the actual source currents

(i_{sa}, i_{sb}, i_{sc}).By comparing these currents hysteresis-band current controller generate PWM pulses for controlling of VSI. The switching logics are formulated as follows:

If $i_{sa} > (i^*_{sa} + HB)$ higher switch is ON and lower switch is OFF for leg "A" (QA = 0).
If $i_{sa} < (i^*_{sa} - HB)$ higher switch is OFF and lower switch is ON for leg "A" (QA=1).

Similar For legs 'B' and 'C' the switching pulses obtained. Where HB is denoting hysteresis bandwidth and QA is the switching function [11]-[13].

Dc-link voltage regulator is used for a good compensation. The error is found by the comparing of actual dc-link capacitor voltage and reference value. This error is processed in a PI controller which generates additional average real power '$\overline{p_{loss}}$'.

3 Result and Discussion

This simulation model has been made for three phase 450V (phase to phase rms voltage), 50 Hz supply system and tested on MATLAB/SIMULINK. The complete simulation model of the proposed theory for the three phase system has been shown in fig.3 it contains three main sections such as load, control circuit, APF. The load circuit is given in fig.4, and the load component is given in the table-1. The three phase nonlinear load is made by using three phase diode bridge rectifier. In beginning only rectifier-I is operated and after 0.2 second rectifier-II is operated and rectifier-I eliminate in system. For opening and closing of the rectifier using three phase breaker and this three phase breaker is controlled by external step signal. In this load model the step signal is set at 0.2seconds. Fig.5 represent's the APF model, in this model VSI is connected with star connected 3 filter capacitors and 3 series inductors The load side current waveform is shown in fig6. The load is changed at 0.2 second as discussed above in the load model description, before load changes the FFT analysis has done in steady state for 5 cycles at 0.1 second is 39.28% and after load change the FFT analysis has been perform for steady state 5 cycles at 0.25 second THD obtained 20.47% as shown in fig.11. The compensating current waveform is shown in fig7. Fig8. shows the source current waveform its nearly about sinusoidal after load change and before load changing condition. Before load changes THD obtained 4.98% and after load change the THD became 3.46% as shown in fig.12. Fig.9 shows that the capacitor charging and discharging with very high frequency. Fig10. represent's voltage and current waveform of phase-a it is clear from this waveform that voltage and current are almost in the same phase. Hence power factor is improved and the harmonic analysis of supply side quantities comply with the IEEE 519-1992 recommendations of harmonic standard limits [14].

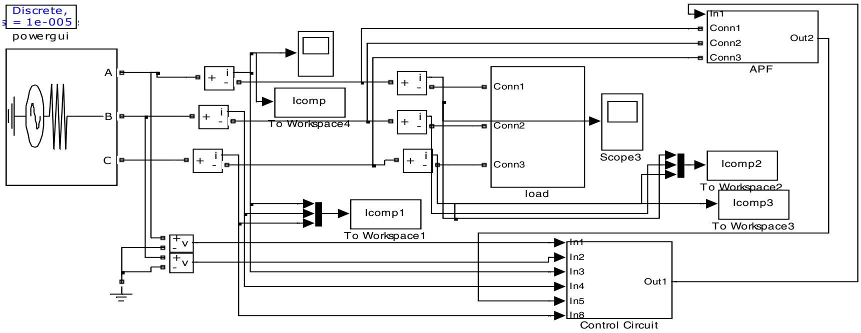


Fig. 3. Complete simulation model

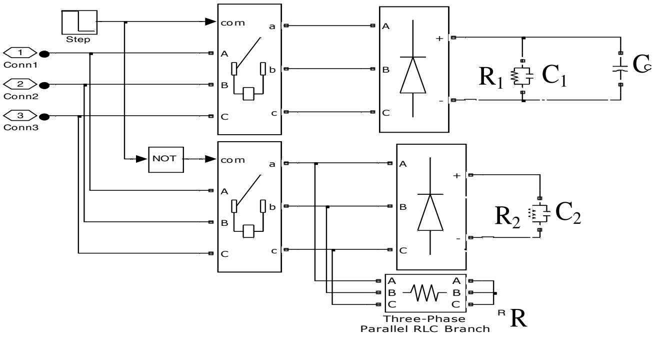


Fig. 4. Load model

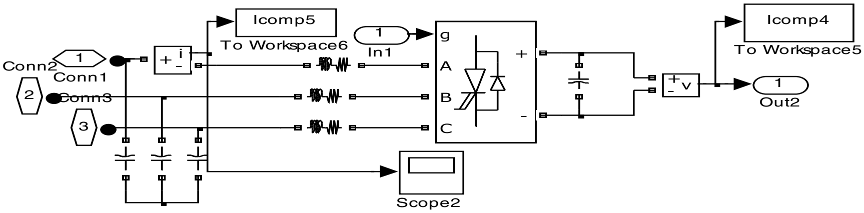


Fig. 5. APF model

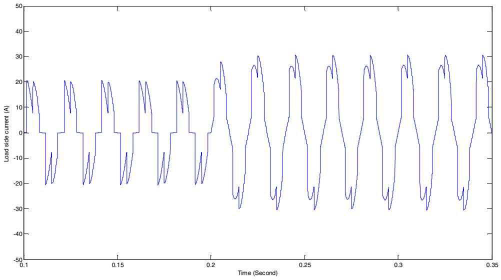


Fig. 6. Load side current

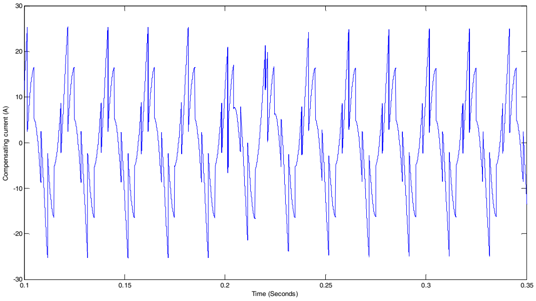


Fig. 7. Compensating current

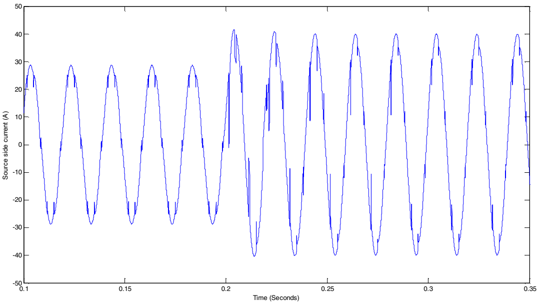


Fig. 8. Source side current

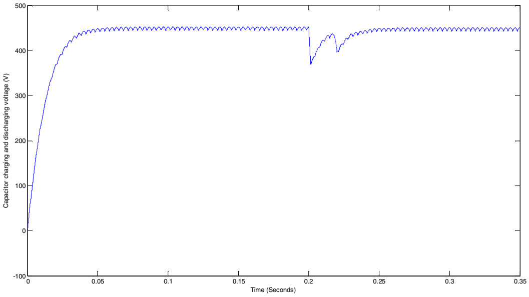


Fig. 9. Capacitor charging and discharging voltage

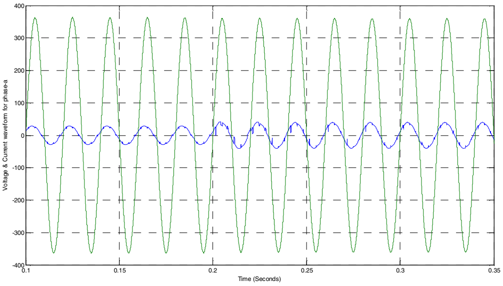


Fig. 10. Voltage & Current waveform for phase-a

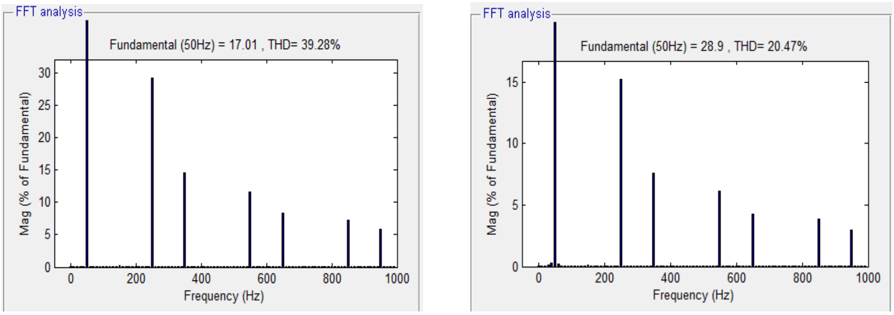


Fig. 11. FFT analysis of THD for load side current before load change 39.28% and after load change 20.47%

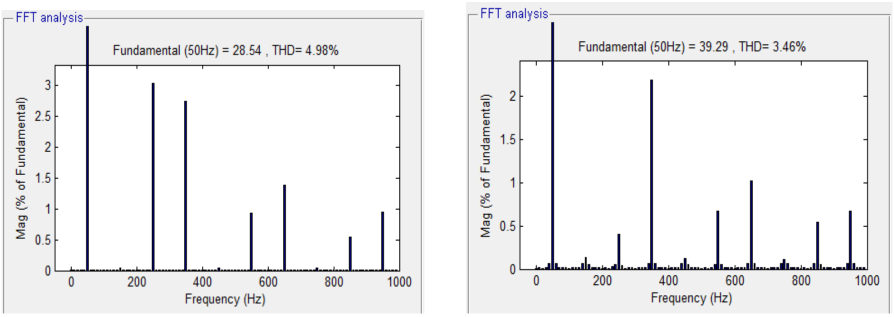


Fig. 12. FFT analysis of THD for supply side current before load change 4.98% and after load change 3.46%

Table 1. Technical Specification of Load

Component	*Quantity*	*Rating for each*
R_1	1	39Ω
C_1	1	46µF
C	1	21µF
R_2	1	39Ω
C_2	1	46µF
R	1	30Ω

R_1,C_1 and C used for rectifier-1, R_2 and C_2 are used for rectifier-2 and R represent's resistance of three phase parallel resistive branch.

4 Conclusion

The modeling and analysis of simplified IRP theory feeding non-linear load has been carried out on three-phase system. The complete system model has been developed in Matlab/simulink. The proposed control strategy has less complexity. The voltage source converter using this control strategy facilitates enhancement of power quality through reactive power compensation and harmonic suppression for nonlinear load. The THD of system quantities with compensator comply with the IEEE 519 standard, which there by validate the satisfactory system performance.

References

1. Cividino, L.: Power Factor, Harmonic distortion; Cause, effects and Considerations. In: IEEE Telecomunications Energy Conference, pp. 506–513 (1992)
2. Melhorn, C.J., McGranaghan, M.F.: Interpretation and analysis of power quality measurements. IEEE Trans. Ind. Appl. 31(6), 1363–1370 (1995)
3. Das, J.C.: Passive filters—Potentialities and limitations. IEEE Trans. Ind. Appl. 40(1), 232–241 (2004)
4. Chen, D., Xie, S.: Review of the control strategies applied to active power filters. In: IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies, Hong Kong, pp. 666–670 (2004)
5. Singh, B., Al-Haddad, K., Chandra, A.: A review of active filters for power quality improvement. IEEE Transactions- Industrial Electronics 46(5), 960–971 (1999)
6. McGranaghan, M.: Active Filter Design and Specification, for Control of Harmonics in Industrial and Commercial Facilities (2001)
7. Watanabe, E.H., Akagi, H., Aredes, M.: Instantaneous p-q power Theory for compensating nonsinusoidal systems. International School on Nonsinusoidal Currents and Compensation, Lagów, pp. 1–10 (2008)
8. Akagi, H., Kanazawa, Y., Nabae, A.: Instantaneous Reactive Power Compensator Comprising Switching Devices Without Energy Storage Components. IEEE Transactions on Industry Applications IA-20(3), 625–630 (1984)
9. Kim, H., Akagi, H.: The Instantaneous Power Theory on the Rotating p-q-r Reference Frames. Proceedings of IEEE PEDS 1, 422–427 (1999)
10. Lada, M.Y., Bugis, I., Talib, M.H.N.: Simulation a shunt active power filter using MATLAB/Simulink. In: 2010 4th International Power Engineering and Optimization Conference (PEOCO), pp. 371–375 (2010)
11. Ozdemir, E., Ucar, M., Kesler, M., Kale, M.: The Design and Implementation of a Shunt Active Power Filter based on Source Current Measurement. In: IEEE International Electric Machines & Drives Conf., IEMDC 2007, vol. 1, pp. 608–613 (2007)
12. Viji, A.J., Sudhakaran, M.: Generalized UPQC system with an improved control method under distorted and unbalanced load conditions. In: International Conference, ICCEET, pp. 193–197 (2012)
13. Malesani, L., Mattavelli, P., Tomasin, P.: High- Performance Hysteresis Modulation Technique for Active Filters. IEEE Transactions on Power Electronics 12(5), 876–884 (1997)
14. IEEE Standard for IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems, IEEE Std. 519 (1992)

Real Time Object Tracking: Simulation and Implementation on FPGA Based Soft Processor

Manoj Pandey[1], Dorothi Borgohain[2], Gargi Baruah[2], J.S. Ubhi[3], and Kota Solomon Raju[2]

[1] B K Birla Institute of Engineering and Technology, Pilani-333031
[2] DSG, Council of Scientific and Industrial research (CSIR) –Central Electronics Engineering Research Institute (CEERI) CSIR-CEERI, Pilani-3330311,
[3] ECE, Sant Longowal Institute of Engineering and Technology, Longowal, Sangrur, Pb
{manoj2pandey,js_ubhi}@yahoo.com, dorothi.s@gmail.com,
solomon@ceeri.ernet.in

Abstract. Adaptive systems are being easy to design using reconfiguration facility on Field programmable gate arrays (FPGAs). In this paper, Kernel based Mean shift algorithm is used for tracking a moving object. First it is simulated on Matlab and then implemented on microblaze soft processor based FPGA board. Tracking is observed for two similar objects crossing each other moving with uniform speed in a stored video as well as real time video. Object tracking, when it comes to implement on pure software (SW) in real time becomes difficult task due to certain limitations of SW. This paper shows how the mean shift algorithm is implemented on Xilinx Spartan 6 FPGA board using EDK. Once the complete algorithm is implemented on microblaze soft processor then some of the mathematical functions of algorithm are calculated on hardware to use HW-SW co-designing methodology to enhance the performance of the system.

Keywords: Kernel, Mean Shift, Real Time Tracking, EDK, FPGA, Spartan6.

1 Introduction

Object tracking is one of the fundamental component of computer vision that can be very beneficial in applications such as unmanned vehicles, surveillance, automated traffic control, biomedical image analysis and intelligent robots, to name a few. Tracking aims to generate the trajectory of objects across video frames. Object tracking is used for identifying the trajectory of moving object in video frame sequences. Like most computer vision tasks, object tracking involves intensive computation in order to extract the desired information from high volume video data. In addition, the real time processing requirements of different computer vision applications stress the need for high performance object tracking implementations. Implementation of vision systems in real time requires high performance HW with flexibility to incorporate the change after the design has been freezed. GPPS and DSPs give flexibility but not high performance while ASICS gives performance but not flexibility.

Latest technologies give performance and flexibility of FPGAs more and rugged with the use of Reconfigurable Computing System (RCS) facilities. Additionally,

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 441–450, 2013.

most FPGAs support dynamic partial reconfiguration (DPR) [1] to provide higher flexibility. With this benefit it is possible to reconfigure a part of the FPGA during run time. The other part of the FPGA is unaffected by this process and continues to run. Reconfigurable Computing offers cost-effective solution for computationally intensive application through reuse of hardware. Using dynamic programming, RCS is efficient in terms of hardware utilization without degrading its performance [2].

In this paper, we presents the simulation of Kernel based mean shift algorithm for object tracking and implementation on FPGA board. Objects as a person is used for the tracking, considering the case of overlapping and scale changing. The aim is to allow designers of applications that benefit from FPGA implementation, to leverage this capability for reconfigurable architecture. In Section 2, the tracking algorithms and its FPGA implementations are presented along with a review of existing hardware (HW) approaches. Section 3 gives basic steps of algorithm and 4 introduce about the HW IPs and their FPGA oriented design. Section 5, presents simulation and implementation results. Finally, Section 6 concludes this paper.

2 Related Work

Several implementations of object tracking for FPGAs exist. A soft-processor based object tracking system on FPGAs are carried out in ref [3, 4]. In this paper Xilinx 32 bit Microblaze soft processor is used to implement the tracking algorithm. The rest of the FPGA is used for the frame grabber and visualization of the video stream. A multi-object-tracking architecture for FPGAs or ASICs based on image segmentation is used in [5]. The algorithm is fixed one for a given constraints. If any constraints change after the design, the system does not work effectively. These systems have some restrictions. The main disadvantage is the restriction of the systems to one algorithm. If the constraints are changing the settings of an algorithm has to be changed. In the worst case the algorithm gets completely improper. To avoid these problems, the implementation of the system with reconfiguration facility is able to provide required change by replacing with additional flexibility in algorithm. Author in [6] implemented a hardware detection system based on the Active Shape Model (ASM) algorithm, and they reported speedup up to 15X compared to software execution. However, their implementation does not include tracking. Schlessman in [7] discusses a practical design based on a hardware/software co-design to realize an optical flow tracking system, which puts the KLT portion that consumes much processing time to FPGA-based hardware implementation as optimization. Christopher in [8] introduces an object tracking hardware design based on color features. Due to the advantages offered by FPGAs in compute intensive applications, several object tracking algorithms have been implemented on reconfigurable devices in recent researches. Nevertheless, one of the biggest challenges of custom hardware implementations is mapping complex algorithms onto reconfigurable fabric architectures that can offer good performance under rigid resource constraints.

For object tracking purposes, numerous algorithms have been proposed in the literatures and are implemented on FPGAs. Object tracking is a complex task which comprises two main subtasks: i) object detection and ii) tracking. In [9] object detection algorithms are classified into point detection, background subtraction techniques, and supervised learning techniques. Furthermore, the tracking portion of object tracking can be performed either separately or jointly with object detection. Alper Yilmaz [10] has characterized tracking algorithms across three main categories: i) point tracking, ii) kernel tracking and iii) silhouette tracking. Using these tracking techniques various tracking algorithms are developed /used for clustering and tracking of non rigid objects. Some of these popular algorithms are KLT-tracker [11], mean shift [12], mean shift with motion vector [13], kernel based Mean shift [14], eigen tracking [15], optical flow tracking [16], fast object tracking using adaptive block matching [17], heuristic methods for object tracking [18], Kalman filter [19], particle filter [20] etc, which are used for various image processing applications instead of only object tracking. The preliminary work for the system presented in this paper was published in [10]. The proposed work was purely SW based and implement on EDK based design.

3 Mean Shift Algorithm

Mean shift is a nonparametric density gradient statistical method which considers feature space as an empirical probability density function (PDF). If the input is a set of points, then mean shift considers them as sampled from the underlying density function. In mean shift trackers, objects of interest are characterized by the probability density functions (pdfs) of their colour or texture features. A spatially-smooth similarity function between the original object as candidate and target are defined as a masking distributions with a monotonically decreasing kernel. Mean shift iterations are then used as a local maximum of this similarity function as an indicator of the direction of target's movement. In order to apply a mean shift calculation, the set of histogram values is weighted by Epanechnikov kernel to yield a smoothed set of values. It defines an ellipsoidal region and gives more weights to pixel closer to the center of the kernel. The rationale for using a kernel to assign smaller weights to pixels farther from the centre is that those pixels are the least reliable, since they are the ones most affected by occlusion or interference from the background. A kernel with Epanechnikov profile was essential for the derivation of the smooth similarity function between the distributions. Since its derivative is constant; the kernel masking lead to a function suitable for gradient optimization, which gave us the direction of the target's movement. The search for the matching target candidate in that case is restricted to a much smaller area and therefore it is much faster than the exhaustive search. The number of bins in the histogram representation for the target is defined by the user. This allows for simpler histograms in cases where the image sequence features highly distinctive colors and is devoid of collision events between like-colored objects. Likewise a larger number of bins may be used if the range of colors is limited or the target and background colors are nearby in normalized RGB. Thus algorithm completes its processing mainly in six steps. In First step we initialize the location of target in current frame. In step two, it computes PDF for target and candidate. In the same step it also computes the similarity between them using Bhattacharya Coefficient. In step third and forth, computes the weights and then apply the mean shift to find the new location respectively. Finally, in step five finds the new location of object in reference to the centre candidate and in step sixth iterate the algorithm for all the frames of video streaming.

4 FPGA Implementation

4.1 Base System Builder

All EDK designs are built on a Base System Builder (BSB) platform which provides a common base and building blocks. Each of the EDK reference designs included with the IVK is built from the base platform. The Base Platform is not a separate design that is delivered with this kit, rather it is the starting point from which all the other designs were built. The board we have used is Xilinx Industrial Video Processing Kit (IVK) Spartan-6 XC6SLX150T-3FGG676C- based embedded platform as shown in fig. 1. It provides two FMC LPC general-purpose I/O expansion connectors, and a memory of 128 MB DDR3 SDRAM. For communication we used RS-232 serial port. The hardware for the implementation of mean shift is made by adding various IPs and peripherals. These are Micro Blaze™ 32-bit soft microprocessor, Local Memory Bus (LMB), LMB Block RAM controller, Block RAM block memory, Processor Local Bus (PLB46), XPS UARTlite, Xilinx Platform Studio (XPS) General Purpose Input/Output (GPIO), XPS Inter-Integrated Circuit (IIC) Controller, External Multi-port Memory Controller (MPMC), MDM MicroBlaze Debug Module, Clock Generator, Processor System Reset and finally the DDR3 block representing the external memory.

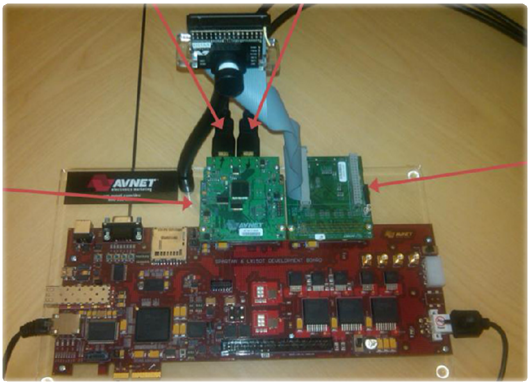


Fig. 1. Xilinx Spartan6 IVK FPGA board setup

The image sensor video input source enters the Camera Input PCORE [21]. This PCORE decodes the BT656 codes to generate synchronization signals and formats the video as an XSVI bus interface. The Video Detect PCORE does not alter the video, but monitors the VSYNC and ACTIVE VIDEO signals to determine the dimensions of the active video streaming through the FPGA. It also generates Video DMA compatible bus interface used to write video data to external memory. The Video DMA PCOREs, in collaboration with the Video Frame Buffer Controller (VFBC) [21] interfaces on the Multi-Port Memory Controller (MPMC), perform the actual transfers to/from external memory. These cores are extremely flexible and are configured via the Micro Blaze processor. The GENLOCK port indicates where the first Video DMA has written the incoming frames. The second Video DMA reads video frames from memory based on the GENLOCK information. After that the histogram calculation IP and later the mean shift block gets the pixel data and takes the RGB values, each of 8 bit. It takes the pre-calculated kernel values and finds out the histogram of the target and the candidate model and they compute the displacement in the mean shift block of the target object in each frame. Since the output frame rate is higher than the input frame rate, frames are duplicated when necessary. The Video Generate PCORE, under control of the Micro Blaze, generates video timing for the output. It also generates a Video DMA compatible bus interface used to read video data from external memory. The DVI Output PCORE takes an XSVI bus interface as input and optionally drives the pins of the DVI output interface. This output to the FMC connector will only be driven once the FMCIMAGEOV module has properly been identified. The video capture is at 1280x720P @ 30Hz and video playback at 1280x720P @ 60Hz. These resolutions are configured by the embedded processor (Micro Blaze) and can be modified to support other resolutions (limited by the image sensor used).

4.2 Compute Displacement (HW)

To accelerate the computing faster, compute displacement function is replaced with HW designed IP compute_dis_25 as shown in fig 2. The displacement dx and dy is calculated with the input functions Kernel derivative function (S1) and weight function (S2). These values are stored in BRAM segmented memory locations (160 X 80) To store values S1 and S2 before computing the displacement. The counter I logic and j logic are used to multiply the element by element of S1 and S2 and then added the pixel values as shown in block diagram in fig 2.

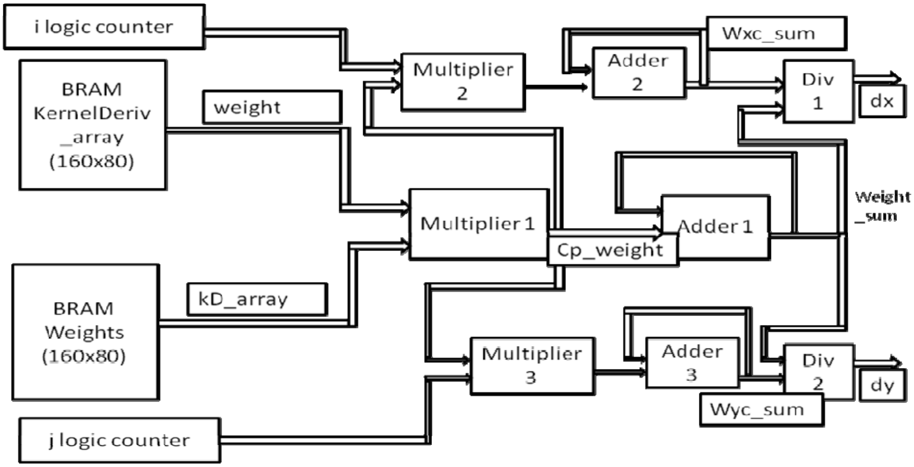


Fig. 2. Data flow diagram of Compute displacement module

5 Simulation and Implementation Results

5.1 Simulation Results

To evaluate and compare the system performance of tracking algorithm, first the algorithm is simulated on Matlab tool. In the simulation of the tracker, we have choose RGB color space as a feature space, in which chosen feature space was quantized into 5 x 5 x 5 bins. The set of histogram values is weighed by Epanechnikov kernel which yields a smoothed set of values. A constant kernel derivative is used in the calculation of the kernel profile. The value of Cd = 2 and d=1 is used. The algorithm is executed comfortably at 75 frames per second (fps) on 2.70 GHz PC, Matlab (version 7.60). The Matlab simulation results are shown in fig 3.1, fig 3.2 and fig 3.3. The first result in fig 3.1, states that when we increase bin number from 5 to 25 to 50 of the target, then there is slight difference between the trajectory of the object in different bins condition but this difference is too less. It means on increasing the bins there is a separation between the object being tracked and background, which shows that algorithm is robust to bin size. Graph in fig 3.2 shows the metric distance between the frames 320 to 340 is very less compare to frames 450 to 470, which means that there is maximum similarity (i.e. minimum distance) between the target window and the candidate window in the successive frames from 320 to 340. It signifies the movement of the object in these consecutive frames is slow in respect to frames 450 to 470. It means that the change in the histogram does also affect the similarity (i.e. distance) between the target model and the candidate model. The graph in fig 3.3 shows the Mean Shift iteration. The maximum distance shows, the more illuminated area is there in the frames ranging from 450 to 500. It means

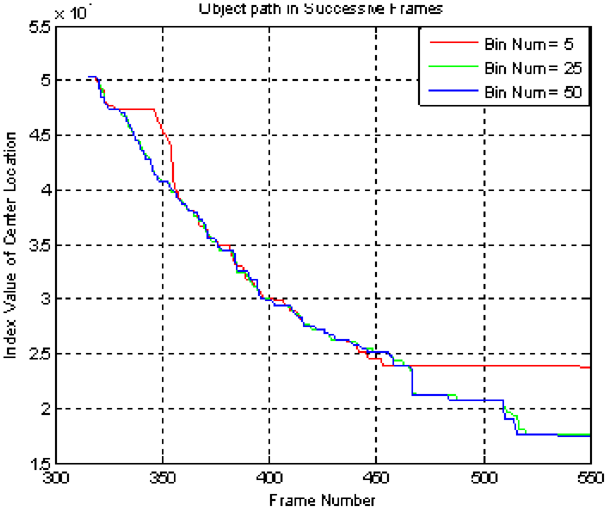


Fig. 3.1. Measurement of estimated location in successive frames

when the object is being tracked and passes from a well illuminated region to variable illuminated region, the model histogram is not indicative of the target very well. Ultimately, tracker shifts from the actual object. As a result, if the two similar objects are overlapped in a scene the tracker move to other similar object. The tracking results of two similar objects are shown in fig 3.4 in three different frames. From top to bottom: (a) before overlapping of objects (b) overlapping of objects in middle and (c) tracking to another similar object at bottom. Object is tracked well before the overlapping of both the objects.

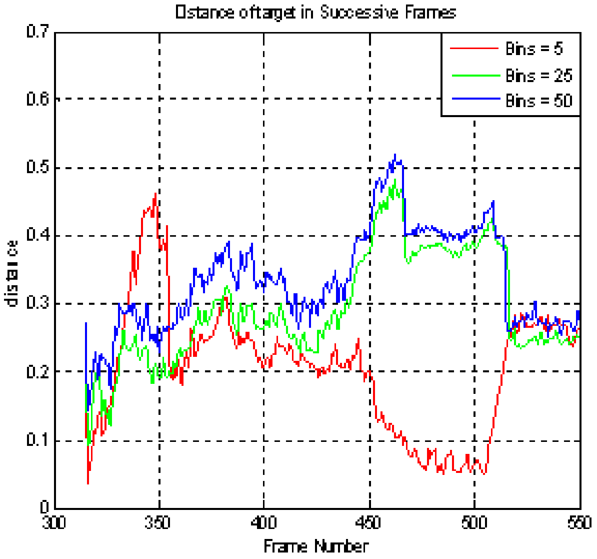


Fig. 3.2. The minimum value of distance function of the frame index

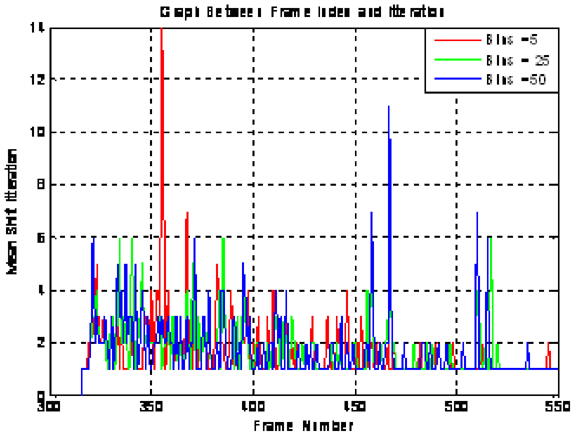


Fig. 3.3. The number of mean shift iterations function of the frame index

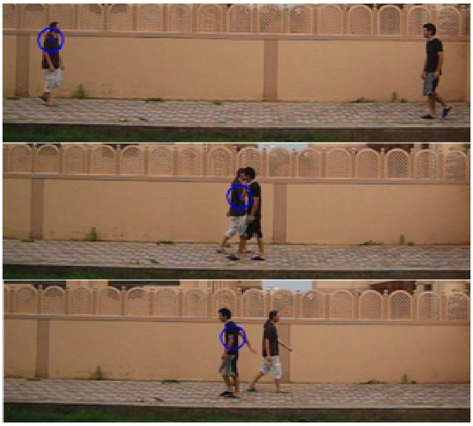


Fig. 3.4. Tracking frames of a moving person in top figure, in middle overlapping of similar object at bottom tracker shifts to other similar object after overlapping

5.2 Implementation Results

Table 1. The table shows the resources used and its percentage utilization

Device Utilization Summary (estimated values)			
Logic Utilization	Used	Available	Utilization
Number of Slice Registers	11,437	184,304	6%
Number of Slice LUTs	10980	92,152	11%
Number used as memory	846	21,680	3%
Number of bonded IOBs	95	386	23%
Number of BUFG/BUFGCTRLs	10	32	31%
Number of DSP48Es	18	128	4%

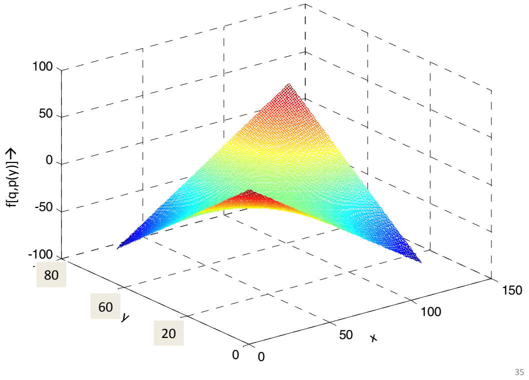


Fig. 4.1. Similarity function f [q, p(y)]

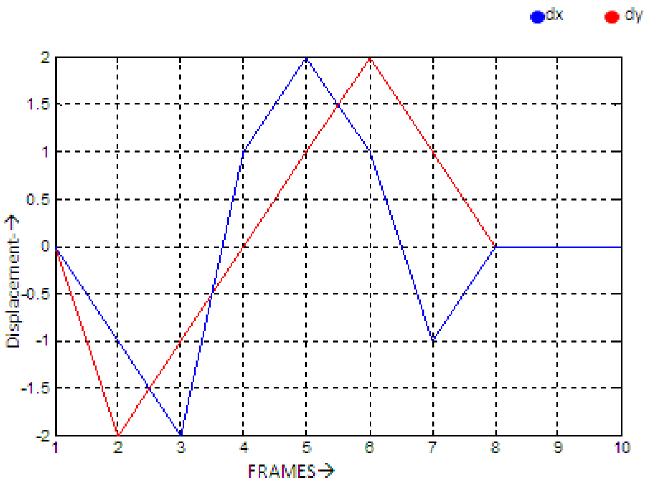


Fig. 4.2. Displacement of the target object in 10 consecutive frames

6 Conclusion

The simulated result of Kernel based mean shift algorithm is found to be smoothly tracking a specified object if there is a good separation between the objects. If the objects are found similar and overlapped then simple Mean Shift may track a wrong object. So this shortcoming may overcome with combination of other algorithms or with filtering in future work. On the other hand, for the implementation in real time tracking some of the frames are lapsed due to high complexity of computation if algorithm is executed on only general purpose soft processor. Solution for real time implementation of algorithm is to bring the complete execution with HW/SW co- design methodology to accelerate the execution. In this reference a compute displacement function is be replaced with hardware which is presented in this paper as an IP.

References

[1] Hsiung, P.-A., Santambrogio, M.D., Huang, C.-H.: Reconfigurable System Design and Verification. CRC Press © Taylor & Francis Group, London (2009)

[2] Compton, K., Hauck, S.: Reconfigurable Computing: A Survey of Systems and Software. ACM Computing Surveys 34(2), 171–210 (2002)

[3] Ali, U., Malik, M.B., Munawar, K.: FPGA/Soft- Processor based real-time object tracking system. In: Proceedings IEEE, Fifth Southern Programmable Logic Conference, pp. 33–37 (2009)

[4] Raju, K.S., Baruah, G., Rajesham, M., Phukan, P.: Computing Displacement of Moving Object in a Real Time Video using EDK. In: International Conference on Computing, Communications, Systems And Applications (ICCCSA), Hyderabad, March 30-31, pp. 76–79 (2012) ISBN:978-81-921580-8-2

[5] Rummele-Werner, M., Perschke, T., Braun, L., Hübner, M., Becker, J.: A FPGA based fast runtime reconfigurable real-time Multi-Object-Tracker. In: IEEE International Symposium on Circuits and System (ISCAS) (May 2011)

[6] Xu, J., Dou, Y., Li, J., Zhou, X., Dou, Q.: FPGA Accelerating Algorithms of Active Shape Model in People Tracking Applications. In: Proc. 10th IEEE Euromicro Conference on Digital System Design Architectures, Methods and Tools (DSD 2007) (2007)

[7] Schlessman, J., Chen, C.Y., Ozer, B., Fujino, K., Itoh, K., Wolf, W.: Hardware/software Co-design of an FPGA based Embedded Tracking System. In: Proceedings of the IEEE Conference on Computer Vision and Pattern 1662 Recognition Workshop, pp. 123–133 (2006)

[8] Johnston, C.T., Gribbon, K.T., Bailey, D.G.: FPGA based Remote Object Tracking for Real-time Control. In: Proceeding 1st International Conference on Sensing Technology, November 21-23, pp. 66–71 (2005)

[9] Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A Survey. ACM Computing Surveys 38(4), Article 13 (December 2006)

[10] Raju, K.S., Baruah, G., Rajesham, M., Phukhan, P., Pandey, M.: Implementation of moving object tracking using EDK. International Journal of Computer Science Issues (IJCSI) 9(3), 43–50 (2012)

[11] Shi, J., Tomasi, C.: Good features to track. In: Proceeding IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593–600 (1994)

[12] Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recongition, Hilton Head, vol. 2, pp. 142–149 (2000)

[13] Tian, G., Hu, R.-M., Wang, Z.-Y., Zhu, L.: Object Tracking Algorithm Based on Meanshift Algorithm Combining with motion vector analysis. In: Proceeding, First International Workshop on Education Technology and Computer Science, vol. 01, pp. 987–990 (2009)

[14] Comaniciu, D., Ramesh, V., Meer, P.: Kernel-Based Object Tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence 25(5), 564–577 (2003)

[15] Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International conference on Artificial Intelligence (IJCAI), August 24-28, pp. 674–679 (1981)

[16] Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. Int. J. Comput. Vision (IJCV) 12(1), 43–77 (1994)

[17] Hariharakrishnan, K., Schonfeld, D.: Fast object tracking using adaptive block matching. IEEE Transaction on Multimedia 7(5) (October 2005)

[18] Ronfard, R.: Region based strategies for active contour models. Int. J. Comput. Vision 13(2), 229–251 (1994)

[19] Zhong, J., Sclaroff, S.: Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In: Proceeding of the Ninth IEEE International Conference on Computer Vision (ICCV), October 13-16, vol. 1, pp. 44–50 (2003)

[20] Zhou, S., Chellapa, R., Moghadam, B.: Adaptive visual tracking and recognition using particle filters. IEEE Transactions on Image Processing 13(11), 1491–1506 (2004)

[21] Spartan-6 Industrial Video Processing Kit – EDK Reference Design Tutorial, Xilinx Inc., http://www.xilinx.com

Reduced Complexity Pseudo-fractional Adaptive Algorithm with Variable Tap-Length Selection

Asutosh Kar[1] and Mahesh Chandra[2]

[1] Dept. of Electronics and Telecommunication Engineering, IIIT, Bhubaneswar, India
[2] Dept. of Electronics and Communication Engineering, BIT, Mesra, India
asutosh@iiit-bh.ac.in, shrotriya69@rediffmail.com

Abstract. The structural complexity and overall performance of the adaptive filter depend on its structure. The number of taps is one of the most important structural parameters of the liner adaptive filter. In practice the system length is not known a-priori and has to be estimated from the knowledge of the input and output signals. In a system identification framework the tap length estimation algorithm automatically adapts the filter order to the desired optimum value which makes the variable order adaptive filter a best identifier of the unknown plant. In this paper an improved pseudo-fractional tap-length selection algorithm has been proposed to find out the optimum tap-length which best balances the complexity and steady state performance. Simulation results reveal that the proposed algorithm results in reduced complexity and faster convergence in comparison to existing tap-length learning methods.

Keywords: Adaptive filter, tap-length, structure adaptation, least mean square (LMS), system identification, mean square error (MSE).

1 Introduction

Inherent stability and tapped delay line (TDL) feed forward structure makes finite impulse response (FIR) adaptive filter widely popular than its infinite impulse response (IIR) counterpart [1], [2]. IIR system is preferred due to its less computational complexity [1]. An FIR system has a finite impulse so an ideal tap-length can be selected to match the system order in an identification application. Whereas the infinity impulse response of an IIR system used in a system identification framework makes the tap-length of the adaptive filter critical one to best adjust the system performance. The performance of the TDL structure of the adaptive filter in which the weights/tap-coefficients are recursively updated by adaptive algorithm such as the least mean square (LMS), recursive least square (RLS) is highly affected by the filter order or in other words the tap-length selection [3]. The LMS algorithm has been extensively used in many applications because of its simplicity and robustness [1]. A too short order filter results in inefficient model of the system and increases the mean square error (MSE) [4], [5]. In principle minimum MSE (MMSE) is a monotonic non increasing function of the filter order but it is not advisable to have a too long order filter as it introduce adaptation noise and extra

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 451–460, 2013.

complexity due to more taps [6]-[9]. Therefore to balance the adaptive filter performance and complexity there should be an optimum order of the filter. More relevant work was proposed in [10] where the filter is partitioned into segments and order is adjusted by one segment either being added or removed from the filter according to the difference of the output errors from the last two segments. This algorithm suffers from the drawback of carefully selecting the segment parameters and use of absolute error rather than MSE i.e. to solve the problem of suitable tap-length estimation it creates another issue of selecting the proper length of the segment.

The fractional tap-length LMS (FT-LMS) algorithm was first proposed in [6]-[8] relaxing the constraint that the filter order must be an integer. This fractional order estimation procedure retains the advantage of both segmented filter and gradient decent algorithm and has less complexity than the previously proposed methods. But it suffers from noise level and parameter variation due to unconstrained and random use of the leaky factor and step size used for order adaptation [11], [12]. An improved variable tap-length variable step LMS (VT-VLMS) algorithm produces better convergence and steady state error performance than the FT-LMS algorithm [11], [13]. But it depends on a careful selection of leaky factor which controls the overall tap-length adaptation. This algorithm [11], [13] is found to be more suitable for the echo cancellation applications with the parameter guidelines it suggest. In this paper a new gradient search method based on the pseudo-fractional order estimation technique is proposed which finds the optimum filter order dynamically with a modified tap-length learning procedure. The filter order can be increased to and decreased from any value to achieve the desired tap-length for structure adaptation. There should be a trade-off between a suitable steady state tap-length and convergence rate. The steady state performance analysis of the proposed algorithm shows the importance of variable error width parameter. The proposed algorithm shows better performance both in convergence as well as MSE in comparison to the famous FT-LMS [6] and VT-VLMS algorithm [11]. It reduces the overall design complexity and hence proves to be a cost saving design.

The paper is organized as follows. The tap-length optimization with an improved pseudo-fractional tap-length selection algorithm has been proposed in Section-2. In Section-3 the computer simulation setup has been designed both for FIR and IIR system identification frameworks. The results and discussion are given in Section-4.

2 Pseudo-fractional Tap-Length Optimization

Selecting the tap-length for any system identification frame work is not a trivial task. The selection depends on the nature of the system to be identified, memory requirement, desired performance, computational complexity, noise level and parameter variation etc. For example, in a multiuser acoustic echo cancellation arrangement the optimum tap-length may vary with the variation in time as echo length keeps changing due to the users entering and leaving the room/system [9]. In most filter designs unfortunately the tap-length is fixed at some fixed value creating the problem of too short and too long filters.

In LMS based algorithms the stability condition needs to be checked each time the order changes. So it is advocated to use the normalized LMS (NLMS) algorithm for better convergence and constant level of misadjustment [2].

$$W(n+1)=W(n)+\frac{\bar{\mu}}{X^T(n)X(n)}X(n)e(n) \tag{1}$$

Here $\bar{\mu}$ is the step size for NLMS algorithm. NLMS converges to mean square for condition [1], 0< $\bar{\mu}$ <2.

In the proposed approach NLMS is used which provides inherent stability and robustness again the modification to it improves the convergence [11],

$$W_{P(n)}(n+1)=W_{P(n)}(n)+\frac{\mu'}{X^T_{P(n)}(n)X_{P(n)}(n)[2+P(n)]}X_{P(n)}(n)e^P_{P(n)}(n) \tag{2}$$

where μ' is a constant, $\sigma_X^2 = X^T(n)X(n)$ is the variance of input signal. $P(n)$ is the instantaneous variable adaptive tap-length obtained from the proposed fractional order estimation algorithm. $W_{P(n)}, X_{P(n)}$ are the weight and input vector pertaining to the order $P(n)$.

$$e^P_{P(n)}(n)=d(n)-W^T_{P(n)}X_{P(n)}(n) \tag{3}$$

$$d(n)=W^T_{P_{Opt}}(n)*X_{P_{opt}}(n)+t(n) \tag{4}$$

where $W_{P_{Opt}}(n), X_{P_{opt}}(n)$ are the weight and input vector pertaining to optimum tap-length P_{Opt} and $t(n)$ is the system noise.

$$\mu(n)=\frac{\mu'}{{\sigma_X}^2[2+P(n)]} \tag{5}$$

Then (2) can be written as

$$W_{P(n)}(n+1)=W_{P(n)}(n)+\mu(n)X_{P(n)}(n)e^P_{P(n)}(n) \tag{6}$$

which forms a variable step NLMS (VNLMS) algorithm where the step size depends on the order estimation. If the difference of the MSE output of any two consecutive taps of the adaptive filter falls below a very small positive value, when the tap-length is increased, then it can be concluded that adding extra taps added to the present order do not reduce the MSE. Let us define $\Delta_P=J_{P-1}(\infty)-J_P(\infty)$ as the difference between the converged MSE when the filter order is increased from *P-1* to *P*. Now the optimum order can be defined as $\bar{P}$ that satisfies,

$$\Delta_P \le \delta \qquad \text{for all } P > \bar{P} \tag{7}$$

where δ is a very small positive number set pertaining to the system requirement. The cost function for tap-length selection can be defined as $\min\{P \mid J_{P-1}-J_P \le \delta\}$.The issue of false optimum tap-length sometimes creates confusion in the search of desired optimum filter length. These pseudo tap-lengths can be defined as follows. Let there exist a positive integer L that satisfies,

$$L < \bar{P} \text{ and } \Delta_L < \delta \tag{8}$$

where L is called the pseudo-optimum filter order . If the above condition is satisfied by a group of concatenated integer L, $L+1$,........ $L+S$-1 then $S+1$ is called the width of the pseudo-optimum filter order. These taps satisfies the optimality condition but cannot be treated as the optimal filter order as it under model the system. The issue of this pseudo-optimum tap-length can be removed by choosing a variable error width Δ which is shown later in this section.

The steady state MSE is not available usually and can be found out by exponential averaging,

$$J(n+1) = (1-H)e^2(n+1) + H\,J(n) \tag{9}$$

where H is the smoothing constant which control the effective memory of the iterative process. A smoothed estimated error can be obtained from,

$$\tilde{e}^{P}_{P(n)}(n) = (1-f)\sum_{i=0}^{n-1} f^{n-i} e^{P(i)}_{P_{Opt}(i)}(i) + t(i) * f^{i} \tag{10}$$

where n is the time index, P_{Opt} is the optimum suitable selection of tap-length. f is a forgetting smoothing factor which can be evaluated as,

$$f = \frac{\log_{10} \Delta_P(n)}{(\Delta_{\max} - \Delta_{\min})} \tag{11}$$

where $\Delta_P(n)$ is the variable error spacing parameter which has been evaluated in the next section, $(\Delta_{\max}, \Delta_{\min})$ can be set according to the system requirements.

The MSE can be written as the sum of excess MSE (EMSE) and the system noise as,

$$E[(e^{P}_{P(n)}(n))^2] = E[J_{ex}{}^2(n)] + E[t^2(n)] \tag{12}$$

where $J_{ex}(n)$ is the EMSE which is used in updating the iteration parameter as it increases to large value in the early stage and later decreases to small value with the variation in tap-length. The performance of the squared smoothed estimated error can be simplified at the steady state as, [16]

$$E[\tilde{e}^{P}_{P(n)}(n)^2] \cong \frac{(1-f)}{(1+f)}\sigma_t^2 \tag{13}$$

where it is assumed that the error signal approximates to the system noise at the steady state. Now the algorithm for tap-length adaptation in a time varying environment can be defined as,

$$P_{nf}(n+1) = \left[P_{nf}(n) - K_n\right] + [(e^{P}_{P(n)}(n))^2 - (e^{P}_{P(n)-\Delta_P(n)}(n))^2]\bar{K}_n \tag{14}$$

Finally the tap-length $P(n+1)$ in the adaptation of filter weights for next iteration can be formulated as follows, [11]

$$P(n+1) = \left\langle P_{nf}(n) \right\rangle \; if \left| P(n) - P_{nf}(n) \right| > \frac{K_n}{\bar{K}_n} \qquad (15)$$

$$P(n) \qquad otherwise$$

$P_{nf}(n)$ is the tap-length which can take fractional values. As the actual order of the adaptive filter cannot be a fractional value so $P_{nf}(n)$ is rounded to the nearest integer value to get the suitable optimum tap-length. In (14) the factor K_n is the leakage factor which prevents the order to be increased to an unexpectedly large value and $\bar{K}_n$ is the step size for filter order adaptation. In [11] the value of $(K_n, \bar{K}_n)$ was based on setting a random leaky factor which performed well for FIR systems especially for the issues of acoustic echo cancellation [13]. In this paper a unique method for setting these parameters has been defined which can find better performance in structure adaptation and hence decreases the overall design complexity.

$$K_n = \min(K_{n,\max}, K_n(i+1)) \qquad (16)$$

where

$$K_n(i+1) = \frac{\tilde{e}^2(i+1)}{\tilde{e}^2(i+1) + \Delta_{P_{ss}}(i)} \qquad (17)$$

$$\tilde{e}^P_{P(n)}(i+1) = f * \tilde{e}^P_{P(n)}(i) + (1-f) e^P_{P_{Opt}(n)}(i+1) \qquad (18)$$

$\Delta_{P_{ss}}$ defines the variable error spacing parameter at steady state tap-length P_{ss}. At the steady state, [16]

$$K_n \rightarrow \frac{(1-f)\sigma_t^2}{(1+f)\Delta_{P_{ss}}(\infty)} \qquad (19)$$

Similarly the adaptation step size depends on the bias between MSE values with a Δ_p difference. If the difference is more, then adaptation should be slow and vice versa.

$$\bar{K}_n = \min(\bar{K}_{n,\max}, \tau * \bar{K}_{n,\max}) \qquad (20)$$

where

$$\tau = \frac{[e^P_{P(n)}(n))^2 - (e^P_{P(n)-\Delta_P(n)}(n))^2]}{(e^P_{P_{Opt}(n)}(n))^2 + f * [e^P_{P(n)}(n))^2 - (e^P_{P(n)-\Delta_P(n)}(n))^2]} \qquad (21)$$

The variable error width parameter Δ_p decides the bias between the unknown optimum tap-length P_{opt} and the steady state tap-length in a system identification framework. It removes the suboptimum values and finds the optimum tap-length. A large value of Δ_p produces large error width and brings heavy computational complexity whereas a small Δ_p slow down the convergence and makes it difficult to overcome the suboptimum values. [14], [15]

The steady state tap-length is approximately equal to $L_{opt} + \Delta_p$.

In order to maintain the trade-off between convergence and steady state error [15]

$$\hat{e}^2 = \rho\hat{e}^2(n-1) + (1-\rho)\hat{e}^2(n) \tag{22}$$

$$\Delta_p(n) = \min(\Delta_{P,\max}, \upsilon * \hat{e}^2(n)) \tag{23}$$

where ρ is the smoothing parameter and υ is a constant which depends on the characteristics of the unknown plant.

Although different $\Delta_p(n)$ are needed for different applications, whereas for a certain application it can be easily decided in advance according to the noise conditions.

3 Simulation Setup

The simulation is performed for adaptive filter system modeling module as shown in Figure.1. MATLAB 7.7 platform has been chosen for simulation purpose. The input samples *x(n)* are from a white process having mean zeros and variance one. The proposed algorithm deals with a modified version of NLMS algorithm to avoid this slow convergence. $x(n)$ is fed to both unknown plant as well as LMS adaptive filter. The output of unknown filter is mixed with a white noise $t(n)$ such that the SNR remains as 40dB throughout the process. The unknown system is modeled as an Infinite Impulse Response (IIR) system. Because of infinite impulse response it is important to measure the optimum order in a system identification framework which can exactly replicate the performance of the IIR system. For simulations purpose $H_1(z) = \dfrac{1}{1-0.75z^{-1}+0.45z^{-2}}$ and has been fixed as the impulse response of the unknown system to be identified in the framework by an adaptive filter. The algorithm has to find out a filter with minimum co-efficient to completely match the impulse response of the plant in a time varying environment so that the structural design complexity can be minimized.

For comparison of the proposed analysis with present variable tap-length estimation algorithms like the FT-LMS, VT-VLMS has also been implemented along with the proposed algorithm under various noise conditions and parameter variations. The value of $\Delta_{P,\max}$ is kept fixed at 100, υ at 0.5 for the IIR system and δ at 1. For FT-LMS the step size is set at 0.005 and for VT-VLMS the leaky factor varies from 0 to 0.6 [11].

4 Results and Discussion

The adaptive filter response let's say $H(z)$ is matched to the unknown IIR plant in the system identification framework. Fig.1 depicts the performance of converged MSE with tap-length variation for the proposed as well as the FT-LMS [6]-[8] and VT-VLMS [11], [13] algorithms at SNR=40dB. For the proposed variable tap-length algorithm the MSE remains constant after the 20th tap whereas we get suboptimum order of 9 which has been removed by the use of variable error width $\Delta_p(n)$ and choice of parameter υ in the proposed algorithm as mentioned in (23). The optimum tap-length that adjusts the IIR system performance for the FT-LMS and VT-VLMS algorithm is obtained approximately at 48th and 35th tap respectively as shown in Fig. 1. The MSE and tap-length optimization performance of VT-VLMS algorithm is proved to be better than the FT-LMS algorithm but far short than the proposed variable tap-length algorithm. This clearly shows that the improved tap-length learning method analyzed in this paper reduces the structural complexity with 15 to 28 fewer taps. On the other hand it achieves the best MSE performance among all the simulated methods. The MSE performance with number of iterations has been shown in Fig. 2 keeping the SNR fixed at 40dB with averaging over 200 independent runs. The MSE decreases with increased number of iterations as per the general convention [1] but the proposed algorithm clearly outperforms its counterparts by achieving the best MSE performance and 5 to 8dB SNR improvement over 10000 iterations.

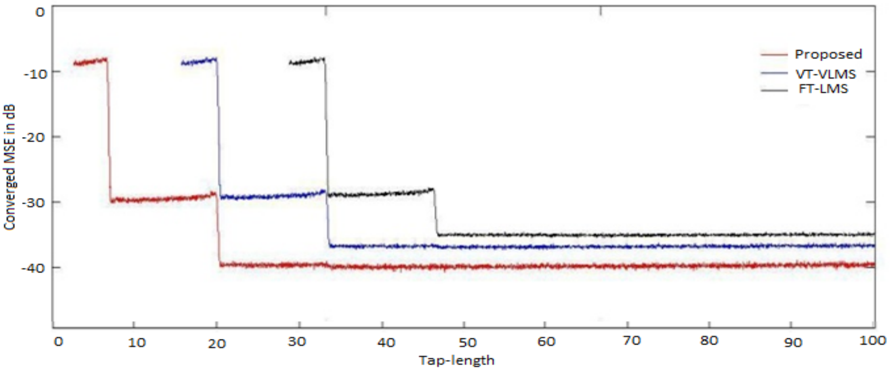


Fig. 1. Converged MSE Vs Tap-length at SNR=40dB

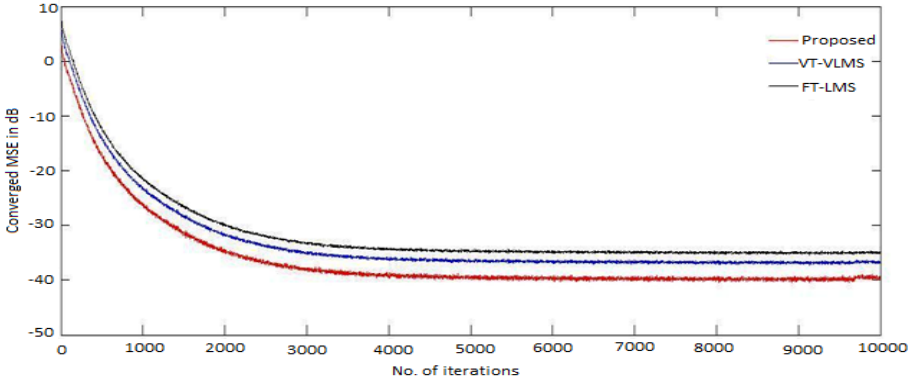


Fig. 2. Converged MSE Vs Iterations at SNR=40dB

The variation of error spacing parameter Δ_P with increased number of iterations, averaged over 200 Monte Carlo runs has been shown in Fig. 3. If Δ_P is being varied with respect to number of iterations then two transient points are noticed between 0 to 10 and 350 to 400 numbers of iterations. These transients are shown in Fig.4 and Fig.5 respectively. It depicts that after some initial transition Δ_P attains steady state value i.e. the optimum tap-length is achieved as the variation between consecutive converged MSE remains at a fixed value. It is discussed and mathematically analysed in Section 2.

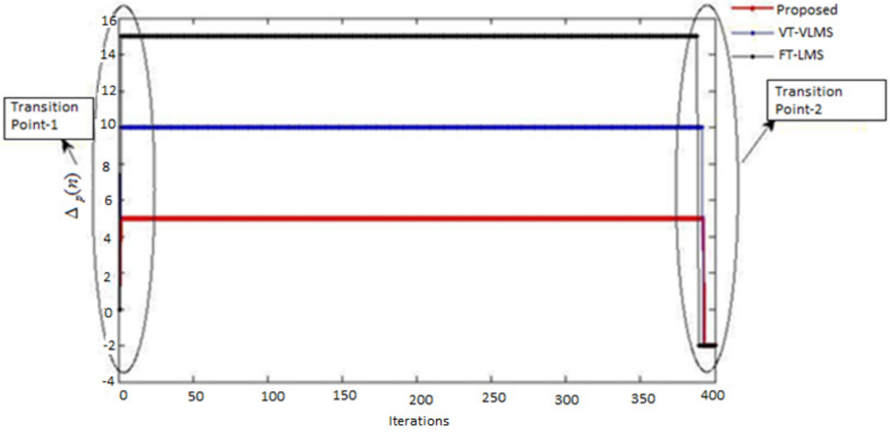


Fig. 3. Δ_P Vs Iterations

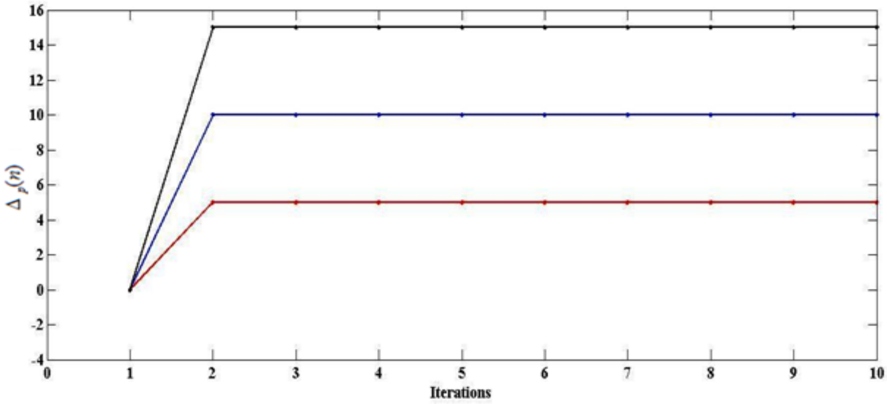


Fig. 4. Δ_P Vs Iterations (Transition point-1)

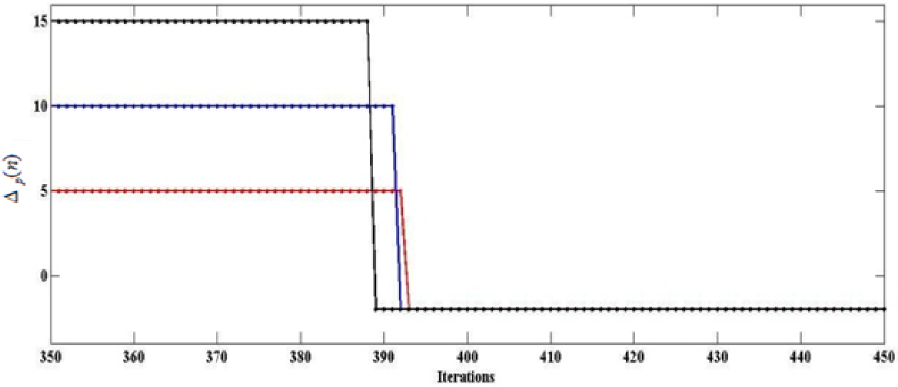


Fig. 5. Δ_P Vs Iterations (Transition point-2)

The variation for the proposed algorithm in comparison to its counterparts is between the minimum if we consider the absolute values. In transition point-1 the Δ_P goes from 0 to 4, 10 and 15 for the proposed, VT-VLMS and FT-LMS respectively in Fig. 5 which shows a steady increase in value before achieving the steady state up to 390 to 400 taps. Then it again decreases to -2 as shown in Fig.5 and attains that value till 5000 iterations which indicates that the desired optimum tap-length has been achieved. The proposed algorithm makes the best use of the variable error spacing parameter which affects the tap-length adaptation up to a large extent.

5 Conclusion

An improved pseudo-fractional tap-length selection for automatic structure adaptation in a dynamic time varying environment has been proposed. The key parameters were set according to the structure adaptation to best adjust the system performance and convergence in an identification framework. The proposed algorithm is compared with the existing tap-length learning algorithms and the improvements are addressed. The computer simulation and results are shown to verify the analysis.

References

1. Widrow, B., Sterns, S.D.: Adaptive Signal Processing. Prentice Hall Inc., Englewood Cliffs (1985)
2. Haykin, S.: Adaptive Filter Theory. Prentice Hall Inc., Englewood Cliffs (1996)
3. Shnyk, J.J.: Frequency Domain and Multirate Adaptive Filtering. IEEE Signal Processing Magazine 9(1), 14–37 (1992)
4. Gu, Y., Tang, K., Cui, H., Du, W.: Convergence analysis of a deficient-length LMS filter and optimal-length to model exponential decay impulse response. IEEE Signal Process. Lett. 10, 4–7 (2003)

5. Mayyas, K.: Performance analysis of the deficient length LMS adaptive algorithm. IEEE Trans. Signal Process. 53(8), 2727–2734 (2005)
6. Gong, Y., Cowan, C.F.N.: A novel variable tap-length algorithm for linear adaptive fitlers. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Montreal, QC, Canada (May 2004)
7. Gong, Y., Cowan, C.F.N.: An LMS style variable tap-length algorithm for structure adaptation. IEEE Trans. Signal Process. 53(7), 2400–2407 (2005)
8. Gong, Y., Cowan, C.F.N.: Structure adaptation of linear MMSE adaptive filters. Proc. Inst. Elect. Eng., Vis., Image, Signal Process. 151(4), 271–277 (2004)
9. Schüldt, C., Lindstromb, F., Li, H., Claesson, I.: Adaptive filter length selection for acoustic echo cacellation. Signal Processing 89, 1185–1194 (2009)
10. Riera-Palou, F., Noras, J.M., Cruickshank, D.G.M.: Linear equalizers with dynamic and automatic length selection. Electronics Letters 37, 1553–1554 (2001)
11. Kar, A., Nath, R., Barik, A.: A VLMS based pseudo-fractional order estimation algorithm. In: ACM International Conference on Communication, Computing and Security, ICSSS 2011 Proceedings, pp. 119–123. NIT RKL (February 2011)
12. Yu, H., Liu, Z., Li, G.: A VSLMS Style Tap-Length Learning Algorithm for Structure Adaptation. In: Proc. IEEE International conference Communication Systems, ICCS 2008, pp. 503–508 (2008)
13. Kar, A., Barik, A., Nath, R.: An Improved Order Estimation of MSF for Stereophonic Acoustic Echo Cancellation. In: Springer International Conference on Information System Design and Intelligent Applications, InconINDIA-12 Proceedings, Vizag, pp. 319–327 (January 2011)
14. Gang, Y., Li, N., Chambers, A.: Steady-State Performance Analysis of a Variable Tap-Length LMS Algorithm. IEEE Transactions on Signal Processing 56(2) (February 2008)
15. Li, N., Zhang, Y., Hao, Y., Zhao, Y.: A new variable tap-length LMS algorithm with variable error width. In: Proc. IEEE Int. Conf. on Signal Processing, pp. 276–279 (October 2008)
16. Li, L., Chambers, J.A.: A Novel Adaptive Leakage Factor Scheam for Enhancement of a Variable-Taplength learning Algorithm. In: Proc. IEEE ICASSP 2008, March 31- April 4, pp. 3837–3840 (2008)

An Efficient and Secure Micro-payment Transaction Using Shell Cryptography

Mayank Tiwari, Rajeshwar Kumar, Shaivya Jindal, Pankaj Sharma, and Priyanshu

ABES Engineering College, Ghaziabad, India
{mayank190590,rajyadav5191,shaivyajindal,priyanshu329}@gmail.com

Abstract. The rapid growth of data communication networks in recent years has led to enormous development .electronic micro payment [6] is one of the most important topics in electronic commerce [4], particularly low cost online payment scenarios and offline payment in rural areas. In this paper we discuss some of the micro payment schemes, observe their merits and demerits and the propose micro payment scheme. we will use shell cryptography in lieu of public key cryptography in e cash schemes and provide security to the micro payment transactions. we compare the improved scheme with others and show that the improved scheme provide better security and efficiency, which enables the schemes viable for real world applications in particular, in resource constraint environment such as mobile payment through handheld devices or customers chip card for debit/credit transaction through point of sale terminal.

Keywords: Micro payment, double spending, electronic commerce, shell cryptography, e cash, Millicent, man in the middle (MitM).

1 Introduction

The electronic payment mechanisms of today have been designed for handling payments of value over five dollars. It, however, seems that these systems cannot manage a large amount of payment transactions below that value level in parallel very well. Together with the envisioned new opportunities in e-commerce, the difficulties have lead to the development of completely new electronic payment [1] mechanisms such as micropayments that have been envisioned to bring solutions to these problems. To meet sufficient security for all participants in electronic commerce, a micropayment system makes it possible to make small payment through electronic communication networks.

Micro payment system provides a means of transferring small monetary amounts and serves as a convenient alternative to traditional payment arrangements. Micropayments refer to low value electronic transactions. Micro-payment system involves: [6], [1]

- A buyer / client
- A vendor / data editor
- One or more brokers / intermediates / billing servers

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 461–469, 2013.

current micro payment systems used by e-commerce sites are not suitable for high volume, low cost transaction, such as charging on a per page basis for website browsing. Micro payment system like e cash suffers from heavy encryption technologies and security risks. There are varieties of micro payment system such as Millicent, e cash, payfair. Most existing micro payment technologies proposed a prototype to date suffer from limitations with communication, security, lack of anonymity or being vendor.Man in the middle attack, high cost cryptographic system (public key) results into the low popularity of micro payment scheme.

To overcome these issues we developed a micro payment scheme by using shell cryptography instead of public cryptography in e cash system. In this type of cryptography key is implanted in the message which results into the high secure transaction. It prevents transaction from man in the middle attack as well as double spending too.

It is noted that sometimes our credit card information, e cash information leaked out due to man in the middle attack which results into the dignity of site, which results into the decrement of customers rate. Some organizations are small, they are not able to use public cryptography for encrypting secret information so we suggest use of shell cryptography which protects man in the middle attack and reduces cost of encrypting secret information.

2 Existing System

We are giving examples of some micro payment system which are considered to be secure.

2.1 E-Cash System

This system [3] is based on what is called a single use token system. The user generates blinded electronic bank notes and sends them to his bank to be signed with his bank's public key (PK)[1]. The bank signs the notes, deducts the amount from the user's account, and sends the signed notes back to the user. The user removes the blinding factor and uses them to purchase at the shop. The shop verifies the authenticity of the bank notes using the bank's corresponding public key and sends them to the bank where they are checked against a list of notes already spent. The amount is deposited into the shop's account, the deposit confirmed, and the shop in turn sends out the goods. All communication over the network is protected by encryption.

The system involves software for both the consumer and the merchant to conduct the transactions. The customer runs a "wallet" program. The user can spend the digital money at any shop accepting e-Cash, without the trouble of having to open an account there first, or having to transmit credit card numbers. Because the received e-Cash is the value involved with the transaction, shops can instantly provide the goods or services requested.

2.2 Disadvantages of E-Cash System

This system uses first the public key for encrypting the message which results into the high processing cost (transaction cost). Small organization is unable to afford it.

Secondly, the E-cash system is also not so secure [2] because it uses public cryptography which is not protected form man in the middle attack (MitM)[8].this attack leads to the threat in the transaction, changes the original content of the message results into the loss of message between merchant and buyeraccount information can also be leaked by MitM attacks.

2.3 Example of an Attack

Suppose Alice wishes to communicate with Bob. Meanwhile, Mallory wishes to intercept the conversation to eavesdrop and possibly deliver a false message to Bob.

First, Alice asks Bob for his public key. If Bob sends his public key to Alice, but Mallory is able to intercept it, a man-in-the-middle attack can begin. Mallory sends a forged message to Alice that claims to be from Bob, but instead includes Mallory's public key.

Alice, believing this public key to be Bob's, encrypts her message with Mallory's key and sends the enciphered message back to Bob. Mallory again intercepts, deciphers the message using her private key, possibly alters it if she wants, and re-enciphers it using the public key Bob originally sent to Alice. When Bob receives the newly enciphered message, he believes it came from Alice.

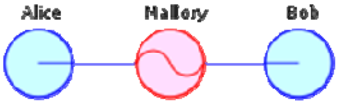


Fig. 1. Illustration of man-in-the-middle attack-[8]

1. Alice sends a message to Bob, which is intercepted by Mallory:

Alice"Hi Bob, it's Alice. Give me your key"-->MalloryBob

2. Mallory relays this message to Bob; Bob cannot tell it is not really from Alice:

AliceMallory"Hi Bob, it's Alice. Give me your key"-->Bob

3. Bob responds with his encryption key:

AliceMallory<--[Bob's_key]Bob

4. Mallory replaces Bob's key with her own, and relays this to Alice, claiming that it is Bob's key:

Alice<--[Mallory's_key]MalloryBob

5. Alice encrypts a message with what she believes to be Bob's key, thinking that only Bob can read it:

Alice"Meet me at the bus stop!"[encrypted with Mallory's key]-->Mallory
Bobit was actually

6. However, because encrypted with Mallory's key, Mallory can decrypt it, read it, modify it (if desired), re-encrypt with Bob's key, and forward it to Bob:

AliceMallory"Meet me in the windowless van at 22nd Ave!"[encrypted with Bob's key]-->Bob

7. Bob thinks that this message is a secure communication from Alice

This example shows the need for Alice and Bob to have some way to ensure that they are truly using each other's public keys, rather than the public key of an attacker. Otherwise, such attacks are generally possible, in principle, against any message sent using public-key technology.

2.4 Millicent:-Millicent

It is a decentralized micro-payment scheme [3], which is designed to allow payments as low as 1/10 of a cent. It uses a form of electronic currency, which is called "scrip". It is designed to make the cost of committing a fraud, more than the value of the actual transaction. It uses symmetric encryption for all data transactions. The principal actors of the scheme are the Broker, the Customer and the Vendor**.** Figure 1 (Source: Pierce (M. Pierce '97) demonstrates the scheme.

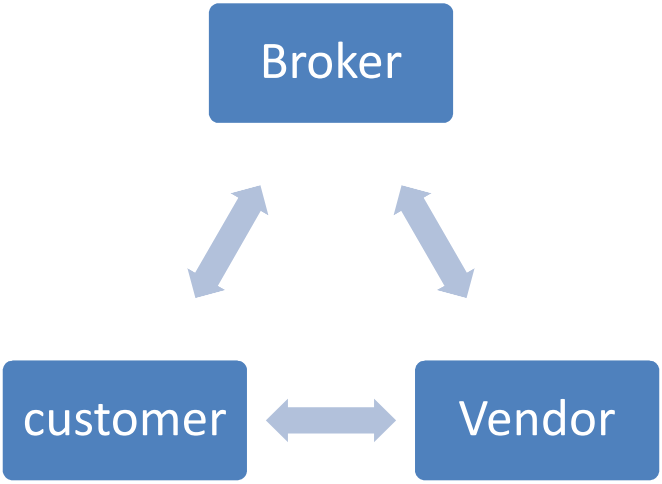


Fig. 2. Millicent's Scheme

1. The Broker: The Broker mediates between the Vendor, using a macro-payment system. The Vendors and Customers in order to simplify the tasks they perform. He acts like a bank and provides the electronic currency ("scrip") for the micro-payments. A Broker, after coming to a deal with the Vendor Broker is then selling the scrip to the customers via macro-payment transactions. As it is seen, Brokers are just

credit intermediates that buy huge amounts of scrip from the Vendors and sell large amounts of scrip to the Customers. During Customer purchases (either from Broker or Vendor), no transactions between the Broker and the Vendors are taking place. Broker mediates between, can either generate his own valid "Vendor-specific" scrip, or buy a large amount of scrip, from

2. The Customer: The Customers buy scrip from the Brokers, using real money, via a macro-payment system. The amount should be sufficient to cover the transaction cost plus to produce financial gain for both the Broker and the Vendor (scrip is Vendor specific). The Customer can then use the scrip to perform micro-payment purchases. No transactions with real money are taking place in any given time between Customers and Vendors.

3. The Vendor: The Vendor is the "data bank". He supplies customers with data, services or both. He accepts his specific scrip as the only method of payment. The scrip was either generated by him (the Vendor) or by a licensed broker. Of course some validation and authentication is necessary to ensure that no double spending will take place. After that the Vendor can transmit the requested data back to the Customer, using a given encryption algorithm for avoiding fraudulent use.

2.5 Disadvantages of Millicent Protocol

Firstly, it uses shared keys. Millicent requires both a vendor and a broker to know the customer shared key. It is okay for the broker to knowabout the key; however, having the vendorknow about CSK requires the vendor to either maintain an extra database or perform an on-line query to the broker.[3]

Secondly, the scrip buyers can be spoofed. Since only the owner of scrip knows about its secret, scripbuyers, including customers, cannot verify scrip.

Thirdly, a long-term relationship is assumed: The Millicent protocol can be inconvenient if acustomer tends to make infrequent purchases with vendors. He needs to go back to hisbroker and exchange for different vendor scrip for each transaction with a new vendor.

3 Proposed System

Security and privacy are very important issues in e transaction and micro payment system. With the advance of technologies cyber-attacks become more proficient. Network security measures are needed to protect data during its transmission.[5]

Cryptography plays a vital role in network security[2] as it allows two parties to exchange sensitive information in a secured manner. In micro payment system, we use public key for encrypting sensitive information during transmission in e cash system, which is considered to be very secure for payment purpose. Intruders become more active and sharp in their technologies so they can easily get our message by man in the middle attack. It results into the loss of privacy and confidentiality of sensitive information which results into the lost to an individual and company as a whole. So we suggest use of imbricate cryptography for sending sensitive information which is less in cost than public key cryptography and secure too.

3.1 Shell Cryptography

Shell cryptography is a new technique that uses the layered approach. It is a typeof symmetric cryptography in which the key is implanted in the message, so the message cannot be recoveredwithout using the correct key. Here the message and the key are inwardly plaited. It involves layers of encryption anddecryption. Since the key is of variable length of the user's choice, it cannot be found by permutation and combination.The algorithm having three layers of encryption, each having its own importance.

Layer1-It is called the mappinglayer and juggles the cracker by jumblingcharacters. Each character will be replace by another one present in the same set. Two type of sets are used

1.repeated character
2.non-repeated character
Equivalence mapping characters are shown in the table

Table For Mapping:

Source file characters	Equivalent mapping char
a/e/i/o/s/t/ {repeated}	p/o/s/i/k/e
b/c/d/e/f/g/h/j/k/l/m/np/q/r/u/v/w/x/ y/z/ {non-repeated}	h/f/b/d/g/c/l/n/j/b/m/u/ y/p/z/q/v/w/x/p
0/1/2/3/4/5/6/7/8/9/ {numerals}	4/6/9/7/0/8/1/3/2/5/
Special characters	Same characters

Layer 2-It is called core encoding layer,The first character of the messageobtained by negation of whole of bijection of char old by negation of ASCII values of key:

$$\text{Char\_new} = \sim[(\text{char\_old}) \texttt{<-->} (\sim k)\ ]$$

Layer 3-It is called the bitmap-conversion layer as it converts ASCII characters into the equivalent binary value and stores the result as a bitmap file.This is done by just obtaining the binary equivalent of the resultant ASCII characters of layer-2 and writing it into a file that is bitmap in nature.

Example how to inbuilt key in message/information-[7]

Message M ={"hello"};
Key K =pei

Layer-1:
From the table, we can replace
M1={“lobbi”};
Layer-2:
M2 =[(M1<-->(~K)];
M2 ={l<-->(~p), e<-->(~o),b<-->(~i), b<-->(~p), i<-->(~o)};
M2 ={“1032~”};

Algorithm for Encryption

1. Get the source file and the password(key) from the user.
2. Choose a mapping character foreach character present in the file usingthe table.
3. Replace the original characterwith the mapping character.
4. Using the password (key) received from the user,encode each character of the message with the successive character of the key.
5. The formula for encoding is:Char_new = ~[(char_old)<-->(~k)]
6. The resultant character is converted into the binary form. This is the end of layer-3.
7. Write the binary values of the new characters in the output bitmapfile.

Algorithm for Decryption

1. Get the bitmap file and the key from the user.
2. Read the binary values from the file and convert back into characters. This is the end of layer-1.
3. From the password (key) received from the user, decode each character with successive character of the key.
4. The formula for encoding is:
   Char_new = ~[(char_old)<-->(~k)]
   ; This is the end of layer-2.
5. Choose a mapping character for each character using the table in the reverse order.
6. Replace the original character with the mapping character. This is the end of layer3.
7. Write the decrypted character in the output file.

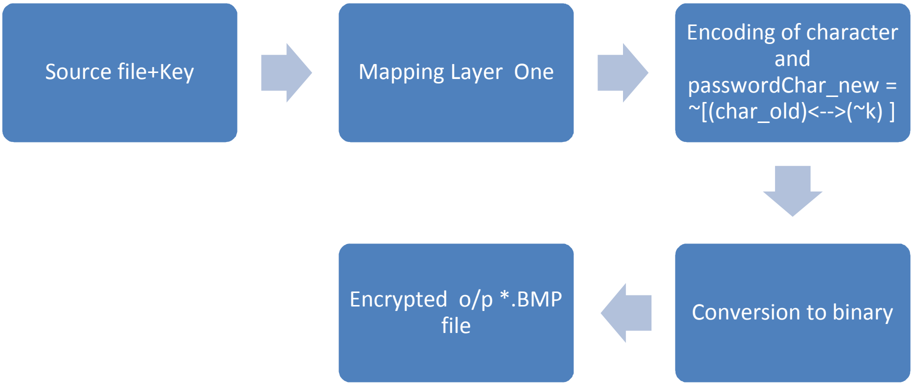

Source file+Key
Mapping Layer One
Encoding of character and passwordChar_new = ~[(char_old)<-->(~k)]
Conversion to binary
Encrypted o/p *.BMP file

Fig. 3. Encryption

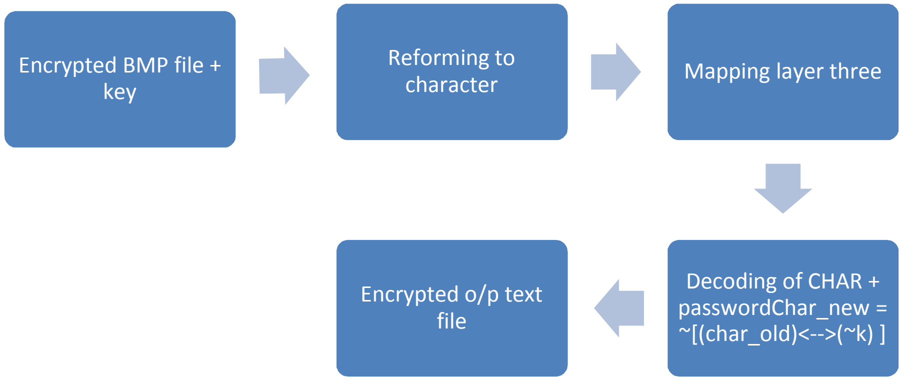


Fig. 4. Decryption

4 System Performance

Any person who wants to crack this system must:

1. Know that the binary values in the bitmap represent ASCII value of the encrypted character.
2. Read the binary values from thebitmap file and convert them into characters.
3. To break the second layer, find the logic that the key is BIJECTED with the characters. (The key should be known.) But finding the key, which is transmitted over a secured channel, is not possible.
4. Then find the mapping characters to break the first layer.Use of the permutation and combination method for finding the key is impossible.Hence the system performance is good.

5 Advantages of Proposed Scheme

1. Confidentiality: No user can access the information without using correct key, it makes e-cash/micropayment system secure.
2. Security: The system is securebecause the key is sent through a secretmedium and the message cannot be recovered without the key.It definitely prevents the fraud of payments and attack of intruders on users account.
3. Protection: It is provided by the key as it controls the access to the message.
4. Incorporated key: Many cryptography techniques use the key for only access control. Our system integrates the key with the message, so the message can be separated from the key only if the correct key is produced.
5. Eonomical: use of public key for encryption in e-cash system is costly so it cannot easily afford by some small companies which makes their payment system insecure, this problem is solved by using shell cryptography, as it is economical and easy to use and is secure too.

6 Conclusion and Future Work

In current scenario many small organizations are growing at very faster rate,they basically use micropayment systems for the payment purpose.User generally trusts them and uses their system for making payments,but sometimes it may happen that more money is deducted from their account ,more money was deducted from their account or loss of personal information of users account.This is due to insecure micropayment scheme.Presently many micropayment schemes are in use like Millicent,E-cash,Cybercoin,Pay-word etc,out of which e-cash is considered to be very secure,but there are many flaws in this scheme also like it is not secure from Man in the Middle attack,as any intruder can easily judge public key and steal users information like of account'personal information etc.it creates bad impact on company's reputation and market value also goes down.To protect MitM we use Shell cryptography[7] in place of public key in E-cash system as it is not Costly as public key,and also secure from Man in the Middle attack i.e intruder cannot guess key easily.This will definitely help small organizations in growing at faster rate,as users trust on them, results into the increase of more profit and more customers.Its best and secure for online micropayments.

References

1. Bayyapu, P.R., Das, M.L.: An Improved and Efficient Micro-payment Scheme.
2. Security Analysis of Micro-Payment Systems, StilianosVidalis School of Computing Technical Report (2004)
3. Comparing and contrasting micro-payment models for E-commerce systems Xiaoling Dai, JohnGrundy, Bruce WN Lo 2
4. Journal of Electronic Commerce Research. 2004 Evaluation of Micropayment Transaction Costs 5(2) (2004)
5. The Siren Song of Internet Micropaymentby: Steve Crocker, Founder of Cyber Cash Article originally posted on iMP: The Magazine on Information Impacts
6. Micropayments Overview by W3C
7. Active Man in the Middle Attacks,A SECURITY ADVISORYA whitepaper from IBM RationaApplication Security Group by Roi Saltzman and AdiSharabani

High Speed Reconfigurable FPGA Based Digital Filter

Navaid Z. Rizvi[1], Raaziyah Shamim[2], Rajesh Mishra[1], and Sandeep Sharma[1]

[1] School of Information and Communication Technology, Gautam Buddha University, India
[2] Department of Electronics and Communication Engineering, JIIT, Noida, India
{navaid,sandeepsharma}@gbu.ac.in, raaziyah.shamim@jiit.ac.in

Abstract. Digital Finite Impulse Response filters are essential building blocks in most Digital Signal Processing (DSP) systems. A large application area is telecommunication, where filters are needed in receivers and transmitters, and an increasing portion of the signal processing is done digitally. However, power dissipation of the digital parts can be a limiting factor, especially in portable, battery operated devices. Scaling of the feature sizes and supply voltages naturally helps to reduce power. For a certain technology, there are still many kinds of architectural and implementation approaches available to the designer. In this paper, a reconfigurable FPGA based pipelined FIR filter is implemented and analyzed. This realized FIR filter is compared for area, power dissipation and data processing rate (throughput). Simulation and compilation of the VHDL code written for the implementation of FIR filters is done using Mentor Graphics ModelSim. For the synthesis targeting to FPGA Xilinx Virtex II Pro XP2VP30 device Xilinx ISE Design Suite 10.1 tool is used. Power estimation is done using Xilinx Xpower tool. FPGA implementation of FIR filter model with respect to power, silicon area, and data processing rate (throughput) is analysed.before and after the abstract. This document is in the required format.

Keywords: DSP, FPGA, FIR Filter, FSM.

1 Introduction

Frequency selective digital filters are required in most DSP (digital signal processing) applications. A typical example is mobile communications where hand-held, battery supplied, devices, such as cellular phones, are used. To obtain a long uptime between recharges of the battery for cellular phones, low power consumption is required. Due to the requirements on high data rates in many communication systems, the corresponding subsystems and circuits must have a high throughput as well. Since significant parts of such communication systems are customer products that are produced in large quantities and are sold at low prices, efficient, fast, and reliable design methods as well as low cost circuit implementations are required. The need for miniaturization of systems also requires subsystems with low power consumption. For such integrated systems, the heat dissipation and the cooling becomes a problem. Low power consumption and circuit area are therefore key design constraints [1].The output data rate of the system also plays a major role in system design. Generally, a tradeoff between different design constraints has to be done to get optimal FIR filter.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 470–478, 2013.

In this paper, a reconfigurable FPGA based pipelined [2] FIR filter is implemented and evaluated for different design constraints such as circuit area, power dissipation and data processing rate. The results are compared with non-pipelined FIR filter.

2 FIR Filter Architecture

The architectures for the FIR filter implementation used are [4]:

- 1-MAC without pipeline FIR filter model
- 1-MAC with pipeline FIR filter model

The design requirements of both the filter implementations are that they must finish 64 tap calculations within time period of 700ns.The description of the two FIR filter architectures implemented is given in the following sections:

2.1 1-MAC without Pipeline FIR Filter Model

The 1-MAC without pipeline FIR filter model implemented comprises of no pipeline as well as no parallel mechanism. The hardware structure of this model is shown in figure 1. There are four main components in this architecture control which are the central control unit, coefficient address RAM, sample RAM and multiplier.

Control component also comprises Finite State Machine (FSM). The function of control unit is to receive the start of the signal of the filter from outside. The data address from the control component is received from coefficient address RAM component and this component also reads out the sample data address and coefficient data from block RAM component, and in turn sends sample data address to the sample RAM component. The function of sample RAM component is to generate sample data in accordance to the address. After this the sample data and coefficient data access the MAC component. Then the process of multiplication of the sample data and coefficient data with MAC component take place and this multiplication process has been repeated exactly the same time as number of taps. Once the control component has been activated, the sending of read address to the coefficient address RAM component does not take place every clock cycle. The sending of second address only take place after the processing of last calculation and then the result is stored in the register of MAC component. This implies that mostly in the clock cycles the most component remains in idle state and earlier process results always remains in the hardware. In this model there is only one data channel between the components. This indicates that for every clock cycle only one sample data and one coefficient data can access the MAC component. This model only comprise of one multiplier accumulator component. The central control unit, implemented, utilizes eight states. The idle state is initial state. After reset signal turns to 1 the FSM turns to reset which in turn assigns in initial value to output. The arrival of new sample set the start coefficient RAM signal active. After the start and FIFO signal have been set to 1, the FSM goes to loadSample state, which makes the reading of new sample from outside and storing it in sample RAM component. After two clock cycle now FSM goes to step1 state, which makes sending of one address to the coefficient address RAM component. In the next clock cycle the reading of coefficient and sample address by the coefficient address RAM component take place and sample RAM component waits for state step1 and step2. In state step1 and step2 enable signal for MAC component has been turn off. In state step3 enable signal to MAC component has been turn on which begins the process of input coefficient and sample data from it. Till the last tap not reached, FSM continues to go to state step1 and processing of new coefficient and sample continues. After reaching the last tap, state waitresult arrived, which implies end of processing period**.** Now the FSM goes to state waitResult for 3 clock cycles and this has been to compromise delay in the MAC component. Finally the last state writeResultToFIFO arise which in return sends the final accumulated result to output register.

2.2 1-MAC Pipeline FIR Filter Model

The 1-MAC pipeline FIR filter model contains pipeline but no parallel mechanism. The hardware structure is depicted in figure 2. The basic principle for this model has been that with the initialization of FIR filter every clock cycle control component sends read addresses to coefficient address RAM component, which in turn discarded the waiting time for the output from previous data. The next read address follows the last read address in the coming clock cycle which implies read address updated continuously every clock cycle till the final stage reached. The data transmission between the components has been continuous. This model contains one MAC component. The basic function and structure of the sampleRAM Coefficient Address RAM and multiplier accumulator components used in this model is the same as for sample RAM component in 1-MAC without pipeline FIR filter model.

The central control unit, implemented by [3], utilizes seven states in this model. After reset signal turns to value 1, the FSM goes to idle state which has been the initial state. In this state, all the initial values have been assigned. After the arrival of new sample, the input signals start and fifoEmpty have been set to value 1. Due to impact of these signals, the FSM changes its state to loadSample state. In loadSample state, new sample has been read from DSP processor and stored into sample block RAM component. Due to pipeline1 state the control component emits read address continuously followed by every clock cycle. These results in forming data flow through address bus which flows through the control component, coefficient address RAM component and sampleRAM component. This state repeats until the proposed tap number processed. After this state, there have been read addresses (as the number of taps) in coefficient address RAM component. The pipeline2 state helps in achieving the delay processing caused from MAC processing. With consideration of delay, the MAC component multiplies the coefficient and sample, and accumulated them from the fifth clock cycle of pipeline1 state till the last cycle of pipeline2 state. The MAC signal is active throughout this both states. When last coefficient data and sample have been sent to MAC, FSM goes waitResult state. Duration of this state has been equal to the delay in the multiplier accumulator component. The last state has been writeResultToFIFO state and then FSM turns to the first state idle until the control signal will reset to 0 value.

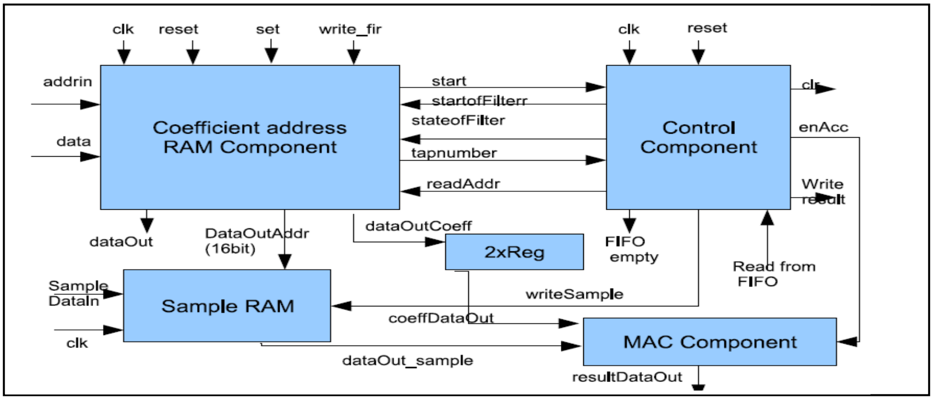


Fig. 1. Hardware architecture of 1-MAC without pipeline FIR filter model

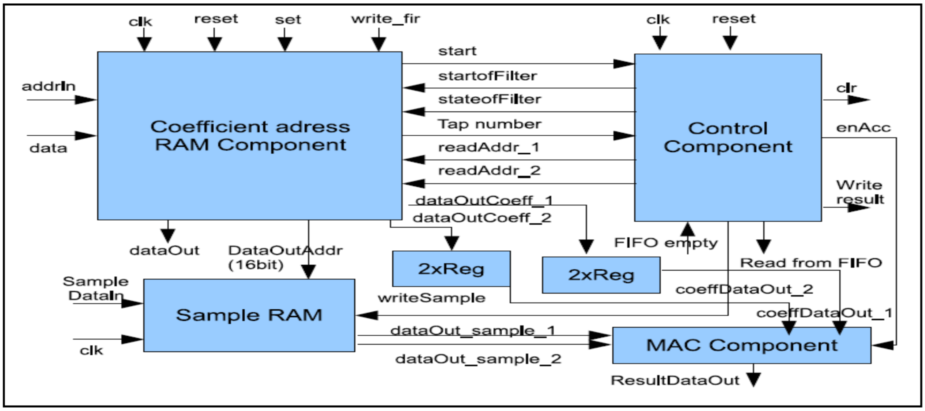


Fig. 2. Hardware architecture of 1-MAC pipeline FIR filter model

3 Simulation and Synthesis of FIR Filters

Firstly, the VHDL code for a particular design, the develop logic is written and tested. Simulation and Compilation is done using Mentor Graphics ModelSim PE Student Edition 6.4a. ModelSim [5] is a simulation and a debugging tool for VHDL, Verilog, and other mixed-language designs from Mentor Graphics [4]. It includes a simulator for VHDL code with complete debugging environment, a source code viewer/editor, waveform viewer, design structure browser, list window, and a host for other features designed to enhance productivity. The design in this paper is targeted for a Xilinx Virtex II Pro XP2VP30 FPGA [6].The Xilinx Virtex-II Pro is a 130nm CMOS nine-layer (copper) FPGA device. It has embedded blocks, consists of a two-dimensional array of configurable logic blocks (CLBs), and programmable interconnect resources. Each CLB has four identical sub-blocks, called slices. A slice comprises two four-input LUTs (whose function are used to map four-input Boolean logic), two Flip

Flops, gates (two AND2, one OR2, two XOR2) which are used to implement arithmetic functions such as carry chains, multiplexers, inverters, and buffers. After synthesis four multiplexers are used as mapped multiplexers and other multiplexers as configurable multiplexers for routing inside the slice. The inverters and the buffers are needed to implement the different clock types allowed in the FPGA. The four slices of one CLB share a common programmable I/O crossbar for communicating with a switch matrix (to realize communication with others CLBs and to route a signal). A CLB, together with its associated switch matrix, is called tile. The Virtex-II Pro contains platform FPGAs for designs that are based on IP cores and customized modules. The family incorporates multi-gigabit transceivers and PowerPC CPU blocks in Virtex-II Pro Series FPGA architecture. It empowers complete solutions for telecommunication, wireless, networking, video, and DSP applications. Here the Block selected RAM and memory modules contribute around 18 Kb storage elements of True Dual Port RAM. The RocketIO Transceiver is a Full-Duplex Serial Transceiver (SERDES) capable of Baud Rates from 600 Mb/s to 3.125 Gb/s. The Select RAM+ Memory module provides upto 1378 Kb of distributed SelectRAM + resources. The Arithmetic functions are performed from 18-bit x 18-bit multiplier blocks. This FPGA contains High-Performance Clock Management Circuitry.

4 Results and Discussions

4.1 Simulation Results for Control Component Used for without Pipeline Mechanism

The testbench of Control component used for without Pipeline mechanism includes starting with start single assigning to value of 1. At 1530 ns the fifoempty has been assigned to 0 value which marks change of idle to loadSample state. After this FSM goes through state step1, waitForCoeff1, step2, and step3. The four taps proccessing implies FSM to process four times. Now the FSM reached to states waitResult and writeResultToFIFO. After crossing these two states one complete cycle is finished. The function of ennAcc signal is to active the MAC component. The simulation result is depicted in figure 3.The output indicates correctness in functionality.

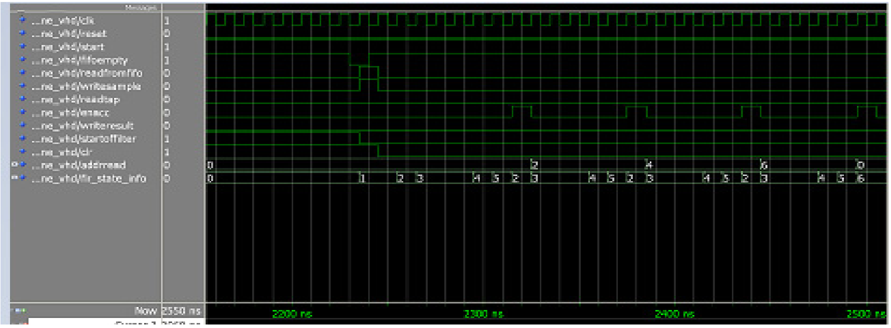


Fig. 3. Simulation waveform for Control component used without Pipeline mechanism

4.2 Simulation Results for Control Component Used for Pipeline Mechanism

Pipeline mechanism only involves the states - idle, loadsample, waitResult and, writeResultToFIFO without pipeline mechanism. The testbench includes starting of start signal. At 2230 ns fifoEmpty signal is assigned value 0, FSM changes its state from idle to loadsample1. After this it passes the state of loadsample1, pipeline1, pipeline2, waitResult and writeResulttoFIFO. The pipeline1 state repeats for 32 times in between 2260 ns to 2580 ns. This state also results in addread to read address from value 0 to 31 in increasing order. At loadsample state readFromFIFO and writeSample signal are also active. The simulation result is depicted in figure 4.

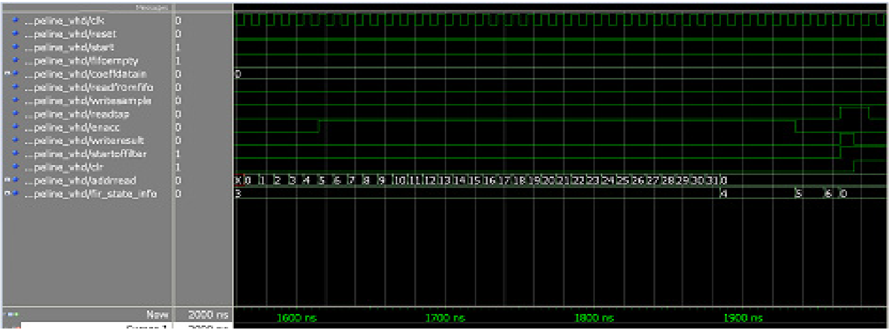

Messages
...peline_vhd/clk 0
...peline_vhd/reset 0
...peline_vhd/start 1
...peline_vhd/fifoempty 1
...peline_vhd/coeffdatain 0
...peline_vhd/readfromfifo 0
...peline_vhd/writesample 0
...peline_vhd/readtap 0
...peline_vhd/enacc 0
...peline_vhd/writeresult 0
...peline_vhd/startoffilter 1
...peline_vhd/clr 1
...peline_vhd/addrread 0
...peline_vhd/fir_state_info 0
0
X 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 0
3 4 5 6 0
Now 2000 ns
1600 ns
1700 ns
1800 ns
1900 ns

Fig. 4. Simulation waveform for Control component used for without Pipeline mechanism

4.3 FIR Filter Model - Simulation Results

For both the filters, the testbench is written considering one sample reads for the input sample file every 700ns. The function of the main model is to process the data and emit them to 32 bit port resultdataout. For every data processing period there is only one new sample to read. Four taps are processed per period. The simulation waveforms are shown in figure 5 and figure 6.

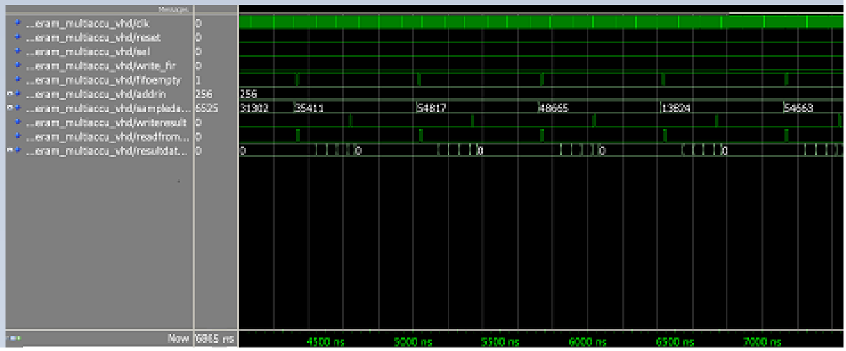

Messages
...eram_multiaccu_vhd/clk 0
...eram_multiaccu_vhd/reset 0
...eram_multiaccu_vhd/sel 0
...eram_multiaccu_vhd/write_fir 0
...eram_multiaccu_vhd/fifoempty 1
...eram_multiaccu_vhd/addrin 256
...eram_multiaccu_vhd/sampleda... 6525
...eram_multiaccu_vhd/writeresult 0
...eram_multiaccu_vhd/readfrom... 0
...eram_multiaccu_vhd/resultdat... 0
256
31302 35411 54817 48665 13824 54563
0 0 0 0 0
Now 6865 ns
4500 ns
5000 ns
5500 ns
6000 ns
6500 ns
7000 ns

Fig. 5. Simulation waveform for 1-MAC without pipeline FIR filter model

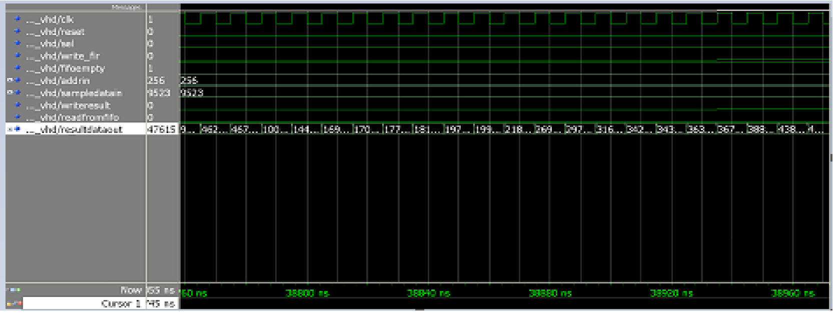


Fig. 6. Simulation waveform for 1-MAC with pipeline FIR filter model

4.4 Synthesis Results Targeted to FPGA for Pipeline and Non-pipeline FIR Filter Models

For The synthesis report, device utilization report in percentage and timing report for pipeline and without pipeline FIR filter models are depicted in table 1, table 2, and table 3 respectively.

Table 1. Timing Report

	Without pipeline	With Pipeline
Minimum period	7.210ns	6.747ns
Minimum input arrival time before clock	4.061ns	4.043ns
Maximum output required time after clock	5.246ns	5.246ns
Maximum combinational path delay	6.654ns	6.654ns
Maximum Frequency	138.706MHz	148.225MHz

Table 2. Device Utilization Report

FPGA Resources	Without Pipeline	Pipeline
Slices	1%	1%
Slices Flip Flops	1%	1%
4 input LUTs	1%	1%
IOBs	44%	44%
BRAM	14%	3%
MULT18x18s	0.74%	0.74%
GCLKs	6%	6%

Table 3. Synthesis Report

FPGA Resources	Without pipeline	Pipeline
RAMs	5	5
-64x16 bit single port RAM	1	1
-64x32 bit dual port RAM	4	4
Multipliers	1	1
16x16 bit multiplier	1	1
Adders/Subtractors	5	5
14-bit addsub	1	1
14-bit subtractor	2	2
3-bit adder	1	1
6-bit adder	1	1
Counters	1	1
14-bit up counter	1	1
Accumulators	1	1
33-bit up accumulator	1	1
Registers	63	63
1 bit register	39	38
14 bit register	2	2
16 bit register	6	7
3 bit register	1	1
32 bit register	11	11
5 bit register	1	1
6 bit register	3	3
14 bit comparator less	1	1
3 bit comparator less	1	2
6 bit comparator less	1	1
32 bit 4 to 1 MUX	1	1

4.5 Power Estimation Results

The power estimation [7] results targeted to FPGA for the two FIR filter models are depicted in table 4. The table shows power (in watts) consumed by clock, logic, signals, IOs as well as Total Quiescent, Total Dynamic and Total Power dissipated by the Control components in the two models. The throughput (time per sample) at 100 MHz clock frequency is calculated to be 70ns/sample for without pipeline and 12.1 ns/sample for with pipeline FIR filter model. Thus, adding pipeline mechanism has increased the data processing rate by five to six times. There is very little difference in the FPGA resources (i.e. area) consumed between the two filter architectures. Adding pipelinism has reduced power dissipation by some amount. The results clearly emphasize that FIR filter with pipeline is a better choice for FPGA filter implementation than FIR filter without pipeline.

5 Conclusion

Single MAC FIR filter models with and without pipeline mechanisms are implemented in VHDL, simulated and synthesised. The two architectures are evaluated for resources (area) consumed, power dissipation and data processing rate. The filter architecture employing pipeline mechanism outperforms the one without pipeline mechanism.

References

1. Khorbotly, S., Carletta, J.E., Veillette, R.J.: A Methodology for Implementing Pipelined Fixed-Point Infinite Impulse Response Filters. In: 41st Southeastern Symposium on System Theory, March 15-17, University of Tennessee Space Institute Tullahoma, TN (2009)
2. Shaw, A., Ahmed, M.: Pipelined recursive digital filters: a general look-ahead scheme and optimal approximation. IEEE Trans. On Circuits and Systems II: Analog & Digital Signal Processing 46(11), 1415–1420 (1999)
3. Wei, C.-H., Hsiao, H.-C., Tsai, S.-W.: FPGA Implementation of FIR Filter with smallest Processor. IEEE (2005)
4. Pirsch, P.: Architectures for Digital Signal Processing. John Wiley & Sons, Chichester (1998)
5. ModelSim User Manual, Software Version 6.4a, Mentor Graphics Corporation (2008)
6. Xilinx. Vitex-II Pro and Virtex-II Pro X FPGA User‘s Guide, Xilinx, Inc. (2007)
7. Xilinx. Xilinx Power Estimator User Guide, Xilinx, Inc. (2007)
8. Xilinx, Jiang, Z., Willson, A.N.: Efficient digital filtering architectures using pipelining/interleaving. IEEE Trans. Circuits Systems–II 44, 110–118 (1997)

Applicability of Rough Set Technique for Data Investigation and Optimization of Intrusion Detection System

Sanjiban Sekhar Roy[1,*], V. Madhu Viswanatham[1], P. Venkata Krishna[1], N. Saraf[1], A. Gupta[1], and Rajesh Mishra[2]

[1] School of Computing Science and Engineering, VIT University,
Vellore, India
[2] School of Information & Communication Technology,
Gautam Buddha University, Greater Noida, India
{s.roy,pvenkatakrishna,vmadhuviswanatham}@vit.ac.in,
{nkhlsrf,akku.gupta,raj25mis}@gmail.com

Abstract. The very idea of intrusion detection can be perceived through the hasty advancement following the expansion and revolution of artificial intelligence and soft computing**.** Thus, in order to analyze, detect, identify and hold up network attacks a network intrusion detection system based on rough set theory has been proposed in this article. In this paper we have shown how the rough set technique can be applied to reduce the redundancies in the dataset and optimize the Intrusion Detection System (IDS).

Keywords: Intrusion Detection, Network Attacks, Rough Set Theory.

1 Introduction

One of the core necessities of modern day is a closely protected internet service. It enables transmission of huge amounts of data every day without proper security parameters in place. Thus, in recent years unauthorized access to information has been one of the biggest concerns for many big organizations by putting them at risk. Therefore, efficient detection of such threats has become a high priority task.

In the midst of the arrival of intrusion detection technology, the activities of networks can be exemplified by means of uncertainty, intricacy, variety and vibrant tendencies. An intrusion detection system (IDS) scrutinizes all inbound and outbound network activities and identifies mistrustful patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. IDS can be sort out into two kinds: Network-based systems (NIDS) and Host-based systems (HIDS). Within a HIDS, the system analyzes the activity on every being computer or host. In a network-based configuration, the entity packets flowing all the way through a network are examined. NIDS be able to perceive malicious packets

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 479–484, 2013.

that are deliberated to be unnoticed by a firewall's basic filtering rules. Here we have concentrated only on NIDS. This paper implements network intrusion detection system using rough set theory. Intrusion detection data sets are generally very large and can result in redundant records which again results in uncertainty, consequently it isn't possible to sift through all the data manually. In order to produce an efficient dataset, reduction techniques of rough sets have been applied. Roy et al. [2] has proposed a technique for IDS, likewise here also KDD'99 Cup Data set has been used for implementation in this paper**.** It is first, analyzed using Rough Sets [1],[3] during which a rule list is generated. Afterwards these rules act as decisive parameters for determining the threat and notify in case of any intrusion. Intrusion data diminution, rule assortment, feature selection by rough set theory is browbeaten to improve detection exactness, preprocess data and trim down false alarm and illusory alarm.

Our article has been prearranged in the following way. Section 1 explains Rough Set theory followed by Intrusion Detection Systems in Section 2. Thereafter Section 2.1 places of interest the KDD'99 Cup Data set that is used. Implementation of the concept is elaborated in Section 2.2 and 2.3, after which the paper is concluded with the advantages of the system implemented and other future prospects.

1.1 An Overview of the Rough Set Theory

The theory of Rough set is a new mathematical tool to deal with intelligent data analysis and data mining [1] proposed by Z Pawlak. Roy et al. [4] has shown that this theoretical framework is based on the concept that every object in the universe is attached with some kind of information. Set theory is a great help to the computer science research and theory of Rough set is an extension to that. It's a mathematical tool to deal with inexact, uncertain or vague data, which are part of artificial intelligent system. It includes algorithms for generation of rules, classification and reduction of attributes. It is hugely used for knowledge discovery [1], [7] and reduction of knowledge. The theory of Rough set has got many important applications.

1. Information/Decision Systems (Tables) :

   IS is a pair (*U, A*). *U* is a non-empty finite set of objects. *A* is a non-empty finite attributes sets such that $a : U \rightarrow V_a$ for every $a \in A$. V_a is called the value set of a[1].
2. Indiscernibility:

   Here we say, *IS = (U, A)* is an information system, such that $B \subseteq A$, there is an connected equivalence relation:

   $$IND_{IS}(B) = \{(x, x') \in U^2 \mid \forall a \in B, a(x) = a(x')\}$$

   where, $IND_{IS}(B)$ is known the *B*-indiscernibility relation.
3. Set Approximation :

Again, if *IS* = *(U, A)* and if $B \subseteq A$ and $X \subseteq U$,then we be able to approximate *X* in *B* by constructing the B-lower and B-upper and boundary region of *X,* can be denoted as

$$\underline{B}X = \{x \mid [x]_B \subseteq X\}$$

$$\overline{B}X = \{x \mid [x]_B \cap X \neq \phi\}$$

$$BN_B(X) = \overline{B}X - \underline{B}X \text{ ,.}$$

4. Reduct and Core :

Let B be a subset of A and let a belong to B.

We say that a is dispensable in B if I(B) = I(B– {a}); otherwise a is indispensable in B.

$$Core\,(B) = \cap Red\,(B),$$

Where Red (B) is the set off all reducts of *B*.

2 Intrusion Detection System

An intrusion detection system (IDS) scrutinizes [5], [6] all inbound and outbound network activities and identifies mistrustful patterns that may indicate a network or system attack from someone attempting to break into or compromise a system. We have shown a pictorial view of an intrusion detection system.

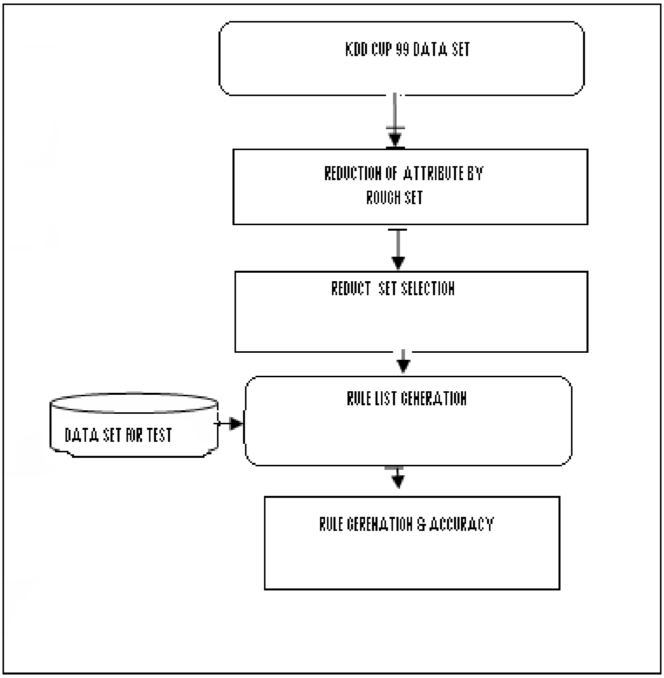


Fig. 1. Intrusion detection system

The following steps are involved in an intrusion detection system.

1. Network data – The network data is collected by analysing and consolidating the network traffic. In this paper, KDD’99 Cup Data set has been used.
2. Attribute reduction and selection of reduct – Since the data set is comparatively huge, it is necessary to filter out the required data from the unclean dataset. This has been done by selecting the core attributes and necessary data from the entire data using rough set reduction theory.
3. Rule generation – Using the reduct and core obtained after applying rough set technique on the data set, decision rules are generated and matched with the already trained history data.
4. Best rule selection – Thus the best rules from the list are selected to form the decision parameters for the system.
5. Alarming – To inform users in case of any intrusion attack or abnormal activity

2.1 KDD’99 Cup Data Set

The tedious job was to create an intrusion detector network, a predictive model having capacity of differentiating between “bad” connections, called intrusions or attacks, and “good” normal connections. This database is consisted of a standard data set required to be audited. This set includes a long range of intrusions replicated within a pro-military environmental network.

The KDD’99 is the most widely used data set for the evaluation of Intrusion Detection Systems. The data set prepared by Stolfo et al. has been populated with about 4 gigabytes of compressed raw tcpdump data captured in DARPA’98 IDS evaluation program. KDD training dataset consists of approximately 4,900,000 single Connection vectors each of which contains 41 features and is labelled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories:

1. Denial of Service Attack (DoS): The legitimate user is denied access in this type of attack or the attacker makes the memory/computing resource too busy to respond.

2. User to Root Attack (U2R): the attacker logs in as a normal user initially. However,, he gradually gains root access to the system thus making it vulnerable.

3. Remote to Local Attack (R2L): an attacker who though having the ability to send packets to a machine over a network tends to exploit susceptibility of that machine for achieving local access as a user, for not having an account on that machine.

4. Probing Attack: a step taken to collect information in order to sabotage the network related security.

2.2 Experimental Result

10% of the KDD Data set has been used for implementation to help form the decision algorithm.

Table 1. We have used three types of attributes and they are as follows

Feature name	Description	Type
Duration	length (number of seconds) of the connection	continuous
Protocol type	type of the protocol, e.g. tcp, udp, etc.	discrete
Service	network service on the destination, e.g., http, telnet, etc.	discrete
Source bytes	Number of data bytes from source to destination	continuous
Destination bytes	Number of data bytes from destination to source	continuous
Flag	normal or error status of the connection	discrete

The Decision Attribute of the dataset was selected to be the FLAG attribute.

It can be classified into two values a) Successful b) Rejected based on the intrusion in the network.

Table 2. Contains the Reduct Set

Size	Pos. Reg.	SC	Reducts
5	0.96	1	{duration, protocol_type, service, dst_host_rerror_rate, count}
3	0.96	1	{ dst_host_rerror_rate, src_bytes, count }
3	0.96	1	{ service, src_bytes, count }
4	0.96	1	{ protocol_type, dst_host_rerror_rate, dst_bytes, count }
4	0.96	1	{ protocol_type, service, dst_bytes, count }

2.3 Decision Algorithms

1. If (service = http) & (dst_host_rerror_rate = 0) then (flag = SF) [34 matches]
2. If (protocol_type = udp) then (flag = SF) [32 matches]
3. If (dst_host_rerror_rate = 1) then (flag = REJ) [27 matches]
4. If (count = 2) then (flag = SF) [15 matches]
5. If (src_bytes = 105) then (flag = SF) [13 matches]
6. If (dst_bytes = 147) then (flag = SF) [7 matches]
7. If (protocol_type = tcp) & (service = private) & (count = 1) then (flag = REJ) [3 matches]
8. If (protocol_type = tcp) & (protocol_type = tcp) & (dst_bytes = 0) & (count = 1) then (flag = REJ) [3 matches]

3 Conclusion

The most essential requirement in these days of globalization is the maintenance of a well guarded internet service as that helps in transmitting huge amounts of data every day. Now, as recently unchecked access to information has turned out to be one of the biggest concerns, so development of an efficient technique to detect such threats has become very necessary. Hence, to analyze, detect, identify and hold up network attacks in an effective manner a network intrusion detection system based on rough set theory has been proposed in this article.

Acknowledgment. We would like to thank to the higher ranking management of VIT University for their kind help and encouragement towards our research.

References

1. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
2. Roy, S.S., Viswanatham, V.M., Krishna, P.V.: Intrusion Detection Data Analysis Using Dominance Based Rough Set Annals. Computer Science Series 10 (2012)
3. Saquer, J., Deogun, J.S.: Concept approximations based on rough sets and similarity measures. International Journal of Applied Mathematics and Computer Science 11, 655–674 (2001)
4. Roy, S.S., Rawat, S.S.S.: Core generation from phone calls data using rough set theory. Annals Computer Science Series 10, 29–32 (2012)
5. Cheng, X., Xiang, B., Zhang, Y.L.: Attribute Reduction Method Applied to IDS, Information engineering Institute, Jingdezhen Ceramic Institute. In: International Conference on Communications and Mobile Computing (2010)
6. Zainal, A., Maarof, M.A., Shamsuddin, S.M.: Feature Selection Using Rough Set in Intrusion Detection. In: IEEE Region 10 Conference, TENCON (2006)
7. Roy, S.S., Gupta, G., Sinha, A., Ramesh, R.: Cancer data investigation using variable precision Rough set with flexible classification. In: Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, ACM Digital Library, pp. 472–575 (2012)

NTTM: Novel Transmission Time Based Mechanism to Detect Wormhole Attack

Kumar Chanchal and D.K. Lobiyal

School of Computer and Systems Sciences,
Jawaharlal Nehru University, New Delhi.
{chanchal.ck1,lobiyal}@gmail.com

Abstract. The cooperative nature and absence of infrastructure gives rise to lot of scope for research in the area of Mobile Ad-hoc Networks (MANETs). The dynamic topology, absence of central control and broadcast nature of communication open security threats for MANETs. Many security attacks have been identified by the researchers, but wormhole attack is one of the most devastating attacks. Novel Transmission Time based Mechanism (NTTM) detects wormhole attacks by keeping every node under the surveillance of its neighbors. Based on the Round Trip Time (RTT) computed by each node on a route, the source node computes RTT between each neighbor. If the RTT between a pair of nodes is more than the threshold value, it is assumed that there is wormhole attack between these nodes. The performance of NTTM is evaluated using dynamic source routing (DSR) protocol under wormhole attack.

Keywords: MANETs, Wormhole attack, RTT, NTTM, DSR.

1 Introduction

A number of threats and their countermeasures in the area of MANETs have been identified by the researchers. In all possible threats, wormhole attack is the most devastating attack and it is lunched at the time of route discovery phase. Two malicious nodes located at different positions form a secret tunnel (wormhole). One malicious node captures the control as well as data packets from the location near the source. It directs these packets to move through the tunnel towards other colluding nodes placed at other locations in the network. These colluding nodes in turn drop or replay back the packets into the network.Although the length of the tunnel is large but it creates an illusion that there exists a shortest path between the source and the destination. Therefore, source node chooses the shortest path through the tunnel to send its data. By using this link, malicious nodes launch variety of attacks against the data flow such as selective dropping, reply attacks, eavesdropping etc.[1,2].

Wormhole or tunnel can be formed either by packet encapsulated channels (also known as In-band-channel) or out-of-band channels. In packet encapsulated channels, malicious node captures the route message and inserts it in data packet payload. This packet is transmitted using legitimate nodes towards other malicious node. The

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 485–495, 2013.

malicious node draws the routing message from packet payload and further braodcast it to the destination. In Out-of-Band channel, a special channel either a direct wired link or a long range wireless link can be used to form tunnel between malicious nodes.

Wormhole attacks can be classified broadly into hidden wormhole attacksand exposed wormhole attacks. In hidden wormhole attacks, legitimate nodes are unware of malicious nodes. The malicious nodes don't upadate hop count field in packet header i.e. only legitimate nodes change the hop length during route establishment. In exposed wormhole attacks, they are aware of the fact that malicious nodes are forwarding packets. But they actually do not know that they are malicious nodes. Here, attackers neither modify packet header nor the content of the packet. The nodes simply add its own MAC address in the header of the packet and forwards it. By extracting information from packet header, the malicious node obtains necessary information about the sender of a packet [2-4].

In this paper we have proposed an effiecent and secure mechanism to detect wormhole attack known as NTTM over DSR protocol. NTTM uses route discovery mechanism of DSR protocol with some modifications. The RTT between the destination and each node in the path is computed by the neighbor nodes of the destination. Finally, RTT value is forwarded to the source that will declare about the presence of wormhole aftersome computation.

The remainder section of this paper is arranged as follows. Section 2 describes the work that has been done related to the detection of wormhole attack. The problem is defined in section 3. Section 4 provides the proposed work in detail. The performance of NTTM evaluated by using simulations is presented in section 5. Section 6 concludes the work carried out along with discussions on possible future extensions.

2 Related Works

Many methods have been proposed to defend against wormhole attacks in which either existing protocol is modified like AODV or special hardware is used such as directional antennas [5,6].

Su et al. [1] introduced Wormhole Avoidance Routing Protocol (WARP) based on AODV protocol. WARP keeps multiple link disjoint paths into consideration. The malicious nodes have great tendency to get involved in the path discovery process. WRAP uses this characteristic of malicious nodes to detect the wormholes attack. Each node records anomaly value of its neighboring nodes. The probability of involvement of a node among the multi joint path is known as its anomaly value. If the anomaly value of a particular node exceeds the threshold value, its neighbor node declares it as a malicious node and it further, discards all the requests coming from that node to form a route. It may be possible that legitimate nodes may be considered as malicious and isolated by their neighbors.

Phuong et al. [3] proposed transmission time based mechanism using Round Trip Time (RTT) of packets between each neighbor to detect wormhole attacks. It is tested over AODV protocol. The destination node modified the format of RREP packet in AODV by adding an extensional part. The size of extensional part is according to hop count field in RREQ packet. A source node calculates RTT between it and eachof theneighbors after obtaining the information regarding the RTT time of each node. If RTT value between any pair of neighbor's nodes exceeds the threshold time limit, it shows the existence of the wormhole on the route. There are the high chances of inserting false information by the malicious nodes.

Hu et al. [7] introduced a general mechanism to defend against wormhole attack in which small amount of information called Leash is added into a packet. Leash restricts a packet's maximum allowed transmission distance. Leash can be geographical or temporal [4, 7]. To form geographical leash, every node should know their positions. After receiving a packet, receiver calculates the maximum distance between sender and itself. It also records its receiving time. If the distance exceeds the maximum limit, the node discards the packets. The temporal leash uses a special off-the-shelf hardware based on LORAN-C, and WWVB in place of loose clock synchronization to provide tight time synchronization. It is implemented through Timed Efficient Stream Loss tolerant Authentication (TESLA) with Instant Key (TIK) disclosure protocol. It requires extremely tight time synchronization and Global Positioning System (GPS).

Wing et al. [8] brought in an End-to-End detection of wormhole attack which calculates "minimum hop count" from every node to the destination. The EDWA can work with both AODV and DSR routing protocols with a constraint that only destination node can reply to the RREQ packet. Based on the position of source and the destination, source node calculates the shortest path in terms of hop length. If the estimate hop count value is greater than hop count value of RREP packet, it is assumed that there is wormhole somewhere on the path. To identify the end point of wormhole, source node sends Trash packet. If the large increment (more than one) is observed in hop count between any pair of neighbors, it shows that this pair of nodes comprises the end points of a wormhole. The EDWA mechanism performs better when the source and destination are not far away.

3 Proposed Work

This section describes the detail of proposed work. The RRT experienced by the packet while travelling through tunnel is large. This characteristic is considered as key point for this proposed model.

3.1 Route Discovery

In Fig. 1 shown below, node *S* wants to send data to node *D*. The sending node *S* triggers route discovery and broadcasts a RREQ packet into network. The source node stores T_{RREQ} (time of broadcasting RREQ packet) of the RREQ packet. Each RREQ packet contains address of sender, receiver as well as broadcast ID which are used to discard the duplicate packets received at a node.

Node 1, 2 and 3 receive RREQ packet broadcast by node S. Each Node matches destination address recorded in RREQ with its IP address (step 6). The steps mentioned in the brackets are of NTTM algorithm described next. None of them (1, 2 and 3) is destination, so they processes the RREQ packet and put their IP address in

RRL. They rebroadcast the RREQ packet in to network. The node *S* hears the RREQ packet again as shown in Fig.1 and find its address at (n-1)th position in RRL (step 9). Therefore, it stores T_{RREQ} of each node and drops the RREQ packets broadcasted by nodes (step10).

At node 4 two RREQ packets are received through node 1 and node 2. Suppose RREQ packet broadcast by node 2 reaches first at node 4, which will discard RREQ packet (step 5) broadcast by node 1 because of duplication (same broadcast ID, same originator ID). Similarly, node 5 receives RREQ packet from node 3 and node 2. RREQ is received first through 2. They are neither destination nor the RREQ packet has their address in RRL. Therefore, they append their address in RRL tail and broadcast RREQ packet again. Now node 2 satisfies the condition of step 9 for both node 4 and node 5. Hence, node 2 stores T_{RREQ} of the both nodes.

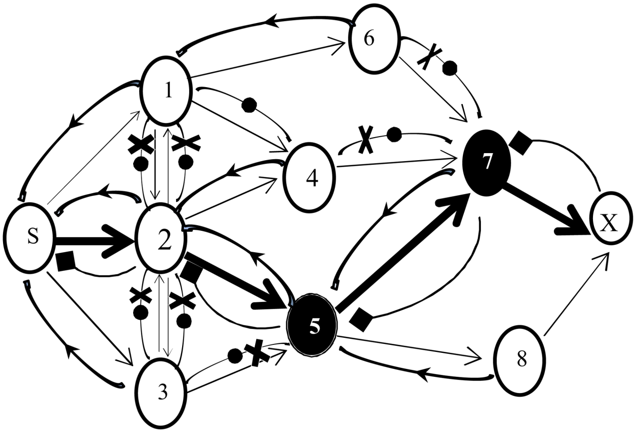


Fig. 1. Route Discovery in NTTM

w	= Wormhole Node
Ⓧ	= Legitimate Node
x ⌒ y	= Node X save broadcast time of RREQ packet by node Y
x ⌒●⌒ y	= Node X reject RREQ packet broadcasted by node Y
x ■⌒ y	= Route Reply sent by node Y to Node X
→	= Broadcasting direction of RREQ packets
➔	= Actual path obtained
✖	= Shows the rejection by any particular node

A node keeps the record of T_{RREQ} of its neighboring node till it gets the information regarding T_{RREP}(Time of receiving RREP at any node) of that particular node. In this way RREQ packet reaches at destination node D through the intermediate nodes (2, 5 and 7). Thus the route established is S---2---5---7---D.

Now destination node responds by generating RREP packet. Node D copies the RRL in RREP packet and attaches an extensional space to record the RRT between each neighbor on the route. The RREP packet is unicast back and received by node 7. This node satisfies the condition mentioned in step 21 and simply put the time of receiving of RREP T_{RREP} at the corresponding space in RREP packet without any computation. When the RREP packet proceeds further according to RRL, it will be received by node 5. Node 5 could not meet step 21, so it extracts$(T_{RREP})_7$ *i.e.* time of receiving RREP at any node7 from extensional part in RREP. The RTT of its neighbor 7 computed by node 5 (through step 22 to 24 of algorithm) and put it back at the place from where $(T_{RREP})_7$ it was extracted. Similarly each intermediate node repeats steps from 20 to 24 till RREP packet reached back at source.

3.2 NTTM Algorithm

1. If a node wants to send a data, it initiates route discovery process.
2. The Source node generates RREQ and put its own IP address into Record Route List (RRL) option in RREQ packet as an originator.
3. Source node broadcasts RREQ packet and store T_{RREQ}.
4. Each node in the transmission range of sender node receives RREQ packet.
5. If a node receives a packet with same source ID, broadcast ID and hop length greater or equal than already received packet, then drop the packet.
6. Otherwise node receiving RREQ matches destination (target) IP address.
7. If target IP address matches with receiver IP address go to step 15.
8. Otherwise each neighbor node starts checking RRL.
9. If IP address at (N-1)th position in RRL matched with receiver IP address. // where N is the number of addresses stored in RRL in RREQ packet at any time T.
10. Then stores T_{RREQ} and drop the packet. //neighbor node hears RREQ broadcasting of other nodes.
11. Otherwise node continuously further searches RRL.
12. If node receiving RREQ packet, IP address matched at other position in RRL, then drop RREQ packet.
13. Otherwise receiver node appends its address at tail of RRL in RREQ packet and broadcast it further.
14. Repeat the step 4 to 13 for each intermediate node till destination.
15. When the RREQ packet arrived at destination.
16. Destination generate RREP packet to respond RREQ packet.
17. The RRL is reversed and copied into RRL of RREP packet.
18. An extensional part is added into basic DSR RREP packet by destination. // To store RTT calculated for each node on the route by its neighbor.
19. Destination unicasts RREP packet along reversed RRL.

20. Next node in RRL receives RREP packet.
21. If receiver node IP address is at 2nd position in reversed RRL (RREP RRL) or second last on RRL of RREQ, then store $(T_{RREP})_x$ at appropriate position in extensional part of RREP and forward back the packet further. // Receiver node is just before the destination node.
22. Otherwise if nodereceiving RREP packet is addressed at kth position on reveres RRL, then receiver node extracts T_{RREP} from extensional part.
23. Then receiver node calculate RTT of its neighbor addressed at (k-1)th position on reverse RRL or at (k+1)th position on RRL of RREQ using T_{RREQ} value stored in step 10 and T_{RREP} extracted by receiver from extensional part using equation 1given below.

$$(RTT)_{x,\,d} = \text{abs}\,((\text{RREQ Time})_x - (\text{RREP Time})_x) \qquad (1)$$

24. The node calculating RTT of corresponding neighbor, stores backRTT valueat position from where T_{RREP} is extracted.
25. Repeat steps 20 - 24 till RREP packet reached at source node.
26. When RREQ packet arrives at source, the source node repeats step 23, 24 for one time and calculate RTT value of its neighbor's node on the route.
27. Now source extracts RTT value of each node from extensional part and calculate RTT between each neighbor node on the route using equation 2

$$(\text{RTTneighbors})_{x\,y\,=}\ \text{abs}\,((RTT)_{x,\,d} - (RRT)_{y,\,d}) \qquad (2)$$

28. The source node compares the value of RTTneighbors calculated in step 27 with threshold value.
29. If RTTneighbors is more than threshold value then source node declares that the presence of wormhole on the route.

3.3 Computation of RTT in NTTM

Each node overhears the RREQ broadcastby its neighbors as shown below in Fig. 2 and keeps record of T_{RREQ} in its cache. While receiving RREP packets every node inserts receiving time T_{RREP} into extensional part of RREP packet for further computation.

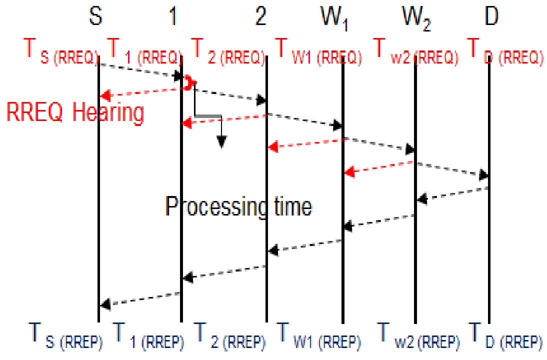


Fig. 2. Timing Diagram of Hearing of RREQ in NTTM

The value of RTT between each node and destination is computed according to the equation 1 mentioned above. The value of RTT between neighboring nodes on the path is computed according to equation 2 given above.

Each node extracts T_{RREP} of its neighbor from extensional part and after calculating its neighbor RTT inserts back RTT in place of T_{RREP}. Here, we assumed that the time at which RREQ sent by a particular node is same as the time of hearing the broadcast of RREQ by its neighbor. The RREQ broadcasting time and RREP receiving time are listed in Table 1.

Table 1. RREQ and RREP Time Record in NTTM

Nodes	Node Hearing RREQ broadcast	RREQ Hearing Time($TH_{(RREQ)}$)	RREP receiving Time ($T_{(RREP)}$)
S	S	-	30
2	S	2.5	27
5	2	5.5	23
7	5	13	15.5
D		-	-

Now the RTT values between each node and the destination is computed using equation 1 and the values are lsited in Table 2.

Table 2. Computation of RRT of Each Node in NTTM

Nodes	Node Computing RTT	RREQ sending Time ($T_{(RREQ)}$)	RREP Receiving Time($T_{(RREP)}$)	RRT of Node
S	S	0	29	29
2	S	2.5	26	23.5
5	2	5.5	23	17.5
7	5	13	16	3

The source node extracts the value of the RTT between each node and the destination node from the extensional part of RREP. It computes the RTT between each neighbors and compares these values with thresold value.The values of RTT between each neighbors are listed in Table 3.

Table 3. RTTneighbors Computation in TTM at Source Node

$(RTT)_{x\,d}$	$(RTT)_{y,\,d}$	$(RTTneighbors)_{x\,y}$
33	28	5 (RTT_{S1})
28	24	4 (RTT_{12})
24	18	6 (RTT_{2W1})
18	4	14 (RTT_{W1W2})

The value of RTT between node 5 and node 7 is comparatively very high as shown in table 3. Therefore, it shows the presence of wormhole attack between node 5 and node 7.

4 Simulations and Experimental Evaluation

To evaluate the performance of NTTM, the simulation is carried out using QualNet 5.0.2 Network simulator. The “All Pass” model is used to launch the wormhole attack. It is assumed that the time of receiving RREQ packet at a node (who calculates RTT) is same as the time of broadcasting RREQ packet by its neighboring node for which RTT is calculated. All nodes are working in promiscuous mode. The diameter of the network is small. The simulation parameters usedare listed in Table 4.

Table 4. Network Simulator Parameters

Parameters	Value
Simulation Time	1000sec
Simulation Repetition	100
Routing protocol	DSR
MAC Layer	802.11
Packet Size	512 bytes
MAC Protocol	802.11
Data Rate	2Mbps
MAC propagation delay	1 μs
Terrain Size	1500 x 1500
Network layer protocol	IPv4
Mobility Model	Random waypoint
Data Traffic Type	CBR
Maximum buffer size forpackets	50 packets
Antenna Model	Omnidirectional
Antenna Height	1.5metres
Noise Factor (SNR)	10.0
Transmission Power	15dBm
Transmission range	367metres

The simulation results were recorded in text file and graphs were generated using Microsoft Office Excel 2007. The trend is observed through the line graph between node’s speed verses packet delivery Ratio (PDR). DSR under the wormhole attack and NTTM under wormhole attack were compared in terms of PDR with different node mobility.

Under the wormhole attack, PDR value decreased consistently as compared to DSR protocol. The NTTM model performed better as shown below in Fig. 3. It showed maximum growth in PDR of 11% at the mobility rate of 35m/s while the worst performance was observed at mobility rate of 25m/s where the growth was only approximately 5%.

At high mobility, the topology changes very rapidly. Therefore, the frequency of route breakage is very high. It is very difficult to build new routes in such conditions. As the speed of nodes increase, PDR falls downs. Initially it fell down very rapidly as shown in Fig. 3. But on further increment in the mobility of nodes, frequency of route breakages get saturated. Therefore, the metric values fell relatively low. From the results, it is evident that the NTTM model performed better and showed a significant growth in PDR as compared to DSR protocol under wormhole attack.

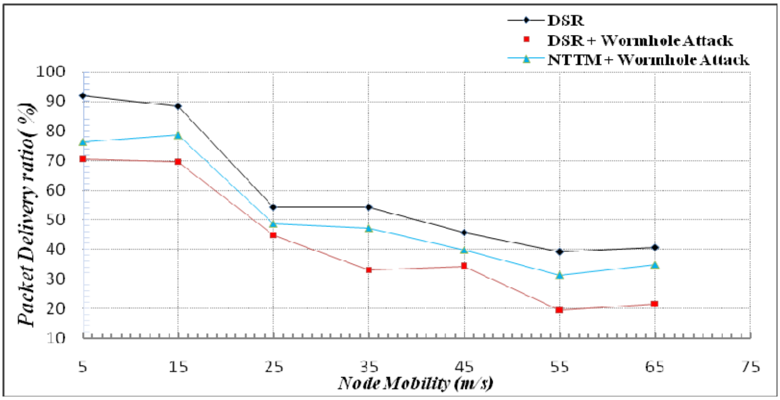


Fig. 3. Packet Deliver Ratio verses Node Mobility

Threshold value played an important role in NTTM model. The threshold value is picked up with respect to the RTT between real neighbors which was observed 14ms and it is then incremented further. If the nodes were at critical position in the network, the nodes experienced large delay could be considered as malicious nodes. Thus, low threshold value result in high false positive. At high threshold value the wormhole attack launched with small tunnel length were undetected and slowly damage the network, therefore false negatives increases.

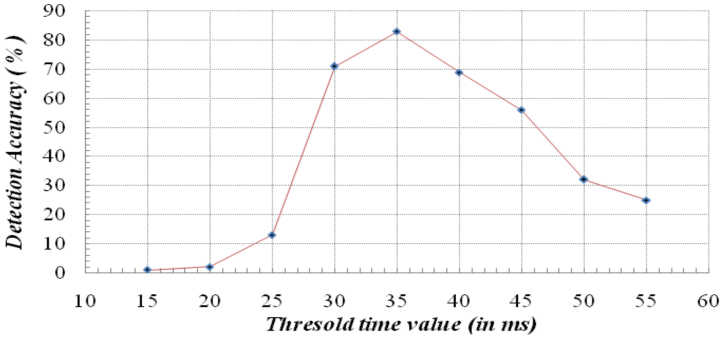


Fig. 4. Detection Accuracy of NTTM

As in the Fig. 4, the detection accuracy graph showed an extreme increment after 25s and achieved best detection accuracy at 35s. At 35ms, both false positive and false negative were low.Therefore,35ms were chosen as threshold limit value.

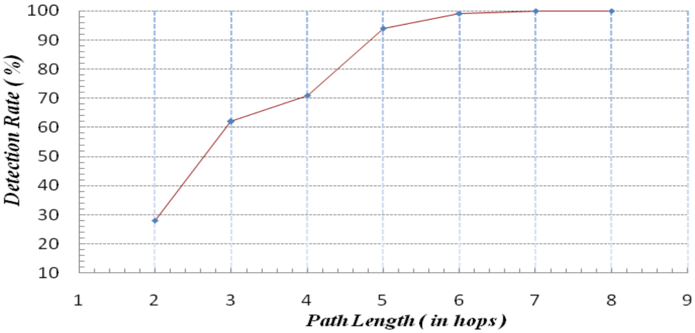


Fig. 5. Detection Rate in NTTM

As shown above in Fig. 5, the detection rate increased exponentially with respect to tunnel length up to 5 hops and saturated after the tunnel length exceeds 6 hops. As the tunnel length increases, the RTT value between the malicious nodes also increases and finally theRTT value between malicious nodes exceeds threshold. Therefore, wormhole attacks can be detected easily and accurately.

5 Conclusion and Future Scope

The study regarding the wormhole attack leads us to draw the conclusion that Wormhole attack is most dangerous attack in MANETs. By identifying the wormhole attack during route discovery, NTTM avoids the chances of damages in the network due to attacks. It was observed that the PDR value on average falls by 17% when DSR protocol under wormhole attack is compared with DSR protocol without wormhole attack at different node mobility. However, the results are improved by 9% under NTTM model. The accuracy and detection rate of NTTM model improved with the incrementin tunnel length.

In future, the difference between the sending time of RREQ packet and receiving time of RREQ packet at its neighbor can be taken into consideration which was assumed negligible in NTTM. NTTM can be implemented over other routing protocols like TORA, DSDV etc.

References

1. Su, M.Y.: WARP: A wormhole-avoidance routing protocol by anomaly detection in mobile Ad-hoc networks. Int. J. of Computer and Security 29, 208–224 (2010)
2. Taheri, M., Naderi, M., Barekatain, M.B.: New Approach for detection and defending the Wormhole Attacks in Wireless Ad Hoc Networks. In: Proc. 18th ICEE, Iran, pp. 331–335 (2010)
3. Van Tran, P., Canh, N.T., Lee, Y.-K., Lee, S.: Transmission Time-based Mechanism to Detect Wormhole Attacks. In: Proc. of IEEE 2nd Asia-Pacific Services Computing Conference, Seoul, pp. 172–178 (2007)

4. Chiu, H.S., Lui, K.S.: DELPHI: wormhole detection mechanism for ad hoc wireless networks. In: Proc. of IEEE 1st Symposium on Wireless Pervasive Computing, China, pp. 6–11 (2006)
5. Perkins, C.E., Royer, E.M., Das, S.R.: Ad hoc on-demand distance vector (AODV) routing. IETF Internet draft, MANET Working Group (2004)
6. Gupta, S., Kar, S., Dharmaraja, S.: WHOP: Wormhole attack Detection Protocol using Hound Packet. In: Proc. IEEE on Innovations in Information Technology, United Arab Emirates, pp. 226–231 (2011)
7. Hu, Y.C., Perrig, A., Johnson, D.B.: PACKET LEASHES: A defense against wormhole attacks in wireless ad hoc networks. In: Proc. IEEE INFOCOM 2003, USA, vol. 3, pp. 1976–1986 (2003)
8. Wang, X., Wong, J.: An end-to-end detection of wormhole attack in wireless Ad-hoc networks. In: Proc. 31st Computer Software and Applications, vol. 1, pp. 39–48. IEEE Computer Society, Washington, DC (2007)

A Privacy Preserving Representation for Web Service Communicators' in the Cloud

D. Chandramohan[1], T. Vengattaraman[1], D. Rajaguru[2], R. Baskaran[3], and P. Dhavachelvan[1]

[1] Department of Computer Science, Pondicerry University, India
[2] Dept.of Inf.Technology, Perunthalaivar Kamarajar Inst.of Eng. & Technology, Karaikal
[3] Department of Computer Science and Engineering, Anna University, Chennai, India
{pdchandramohan,vengat.mailbox,raja.guru42,
dhavachelvan}@gmail.com, baaski@cs.annauniv.edu

Abstract. The present paper focuses on maintaining one's data secrecy in cloud storage area and it may depend on their privacy policy and standards. The effectiveness and efficiency of preserved data from different cloud providers should maintain the integrity of original data stored in it. In the era of could computing, stored information's value, proficiency and optimization of data retrieving spotlights the importance of maintaining cloud users data, privacy, identity, reliability and maintainability it may vary for different Cloud providers (CP). Giant CP ensures their user proprietary information's are sustained more secretly using cloud technologies. During third party cloud services and exodus between inter cloud providers may lead to data portability privacy issue. More remarkable event in this case, even the cloud providers don't have implication about the information and records where it's stored and maintained in their own cloud. This is one of the obligatory research issue in cloud computing. We came frontward by proposing (EMPPC) an Evolutionary model based privacy Preserving technique and try to hold user's, to have trust in providers for maintaining their confidentiality in cloud. This proposal helps the CR (Cloud Requester and Users) to mark trust on their proprietary information and data's stored in cloud.

Keywords: Cloud computing, Web Service, Privacy, Security, Intelligent Computing, Data Portability, Intrusion Detection System, lattice.

1 Introduction

Cloud is one of the massive and major research areas for both industrial and academic field for research, many researchers have been working towards its research issues. As cloud came into existence lot of issues also surrounded to it. Normally cloud computing have most common and general issues like Interoperability, SLA-(Service level Agreement), [3-10] universal standards, unique approach for all cloud providers, data portability among different clouds, various privacy and security issues etc. Cloud computing is not a one-size-fits-all elucidation and companionship required to

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 496–506, 2013.

uncover the need earliest before business into such solutions. CP consists of different layer for information dealing out with an on-demand provisioning of computational resources.

Data Portability Policy and Data Freedom

A cloud provider came across the issues of data portability in the way that users have a request of it. We are initiating a model which may help to frame an open standard because we believe in advancing this open effort. [1-6], However the CP should not permitted to change their policy on demand of their own. Very few cloud provider companies has already launched portability policies. Fig.1 the portability policy proposal is still in its preschooler stages and will nurture as awareness increases, with more unambiguous questions emerging when issues are recognized. CP and SN (social network) providers will need to pay finicky consideration to the projected right for users to port their personal information to another CP, as well as their right to erase their information. Fig.1. [9-21], The right for CR to port their data to a new CP will also be of an explicit anxiety to SN whose servers continue to edge over with user information. The right for CR to involve along with CP to relocate their data to a new CP and it should promote cloud shopping. This will encourage larger antagonism between cloud providers. One of the most valuable weapons for CRs have in their hand is to switch different providers. This is an idyllic policy that should be pursued by all CP. The initiative by means of data Liberation campaign is to be welcomed, even if it residue vague whether the actual motive is to reduce the cloud competitors in maintaining their data container, rather than making customers free from CP data. There are some considerable difficulties to truly liberal cloud users due to data migration policy. [19-26], The policy framing bench is responsible to act in response to its customers if any information exploitation happens.

Practical Challenges

a) Any policy right must illustrate a fetish between a CR data and their fundamental rights to use of it. [5-9]
b) The policy right should be limited to liberate the data held by the cloud provider. [3-7]
c) Whatever the data transferred between different CPs will relay on the configuration of the data format and interfaces they are using, some CPs store data in the favored format for data exchange internally. [11-18]
d) Different CP data format is harder to understand and it may not be portable while transferring data to a new CP. [20-26]
e) The exact raw data is easier to transfer but difficult to maintain its secrecy.

Promoting a Privacy Market and Industry Standards

Cloud requestors very frequently may change their provider based on the providers advance privacy preservation techniques adopted and assurance to CR, [22] there may be some Pit fall if some exploration happens to user's data after adapted to new privacy policy of CPs. Many researchers have been repeatedly initiating to frame an [17-19] universal standard format which helps a provider to permit other business competitive provider to gasp different CPs technology, and it should be compatible to new CP technology which makes an effortless transform, can be achievable to normalize the user personal information and data. These migration possibilities should be informed to all CR and users as different cost for porting their data.

2 Background and Related Work

In his approach Anna et al [1] the author addresses the quandary of safeguarding privacy in trust consultation. He initiates the notion of privacy preserving discovery, with a set that does not include attributes or credentials, or combinations of these, which may negotiate privacy. [3] To obtain privacy preserving disclosure sets, he proposed two techniques based on the notions of substitution and generalization. Keith Frikken et al [2006] in his work he presented few protocols that protect both sensitive credentials and sensitive policies as privacy preserving mechanism. Travis D et al [2008] his research team propose a method to support the software engineering effort to derive security requirements from regulations; in which the methodology for directly extracting access rights and obligations from regulation texts. [8-9] The methodology provides statement-level coverage for an entire regulatory document to consistently identify and infer six types of data access constraints and assign required priorities between access rights and obligations to avoid unlawful information disclosures. Alex X et al and team in [2011] they propose a VGuard framework with efficient protocol that allows a cloud policy owner and a cloud request owner to collaboratively determine whether the request satisfies the policy without the policy owner knowing the request and the request owner knowing the policy. Pengcheng Xiong et al and team [6] proposes a cost-aware resource management system based on SLA- service level agreement termed as SmartSLA which consists of two main components: the system modelling module and the resource allocation decision module. To prevent the online social community the author Dongsheng Li et al [7] came forward to propose an interest group based privacy-preserving recommender system called Pistis. By identifying inherent item-user's interest groups and separating user's private interests from their public interests. Multi-agent based service accessing and security system flow discussed by authors et al [2], [4], [5], [12], [15-16] The author Xiaohui Liang et al [12] proposed the privacy preserving emergency call scheme by enabling patients in life-threatening emergencies to fast and accurately transmit emergency data to the nearby helpers via mobile healthcare social networks. The author Chien-Ding Lee et al [2011] similarly the authors discussed the necessity of secure web services and mitigation of the same in cloud [10-14], [17-22] in his approach he made few regulation to comply with the HIPAA (Health Insurance Portability and Accountability Act), a flexible cryptographic key management solution is proposed to facilitate interoperations among the applied cryptographic mechanisms. Author et al proposed an efficient service cache in an peer to peer networking which simulates the idea of deliver the service in most stipulated period of time [23], Sachin Kadloor et al [16] proposal to develop a dynamic program to compute the optimal privacy preserving policy that minimizes the correlation between user's traffic and adversary's waiting times of cloud user. Chun-Tao Hong et al [18] propose a new MapCG model as a Map-Reduce framework to provide source code level portability between CPUs (Central processing units) and GPUs (Graphics processing units). D.Chandramohan et al [15] proposed a testbed for evaluating the efficiency of services and these features are inherited in our proposed model for driving new parameters and functions to maintain the privacy and its security in cloud environment during data portability or migration between different cloud Providers'.

3 Proposed Model

This approach focus on portability issue in cloud, a user can hold their account details and information along with respective trusted cloud providers, it get pursued until the user mark their uncomfortable with particular CP. Even cloud provider suggestion may be unsuccessful during back tracking the client information maintained by them. They do not have a clear identification where the actual data resides inside their CP cloud. By using this proposal our main objective is to resolve this issue at a minimum risk and maximum benefit to both the providers and users.

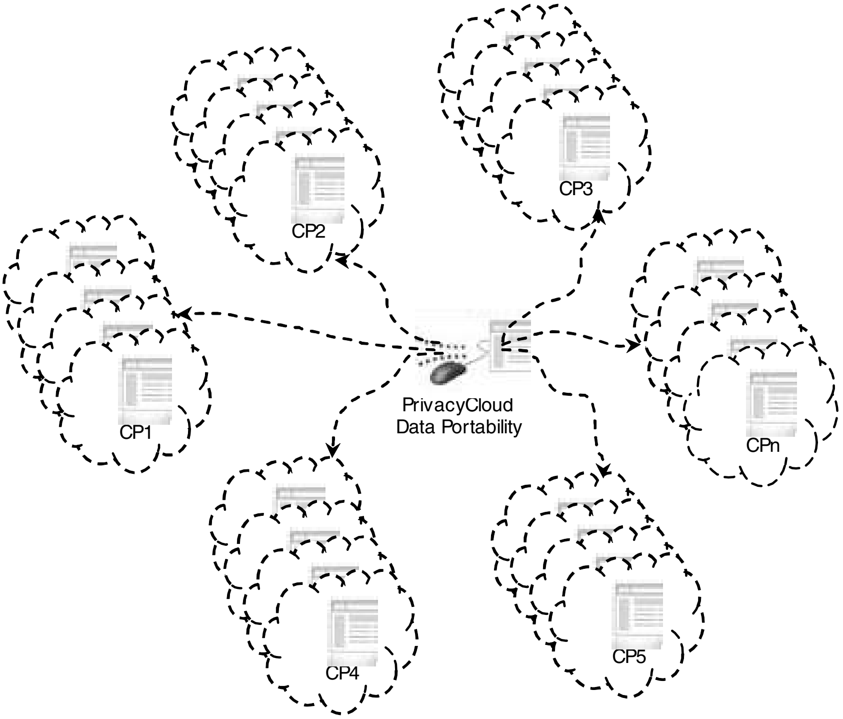


Fig. 1. Privacy Breach in Inter Cloud Portability and Data Migration

The proposed impend strenuous on resolving issues such as adaptability, scalability, reliability, privacy and security, to access the client information as ubiquitous local services virtually in any system. If a user what's to withdraw all his data from one cloud and wish to transfer his data to another cloud, here comes the data portability and its privacy issues. With an open standard and privacy policies cloud computing can able to achieve portability with a huge percentage as freely and loosely coupled with all cloud providers with standard application via internet with in user authorized control. Security is one of the most important frames for cloud provider it will utilize data storage and transmission encryption, user authentication, and authorization, all cloud user concern about the liability of isolated data accessed by criminals like hackers, intruders, and annoyed employees. Cloud providers are extremely aware to this problem and applied extensive possessions to extenuating this kind of distress. Reliability also one of the main issues to feel uncomforted with cloud providers both financially and technologically trustworthy in current market. By using superfluous storage technique some CPs modifying the original data stored with them and lead to signing off from one provider to another. Ownership to CR data has been transferred to the cloud; some users concern that they could lose several data or CP thinks all of their rights are incapable to protect the rights of their beloved customers. Many CP are concentrating on this issue with full standard policy for user and providers benefited agreements.

$$\text{Sim}\ (C_1, C_2) = e^{-ch}.\frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} \tag{1}$$

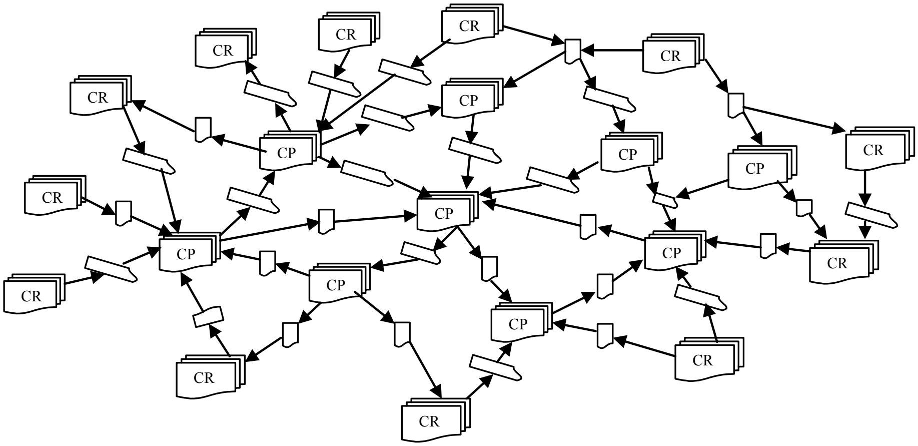


Fig. 2. Privacy EMPPC Intelligent model for Preserving Cloud Users Data

Like this agreement users would be prudent to seek recommendation from their beloved authorized delegates. Data backup CP surplus servers and regulates data backing processes, but some CP worry about being able to manage their own support. Many cloud providers are now presenting data chuck onto medium or allowing cloud users to back up data through ordinary downloads. Data portability and conversion some CR and users are worried that they wish to switch cloud providers; they may have complicated in relocating their data. Porting and exchanging data is highly reliant on the environment of the CPs data reclamation arrangement, fussy in personal belongings where the configuration cannot be effortlessly revealed. As service antagonism nurtures until some open standards befall established, the data portability issue will be more ease, and adaptation progression will be offered by sustaining more accepted cloud providers with some conditions like a cloud users or CR should pay for some ritual data exchange. Supporting multiplatform and more are big issue for IT sector using direct services, the cloud-based service assimilates athwart different environment and operating systems, and some personalized amendment of the service acquire of any new problem into it. Multiplatform sustainment and its necessities will show the simplicity as more user edges are converted into normal web-based system supports. Intellectual Property originates few new features and to use CP as part of the discovery. Once some CP recognizes that computing makes potentially experiences much more of the same fortune as owned new faceplates the proprietary systems, for low-risk processes and for insensible information, the cloud-based services can be endorsed, established, tartan, and made more protected by merging them with habitual non-cloud IT practices.

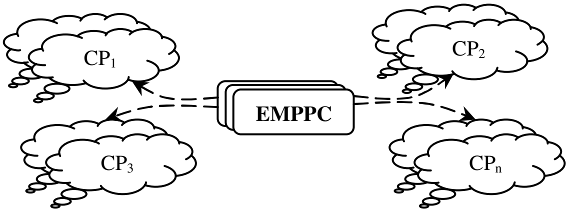


Fig. 3. An Evolutionary Model Based Communicators' Privacy Preserving in the Cloud

In this innovative and competitive cloud world a mammoth development in all related areas. There presents 200 percentage chance of switch over from one provider to some other providers. If a user what's to withdraw all his data from one cloud and wish to transfer his data to another cloud, here comes the data portability and its privacy issues. Designing an Evolutionary model using Intrusion detection system protocol is developed by sharing information based intrusion detection system and the proposed system is embedded in all cloud layer and its neighborhood nodes to provide privacy and security to those data.

$$Cd_A\ (x_i,\ x_j) = (x_i\text{-}\ x_j)^I\ A\ (x_i\text{-}\ x_j) \tag{2}$$

$$CS_A\ (x_i,\ x_j) = x_i^T\ A\ xj \tag{3}$$

3.1 Some Properties of ID-Intrusion Detection System Used in EMPPC

Alert / Alarm, True Positive, False Positive, False Negative, True Negative, Noise, Site policy, Site Policy awareness, Confidence value, Alarm filtering, Attacker identification, Masquerader (Duplicate), Misfeasor, Clandestine user (User act as an Administrator)

$$CD(x,y)=(max\ \{log\ f(x), log\ f(y)\}\ log\ f(x,y))/$$
$$log〖M\ min〖\{\ 〖log\ 〗〖f\ (x), log〖f\ (y)〗\ 〗\}〗\ 〗 \tag{4}$$

$$\frac{\sum_{(n.m)€A} Sim\ l(n.m)}{|P_1| + |P_2|} \tag{5}$$

$$CC_{IN}\left(p_i, t_j\right) = K, \forall\ k > 1 \tag{6}$$

CP- Cloud Provider, CR- Cloud Requestors Fig.3. EMPPC -The proposed Evolutionary model illustrates many preprocessed approaches to check all promising

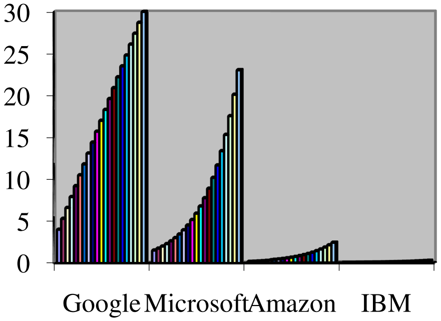


Fig. 4. Cloud Service Privacy Assessment of different providers and Interoperability evaluation in normal scenario

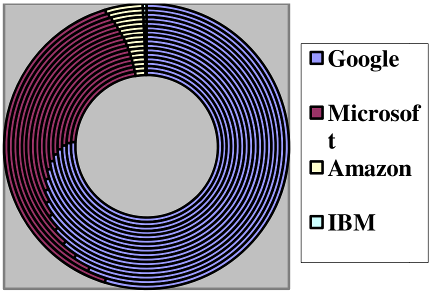


Fig. 6. Cloud Service Privacy with different provider's Evaluation in Medium scenario

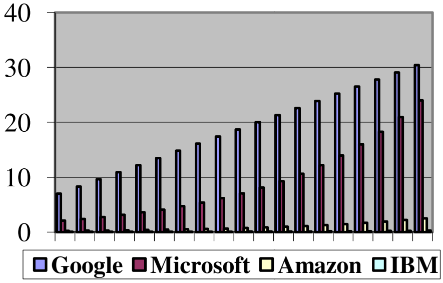


Fig. 5. Cloud Service Privacy Breach Evaluation in Typical Mode

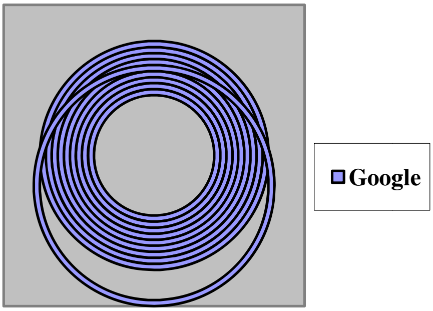


Fig. 7. Cloud Service Privacy Evaluation in Custom Mode with Proposed System Approach from providers perception

Table 1. Cloud Service Privacy assessment of different providers evaluation in normal scenario

Google	Microsoft	Amazon	IBM
4.0	1.5194	0.1589	0.0166
5.3	1.7406	0.1820	0.0190
6.6	1.9941	0.2085	0.0218
7.9	2.2845	0.2389	0.0250
9.2	2.6171	0.2737	0.0286
10.5	2.9982	0.3135	0.0328
11.8	3.4348	0.3592	0.0376
13.1	3.9350	0.4115	0.0430
14.4	4.5080	0.4714	0.0493
15.7	5.1644	0.5401	0.0565
17.0	5.9165	0.6187	0.0647
18.3	6.7780	0.7088	0.0741
19.6	7.7651	0.8120	0.0849
20.9	8.8958	0.9303	0.0973
22.2	10.1912	1.0657	0.1114
23.5	11.6752	1.2209	0.1277
24.8	13.3754	1.3987	0.1463
26.1	15.3231	1.6024	0.1676
27.4	17.5544	1.8357	0.1920
28.7	20.1106	2.1030	0.2199
30.0	23.0391	2.4093	0.2519

Table 2. Cloud Service Privacy evaluation in Typical Mode

Google	Microsoft	Amazon	IBM
10.0	2.8455	0.2976	0.0311
11.3	3.2598	0.3409	0.0356
12.6	3.7345	0.3905	0.0408
13.9	4.2783	0.4474	0.0468
15.2	4.9014	0.5126	0.0536
16.5	5.6151	0.5872	0.0614
17.8	6.4327	0.6727	0.0703
19.1	7.3695	0.7706	0.0806
20.4	8.4426	0.8829	0.0923
21.7	9.6720	1.0114	0.1058
23.0	11.0805	1.1587	0.1212
24.3	12.6940	1.3274	0.1388
25.6	14.5425	1.5208	0.1590
26.9	16.6601	1.7422	0.1822
28.2	19.0861	1.9959	0.2087
29.5	21.8655	2.2865	0.2391

Table 3. Cloud Service Privacy breach with different providers evaluation in Medium scenario

Google	Microsoft	Amazon	IBM
7.0	2.0793	0.2174	0.0227
8.3	2.3820	0.2491	0.0260
9.6	2.7289	0.2854	0.0298
10.9	3.1263	0.3269	0.0342
12.2	3.5815	0.3745	0.0392
13.5	4.1031	0.4291	0.0449
14.8	4.7006	0.4916	0.0514
16.1	5.3851	0.5631	0.0589
17.4	6.1692	0.6451	0.0675
18.7	7.0676	0.7391	0.0773
20.0	8.0968	0.8467	0.0885
21.3	9.2758	0.9700	0.1014
22.6	10.6265	1.1112	0.1162
23.9	12.1739	1.2731	0.1331
25.2	13.9467	1.4585	0.1525
26.5	15.9776	1.6708	0.1747
27.8	18.3043	1.9141	0.2002
29.1	20.9697	2.1929	0.2293
30.4	24.0233	2.5122	0.2627

Table 4. Cloud Service Privacy evaluation in Custom Mode with different providers

Over all Web Service Suitability x; f(x); f'(x); f"(x)		
No of services	**Suitability**	**Providers**
100	10	Google
100	25.2	Microsoft
100	40.4	Amazon
100	55.6	IBM
100	70.8	Sales force
100	86	VMware
100	101.2	Verizon
100	116.4	Accenture
100	131.6	Sodexo
100	146.8	Infosys Technologies

customs to locate the portability between cloud providers. It will act as a gateway for all providers and reveille the privacy actions throughout some malicious attacks experience for the period of portability surrounded by different cloud providers. This paper discusses about portability issues and privacy technique to solve those portability problem occurring for users and cloud providers. Above properties illustrates the technique adopted to maintain the secrecy of any proprietary information. Fig.2 and fig. 3 will give an apparent idea to researches about the proposed work. Equation (1-6) describes the privacy implications in the proposed system. Data migration and its possibilities are expressed if a user currently using CP_1 he can able to migrate to CP_n possibilities as per the compatibility of different cloud providers, the proposed system handles the situation more appropriately and it notifies to all CPs to maintain a standard format to avoid the compatibility issues.

Different privacy invasion have been tremendously increasing in day today life. As privacy breach protection and mitigation stagey the proposed approach acts accordingly. The service provider utilized various privacy techniques to drop down these issues and those representations are tabulated in Table 1, 2, 3 and 4. These invasion handling are plotted and expressed in fig.4, fig.5, fig.6 and fig.7. In vast storage data centers like grid, cloud and distributed storage area are identified as an enormous data breach happening gradually. Table 1 explains the different cloud providers privacy policy, have get varied as per their norms, most of the factors are enriched to be hidden and almost all leading providers are committed to persuade their customers with their attractive policy.

4 Conclusion

This paper proposes an evolutionary privacy model for data portability privacy in cloud which persuades the subsistence of cloud users to have enough trust on cloud providers and co-cloud users. It encompasses both CP and Cloud User (CU) with self and mechanized systems. Most cloud providers don’t put forward and missed to mark the practice of privacy techniques. In our paper we came up with various privacy protection techniques and explored them as survival of the fittest in cloud environment. Similarly table 2, 3 and 4 explore the privacy breach of providers taken into consideration according to its recognition. Proposed approach get fulfilled only when both cloud providers and cloud requestors / end-users ensures all their data have their own privacy in different cloud provider during data portability. The future research focuses on portability in interoperable privacy issues, researchers can hope this proposal will prove to be a useful foundation for solving their issues on privacy for cloud layer in all stipulated areas.

Acknowledgement. This work is a part of the Research Project sponsored under the Major Project Scheme, UGC, India, Reference No: F. No. 40-258/2011 (SR), dated 29 June 2011. The authors would like to express their thanks for the financial support offered by the Sponsored Agency.

References

1. Squicciarini, A.C., Bertino, E., Ferrari, E., Ray, I.: Achieving Privacy in Trust Negotiations with an Ontology-Based Approach. IEEE Transactions on Dependable and Secure Computing 3(1), 13–30 (2006)
2. Victer Paul, P., Saravanan, N., Jayakumar, S.K.V., Dhavachelvan, P., Baskaran, R.: QoS enhancements for global replication management in peer to peer networks. Future Generation Computer Systems 28(3), 573–582 (2012)
3. Vengattaraman, T., Abiramy, S., Dhavachelvan, P., Baskaran, R.: An Application Perspective Evaluation of Multi-Agent System in Versatile Environments. International Journal on Expert Systems with Applications 38(3), 1405–1416
4. Abirami, S., Baskaran, R., Dhavachelvan, P.: A survey of Keyword spotting techniques for Printed Document Images. Artificial Intelligence Review 35(2), 119–136 (2011)
5. Victer Paul, P., Vengattaraman, T., Dhavachelvan, P.: Improving efficiency of Peer Network Applications by formulating Distributed Spanning Tree. In: Proceedings - ICETET 2010, pp. 813–818 (2010)
6. Xiong, P., Chi, Y., Zhu, S., Moon, H.J., Pu, C., Hacıgumus, H.: Intelligent Management of Virtualized Resources for Database Systems in Cloud Environment. In: IEEE ICDE Conference 2011, pp. 87–98 (2011)
7. Li, D., Lv, Q., Xia, H., Shang, L., Lu, T., Gu, N.: Pistis: A Privacy-Preserving Content Recommender System for Online Social Communities. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 79–86 (2011)
8. Venkatesan, S., Dhavachelvan, P., Chellapan, C.: Performance analysis of mobile agent failure recovery in e-service applications. International Journal of Computer Standards and Interfaces 2(1-2), 38–43 ISSN:0920-5489
9. Chandramohan, D., Veeraiah, D., Shanmugam, M., Balaji, N., Sambasivam, G., Khapre, S.: SVIP-enhanced security mechanism for SIP based voIP systems and its issues. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds.) Advances in Computing & Inform. Technology. AISC, vol. 176, pp. 81–86. Springer, Heidelberg (2012)
10. Vengattaraman, T., Dhavachelvan, P.: An Agent-Based Personalized E-Learning Environment: Effort Prediction Perspective. In: IEEE-IAMA 2009 (2009)
11. Dhavachelvan, P., Uma, G.V., Venkatachalapathy, V.S.K.: A New Approach in Development of Distributed Framework for Automated Software Testing Using Agents. International Journal on Knowledge Based Systems 19(4), 235–247 (2006)
12. Liang, X., Lu, R., Chen, L., Lin, X. (Sherman) Shen, X.: PEC: A Privacy-Preserving Emergency Call Scheme for Mobile Healthcare Social Networks. IEEE Journal of Communications and Networks 13(2), 102–112 (2011)
13. Dhavachelvan, P., Uma, G.V.: Reliability Enhancement in Software Testing – An Agent-Based Approach for Complex Systems. In: Das, G., Gulati, V.P. (eds.) CIT 2004. LNCS, vol. 3356, pp. 282–291. Springer, Heidelberg (2004)
14. Dhavachelvan, P., Uma, G.V.: Multi-agent Based Integrated Framework for Intra-class Testing of Object-Oriented Software. In: Yazıcı, A., Şener, C. (eds.) ISCIS 2003. LNCS, vol. 2869, pp. 992–999. Springer, Heidelberg (2003)
15. Chandramohan, D., Jayakumar, S.K.V., Khapre, S., Nanda Kishore, M.S.: DWSE-Simulator For Distributed Web Service Environment. IEEE ICRTIT 2011, 1203–1208 (2011)

16. Kadloor, S., Gong, X., Kiyavash, N., Venkitasubramaniam, P.: Designing Router Scheduling Policies: A Privacy Perspective. IEEE Transactions on Signal Processing 60(4), 2001–2012 (2012)
17. Carlson, M.: Systems and Virtualization Management: Standards and the Cloud (A report on SVM 2011). Journal of Network and Systems Management (2012)
18. Hong, C.-T., Chen, D.-H., Chen, Y.-B., Chen, W.-G., Zheng, W.-M.: Providing Source Code Level Portability Between CPU and GPU with MapCG. Journal of Computer Science and Technology 27(1), 42–56 (2012)
19. Chandramohan, D., Vengattaraman, T., Basha, M.S.S., Dhavachelvan, P.: MSRCC – mitigation of security risks in cloud computing. In: Meghanathan, N., Nagamalai, D., Chaki, N. (eds.) Advances in Computing & Inform. Technology. AISC, vol. 176, pp. 525–532. Springer, Heidelberg (2012)
20. Nanda Kishore, M.S., Jayakumar, S.K.V., Satya Reddy, G., Dhavachelvan, P., Chandramohan, D., Soumya Reddy, N.P.: Web Service Suitability Assessment for Cloud Computing. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) NeCoM/WeST/WiMoN 2011. CCIS, vol. 197, pp. 622–632. Springer, Heidelberg (2011)
21. Saleem Basha, M.S., Dhavachelvan, P.: Web Service Based Secure E-Learning Management System- EWeMS. International Journal of Convergence Information Technology 5(7), 57–69 ISSN: 1975 9320
22. Dhavachelvan, P., Uma, G.V.: Complexity Measures For Software Systems: Towards Multi-Agent Based Software Testing Proceedings. In: ICISIP 2005, pp. 359–364 (2005)
23. Victer Paul, P., Saravanan, N., Baskaran, R., Dhavachelvan, P.: Efficient service cache management in mobile P2P networks. Future Generation Computer Systems (2012) ISSN 0167-739X, doi:10.1016/j.future.2012.12.001

Detailed Dominant Approach Cloud Computing Integration with WSN

Niranjan Lal[1], Shamimul Qamar[2], and Mayank Singh[3]

[1] MODY Institute of Technology and Science, Laxmangarh, Sikar (Raj.) –India
niranjan_verma51@yahoo.com
[2] Noida Institute of Engineering and Technology, Greater Noida (UP) - India
drsqamar@rediffmail.com
[3] THDC Institute of Hydropower Engineering & Technology Tehri (UK) -India
mayanksingh2005@gmail.com

Abstract. The maximum benefit out of the recent developments in sensor networking can be achieved via the integration of sensors with Internet. The real-time specific sensor data must be processed and the action must be taken instantaneously. This distributed architecture has numerous similarities with the wireless sensor networks (WSN) where lots of motes, which are responsible for sensing and preprocessing, are connected with wireless connection in the real-time. Since wireless sensor networks are limited in their processing power, battery life, communication speed and storage resources , cloud computing offers the opposite , which makes it fetching for endless observations, analysis and use in different sort of environment.

In this paper we proposed an architecture, which integrates the Cloud computing technology with the wireless sensor network. In this paper we also discussed some research challenges with respect to cloud computing and wireless sensor networks, and important key component of sensor cloud

Keywords: Cloud computing, Distributing computing, Wireless sensor networks, Sensor cloud, Research challenges of cloud computing and Internet.

1 Introduction

Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications. It allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access, with more efficient computing by centralizing storage, memory, processing and bandwidth. It can be securing immense amounts of data which is only accessible by authorized users. Cloud computing in broad way is shows in Figure 1 and Figure 2.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 507–516, 2013.

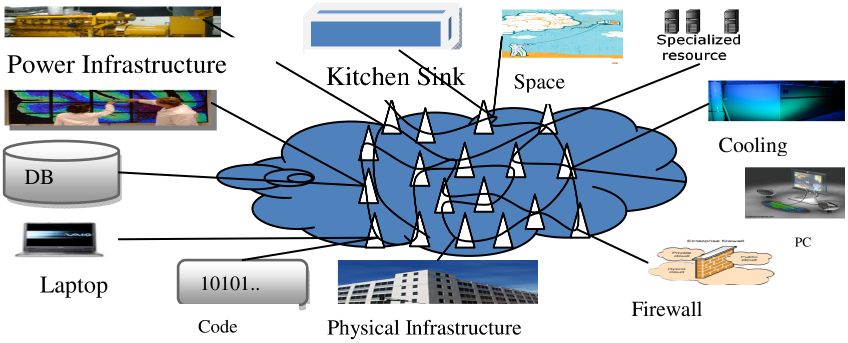


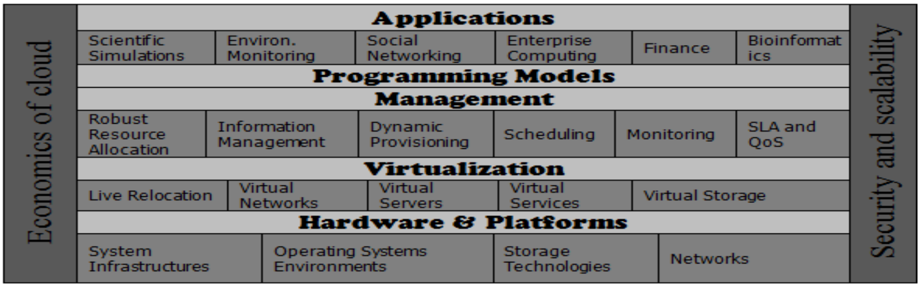
Fig. 1. Cloud Computing

Cloud computing is the technology that enables functionality of an IT infrastructure, IT platform or an IT product to be exposed as a set of services in a seamlessly scalable model so that the consumers of these services can use what they really want and pay for only those services that they use (Pay per use) [2].

"Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [16]. "

Wireless sensor network consists of a large number of such sensor nodes that are able to collect and disseminate data in areas where ordinary networks are unsuitable for environmental and/or strategic reasons. Each sensor node comprises sensing, processing, transmission, mobilize, position finding system (Such as GPS) and power units [8].

The system architecture of wireless sensor network is shown in Figure 3. In other ways Wireless sensor a network is seamlessly couples the physical environment with the digital world. Sensor nodes are small, low power, low cost, and provide multiple functionalities sensing capability, processing power, memory, communication, bandwidth, battery power. Useful in many application domains.



Economics of cloud

Applications

Scientific Simulations | Environ. Monitoring | Social Networking | Enterprise Computing | Finance | Bioinformatics

Programming Models

Management

Robust Resource Allocation | Information Management | Dynamic Provisioning | Scheduling | Monitoring | SLA and QoS

Virtualization

Live Relocation | Virtual Networks | Virtual Servers | Virtual Services | Virtual Storage

Hardware & Platforms

System Infrastructures | Operating Systems Environments | Storage Technologies | Networks

Security and scalability

Fig. 2. Architectural Level Cloud Computing

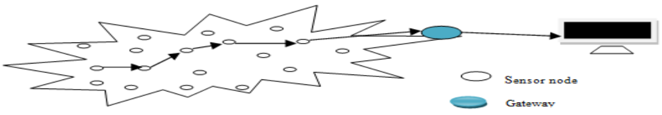


Fig. 3. Architecture of Wireless Sensor Network

The organization of our paper is as follows. In section 2, section 3 and section 4 we have discussed key features of our interest and Limitations of cloud computing and sensor networks. In section 5 describe need to integrate cloud computing with wireless sensor networks? In section 6 we describe research challenges, where some research works can be done in these areas. In section 7 we present a proposed architecture of cloud computing with WSN. In section 8 point out some key components of the proposed architecture and Section 9 conclude and future work.

2 Key Features of Our Interest

There are some key features that are useful for everyone to use internet. i) Immense computational and storage resources that are collocate. ii) Very high speed data processing and movement. iii) Accessibility over the Internet Service-Oriented Architecture and virtually from any platform.

3 Limitation of Cloud Computing

Cloud computing is limited – as of now: i) The immense power of the Cloud can only be fully exploited if it is seamlessly integrated into our physical lives. ii) It is providing the real world's information to the Cloud in real time and getting the Cloud to act and serve us instantly so it need to add the sensing capability to the Cloud.

4 Limitation of Sensor Networks

Sensor networks are limited too: i) It is very challenging to scale sensor networks to large sizes with proprietary vendor-specific designs, which is difficult for different Sensor networks to be interconnected. ii) We know sensor networks is operate in separate silos, so sensor data cannot be easily shared by different groups of users. iii) Sensor network is used for fixed and specific applications that cannot be easily changed.

5 The Missing Piece

The missing piece of cloud computing is shown in the Figure 4. In which cloud computing is would be integrate with sensor networks.

5.1 A Scenario

A scenario is a description of a flow of messages in the network via cell phone shown in the Figure 5. An Insight into the Scenario has some steps shown on next page.

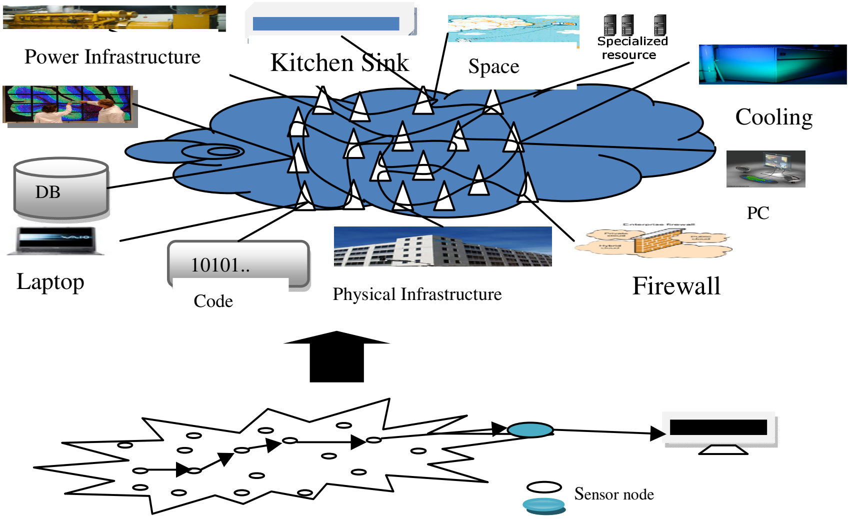


Fig. 4. The Missing Piece

Step 1. Cell phone records the tourist's gestures and activates applications such as camera, microphone, etc.
Step 2. The cell phone produces very swift responses in real time after: i) Processing geographical data. ii) Acquiring tourist's physiological data from wearable physiological sensors (blood sugar, precipitation, etc) and cross-comparing it with his medical records. iii) Speech recognition. iv) Image processing of restaurant's logos and accessing their internet-based profiles. v) Accessing tourist's social network profiles to find out his friends.
Step 3. Fact: the cell phone cannot perform so many tasks!

5.2 Need to Integrate Cloud with Sensors

These are the some point why need to integrate cloud with sensors : i) Acquisition of data feeds from numerous body area (blood sugar, heat, perspiration, etc) and wide area (water quality, weather monitoring, etc) sensor networks in real time. ii) Real-time processing of heterogeneous data sources in order to make critical decisions. iii) Automatic formation of workflows and invocation of services on the cloud one after another to carry out complex tasks.

5.3 The Sensors Cloud

“An infrastructure that allows truly pervasive computation using sensor as interface between physical and cyber world, the data compute cluster as the cyber backbone and the internet as the communication medium.”

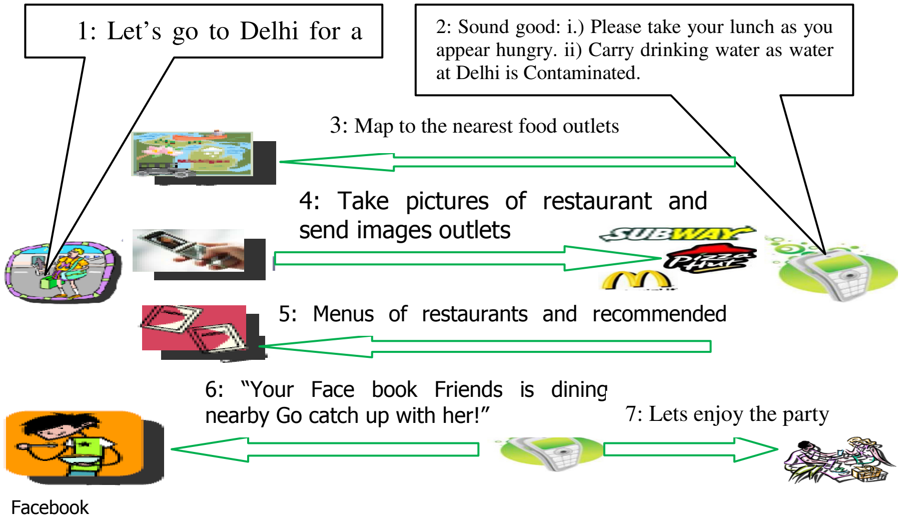

1: Let’s go to Delhi for a
2: Sound good: i.) Please take your lunch as you appear hungry. ii) Carry drinking water as water at Delhi is Contaminated.
3: Map to the nearest food outlets
4: Take pictures of restaurant and send images outlets
SUBWAY
Pizza Hut
5: Menus of restaurants and recommended
6: “Your Face book Friends is dining nearby Go catch up with her!”
7: Lets enjoy the party
Facebook

Fig. 5. A scenario

Sensor cloud integrates large-scale sensor networks with sensing applications and cloud computing Infrastructures , which collects and processes data from various sensor networks, that will enables large-scale data sharing and collaborations among users and applications on the cloud, then it will delivers cloud services via sensor-rich mobile devices, which allows cross-disciplinary applications that span organizational boundaries, enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications. It also Supports complete sensor data life cycle from data collection to the backend decision support system.

6 Research Challenges

Research challenges in which areas research works can be done: i) Complex Event Processing and Management. ii) Massive Scale and Real Time Data Processing. iii) Large Scale Computing Frameworks. iv) Harvesting Collective Intelligence

7 The Proposed Architecture of the Cloud Computing with Sensor Networks

The architecture of the cloud computing with sensor network is shown in the Figure 6 [1], this architecture Enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications. Supports complete sensor data life cycle from data collection to the backend decision support system. Vast amount of sensor data can be processed, analyzed, and stored using computational and storage resources [3] of the cloud.

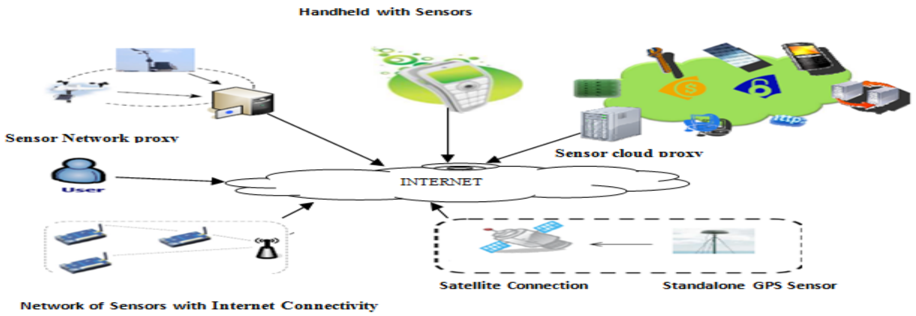


Fig. 6. Sensor Cloud Architecture

7.1 Requirements for the System

Cloud computing model is mainly based on pipes and filters [17] .The pipes and filter design (see Figure 7) is used in digital processing applications, which also used in wireless sensor networks [18].

Pipes are used to buffer data and provide uniform interconnection mechanism of filters. Filters process and transform input data and deliver it to an output port.

The general system architecture, which integrates cloud computing with wireless sensor networks, contains several basic services.

The main requirements for the system are: i) Receive and manage sensor data from heterogeneous motes. ii) Manage a set or chain of filters that perform on-line analysis on sensor data. iii) Run filters offline on a given set of sensor data. iv) Permanently run filters on a given set of sensoric data. v) Provide different, user definable views and visualizations on the sensor data and calculation results. vi) Provide an interface for changing existing filters or to develop new filters out of an existing domain specific modeling tool. vii) Provide an interface for data export, so that the stored data can easily be taken from the Cloud storage and to be used in non-cloud solutions. viii) Provide a notification service, e.g. a filter or machine learning component identifies a specific situations or has finally calculated a specific result. ix) Provide data access rules x) Provide configuration capabilities for filter chains, (web) services, notifications, and data access rules. xi) Provide a management console for the configuration of the whole system.

7.2 Proposed System Level Architecture of Cloud Computing Integration with Wireless Sensor Network(WSN)

The system level architecture of cloud computing integration with wireless sensor network in Figure 8 shows the collection of base services provided to the user as SaaS (Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service) applications as cloud service providers. The Wireless Sensor Network Analytical Services and Cloud groups all services that are necessary to fulfill the requirements that are integrate cloud computing with wireless sensor networks. The following main services are necessary to collect and analyze sensor network data within the cloud: Necessary Services for wireless sensor network and Cloud

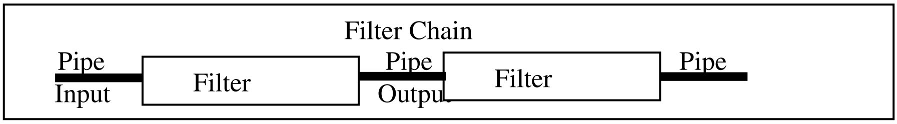


Fig. 7. Pipe, Filter, and Filter Chain

Global Sensor Data Management or Sensor Metadata Management or Cloud Based Management System: It is responsible for the management of the sensor data within the Cloud Computing environment. Since the cloud computing environment provide several ways for storing data, e.g.: Google AppEngine offers Bigtables [4,15] for data persistence and Microsoft Azure provides BLOBs, queues and tables, it's necessary to have a flexible data access layer, which raises the level of abstraction, so that persistence mechanism can be easily exchanged to the global sensor networks.

Provide Runtime for Filter Chains: Filters are usually configured for a specific filter chain. A filter chain is a reliable runtime environment for filters and it executes various user defined filters.

User Interface via Web Browser: A user interface provide the interaction between people (users) with a machine via Web Browser. The user interface includes hardware (physical) and software (logical) component.

Filter Chain and Management of the Filter: Filters and filter chains we need some management. This management service allows the administration of filters, so that a user is able to add new filters, delete filters and aggregate existing filters to combined filters. Since many filter chains are executed in parallel, it is necessary to offer a flexible configuration mechanism.

Visualization / Views for Data Analysis: The visualization service provides various predefined and user-defined views on the data and analysis results. The visualizations and views can be implemented with languages like data warehouses, OLAP, The spatial OLAP Visualization and Analysis Tool (SOVAT), NET Reporting and Heat map. So a powerful visualization is necessary to manage the sensor data.

Notification Service: This service is a mechanism to inform external applications and services about specific situations, where it fires an event. This could be for example, an indication that a gas pump failure is approaching

8 Key Components of the Proposed Architecture of the Cloud Computing with Sensor Networks

There are some key components of the cloud computing with sensor Networks as follows

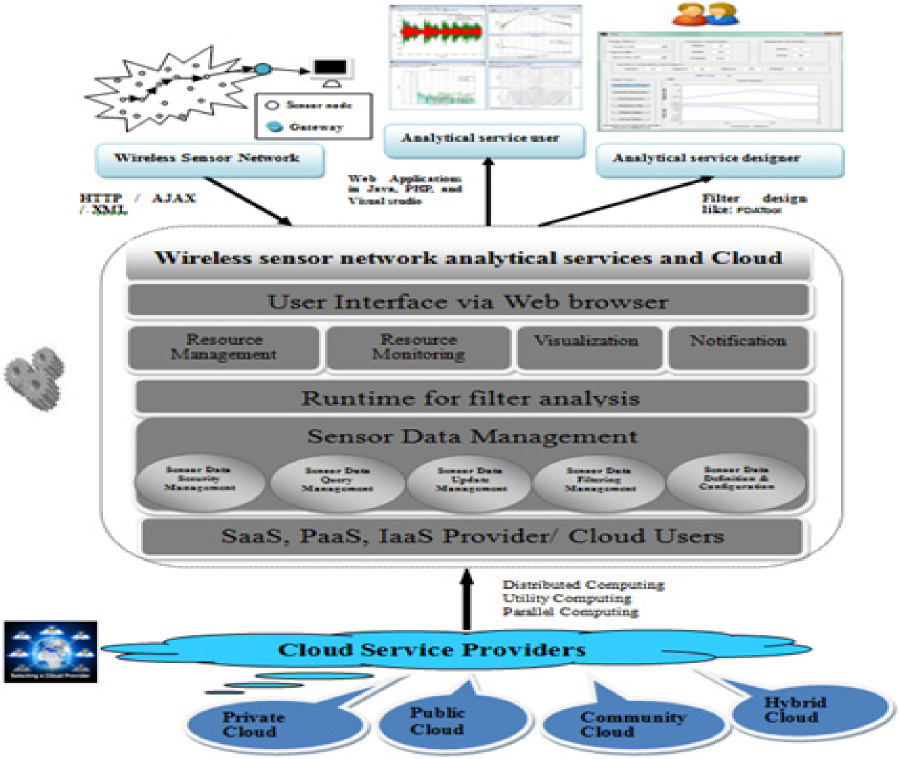

Sensor node
Gateway
Wireless Sensor Network
Analytical service user
Analytical service designer
HTTP / AJAX / XML
Web Applications in Java, PHP, and Visual studio
Filter design like: FDATool
Wireless sensor network analytical services and Cloud
User Interface via Web browser
Resource Management
Resource Monitoring
Visualization
Notification
Runtime for filter analysis
Sensor Data Management
Sensor Data Security Management
Sensor Data Query Management
Sensor Data Update Management
Sensor Data Filtering Management
Sensor Data Definition & Configuration
SaaS, PaaS, IaaS Provider/ Cloud Users
Distributed Computing
Utility Computing
Parallel Computing
Cloud Service Providers
Private Cloud
Public Cloud
Community Cloud
Hybrid Cloud

Fig. 8. Proposed System Level Architecture of Cloud Computing Integration with Wireless Sensor Networks

8.1 Sensor-Cloud Proxy

Sensor cloud proxy provide the interface between sensor resources and the cloud fabric, which Manages sensor network connectivity between the sensor resources and the cloud., that exposes sensor resources as cloud services, to manages sensor resources via indexing services. Cloud discovery services used for resource tracking. for manages sensing jobs for programmable sensor networks, manages data from sensor networks, data format conversion into standard formats (e.g. XML).

8.2 Sensor-Network Proxy

Sensor network proxy provides the connection to sensor resources that do not have direct connection to the cloud, sensor network is still managed from the Sensor-Cloud Interface via Sensor Network proxy, which collects data from the sensor network continuously or as and when requested by the cloud services to enhances the scalability of the Sensor Cloud, finally sensor cloud proxy provides various services for the underlying sensor resources, e.g. power management, security, availability, Quality of Services(QoS).

9 Conclusion and Future Work

The communication among sensor nodes using Internet is a challenging task since sensor nodes contain limited band width, memory and small size batteries. The issues of storage capacity may be overcome by widely used cloud computing technique with sensor networks. Cloud integration with WSN mechanism may provides dynamic collaboration between clouds to enable many services. We also conclude that the cell phone cannot perform so many tasks! There for cloud computing is useful with sensor networks. As we know "Sensor networks are distributed across extended terrain so they open up an entirely new scope of applications. Another critical feature of this technology is that it has a very light footprint, it can be installed using fairly non-intrusive methods, and as a result, we do not impact the environment we are trying to observe." Everyone should interest to discuss about the use of cloud computing for real-world applications, and explore opportunities for collaboration which lead to the intelligence integration into the Internet. This solution has been extended to sensor clouds, which leads to high availability and hence reliability is achieved.

Acknowledgment. We would like to acknowledge the support and input from our family, Mr. Munish Kumar (HOD IT, CDAC Noida,) –INDIA, Dr. J.V. Desai (Dean), Prof. Prema K.V. (HOD) CSE, and our colleagues from FET, MITS Laxmangarh, Sikar (Raj.) India, NIET Greater Noida Uttar Pradesh and THDC Institute of Hydropower Engineering & Technology Tehri Uttarakhand –India.

References

1. Beng, L.H.: Sensor cloud: towards sensor-enabled cloud services. Intelligent Systems Center. Nanyang Technological University (April 13, 2009)
2. Introduction to Cloud Computing architecture White Paper on sun Microsystems, 1st edn (June 2009)
3. Ulmer, C., Alkalai, L., Yalamanchili, S.: Wireless distributed sensor networks for in-situ exploration of mars, Work in progress for NASA Technical Report, `http://users.ece.gatech.edu/`
4. Chang, F., et al.: A Distributed Storage System for Structured Data. In: Seventh Symposium on Operating System Design and Implementation, OSDI 2006, Seattle, WA (2006)
5. Lal, N.: A novel survey on Cloud Computing Issues. Is published in International Journal of Computer Information Systems (IJCIS) 01(02), 18–21 (2010)
6. Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud Computing. University of California, Berkeley (2009), UCB/EECS-2009-28
7. Zhao, F., Guibas, L.: Wireless Sensor Networks - An Information Processing Approach. Morgan Kaufmann (2004)
8. Joseph, J.: Cloud Computing: Computing: Patterns For High Availability, Scalability, And Computing Power With Windows Azure. MSDN Magazine (May 2009)
9. Kurschl, W., Mitsch, S., Schönböck, J.: Modeling Distributed Signal Processing Applications. In: Proceedings of 6th International Workshop on Body Sensor Networks, Berkeley, USA (2009)
10. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: 'Wireless Sensor Networks: A Survey. Computer Networks (Elsevier) Journal, 393–422 (March 2002)
11. Shi, J., Liu, W.: A Service-oriented Model for Wireless Sensor Networks with Internet. Proceedings of the Fifth International Conference on Computer and Information Technology (CIT 2005) (2005)
12. Cloud Computing Conference, Jayshree Ullal, President and Chief Executive Officer, Arista Networks. Abstract (2009)
13. Madden, S., Franklin, J., Hellerstein, J.M., Hong, W.: TinyDB: An Acqusitional Query Processing System for Sensor Networks. ACM Transactions on Database Systems, 47 (2005)
14. Levis, P., et al.: TinyOS: An Operating System for Wireless Sensor Networks. Ambient Intelligence (2005)
15. Severance, C.: Using Google App Engine. O'Reilly (2009)
16. Mell, P., Grance, T.: Draft nist working definition of cloud computing - v15. 21 (2005, 2009)
17. Gamma, E., Helm, R., Johnson, R.E.: Design Patterns. Elements of Reusable Object-Oriented Software. Addison-Wesley Longman (1995)
18. Kurschl, W., Mitsch, S., Schonbock, J.: Modeling Distributed Signal Processing Applications. In: Proceedings of 6th International Workshop on Body Sensor Networks, Berkeley, USA (2009)

Secret Image Sharing Scheme Based on Pixel Replacement

Tapasi Bhattacharjee[1] and Jyoti Prakash Singh[2]

[1] Techno India, Salt Lake, Kolkata, India
tapasi.dgp@gmail.com
[2] National Institute of Technology Patna, Bihar, India
jyotip.singh@gmail.com

Abstract. Dividing an image in several components and sharing through different channel is popular way of sharing and storing sensitive image data. We proposes here a simple image secret sharing method based on random matrices. These random matrices act as key for secret sharing. The technique allows a secret image to be divided into three image shares where each share individually looks meaningless. To reconstruct the secret image all three shares have to be used. This method has no pixel expansion and can reconstruct the secret image precisely. This scheme can be directly applied on gray scale images and can easily be extended to binary and color images. Experimental results prove that this scheme can generate good quality of reconstructed images.

Keywords: Secret sharing, Visual Secret Sharing, Security, Structured Similarity Index Metric, Peak Signal to Noise Ratio.

1 Introduction

With the development of network technology, information can be distributed and transmitted over the internet rapidly and conveniently. Replicating the important information will offer more chances to intruders to gain access to it. On the other hand, having only one copy of the information means that if this copy is destroyed there is no way to retrieve it. Thus, there is a great need to keep information in a secure and reliable way. Hence, secret sharing came to use. Secret sharing method divides a secret into some shares called shadows, where each shadow looks meaningless and individual shares are of no use on their own. These shadows are distributed to the participants. Only a set of qualified participants can recover the secret, while the non-qualified participants can not even get a clue of the secret information. The concept of secret sharing scheme was first introduced by Blakley [1] and Shamir [2] independently. Both the schemes were (k, n) threshold secret sharing schemes. Shamir [2] used polynomial-based technique to share the secret among n participants and Blakley [1] used a geometric approach. Shamir's technique creates a $(k > 1)$ degree polynomial with random coefficients in the range $(0 \cdots p)$, where p is a prime number. The constant term of this polynomial is the secret message. Lagrangian interpolation

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 517–526, 2013.

technique is used for the reconstruction of the secret from any k or more shares. Blakley's [1] technique assumes that secret is a point in a k-dimensional space. Hyper planes intersecting at this point are used to construct the shares. Coefficients of n different hyper planes constitute the n shares. Karnin et al. [3] suggested the concept of perfect secret sharing (PSS) where zero information of the secret is revealed for an unqualified group of $(k-1)$ or fewer members. The unqualified group cannot obtain any information about the secret and the unqualified group cannot reconstruct the secret. Brickell [4] was the first who introduced the notion of ideal structures of secret sharing scheme. A secret sharing scheme is called ideal if the shares are taken from the same domain as the secret. Cimato et al. [5] proposed (n,n) threshold Visual Secret Sharing (VSS) schemes. But this scheme had the disadvantages of pixel expansion and low contrast. Thien and Lin [6] proposed a (k,n) threshold-based image Secret Sharing Scheme based on Shamir's Secret Sharing Scheme [2] to generate image shares. This method reduces the size of image shares to become $\frac{1}{k}$ of the size of the secret image. Bai [7] developed a secret sharing scheme using matrix projection. The idea is based upon the invariance property of matrix projection. This scheme can be used to share multiple secrets. Although the reconstructed images in these schemes can be revealed by simply stacking the collected shadows but the pixel expansion problem occurred. Later on Tuyls et al. [8] proposed (n,n) Secret Sharing scheme for binary images with no pixel expansion and precisely reconstructed image. Yi et. al. [9] presented two (n,n) schemes for color image. The schemes also have no pixel expansion but the secret image was not precisely reconstructed. Wang et al. [10] proposed (n,n) scheme for gray scale image. The scheme has no pixel expansion and gives an exact reconstruction. All schemes in [8], [9] and [10] are constructed based on Boolean operation, which need bit-wise operation when sharing gray scale and color images. To share a lossless secret image, the schemes [11,12] used two pixels to represent the exceeding gray values. Nevertheless, this resulted in the expansion of the secret image and reduced the sharing capacity as well as distort the quality of the shadow image. Chao et al. [13] proposed a method to extend (n,n) scheme to (k,n) scheme by using shadows-assignment matrix. Dong and Ku [14] proposed a new (n,n) secret image sharing scheme with no pixel expansion. In their scheme reconstruction is based on addition which has low computational complexity. Singh et al. [15] proposed an image secret sharing method based on some random matrices that acts as a key for secret sharing. The technique allows a secret image to be divided into four image shares with each share individually looking meaningless. The share generation algorithm works by converting three pixels of the secret image to one pixel each of four different shares based on four random matrices. So, each share is reduced by $1/3^{rd}$ of the original secret image. Peng Li et al. [16] proposed (n,n) visual secret sharing scheme without distortion by computation and get a better visual quality of the reconstructed image. But pixel expansion problem was still a major problem in this scheme.

We propose here a simple secret image sharing scheme based on two random matrices. The random matrices was used to scramble the pixel positions of the original secret images. For this scrambling operation, first random matrix give the row value and second random matrix give the column value where to place the pixel of the original secret image. This scheme has no pixel expansion and can reconstruct the secret image precisely. The proposed scheme can be directly applied on gray scale images and can be easily extended on binary and color images.

The rest of the paper is organized as follows: section 2 describes the idea of the proposed scheme. The results of this scheme and comparisons with other similar works are discussed in section 3. Finally, Section 4 summarizes the paper and gives the concluding remarks.

2 Proposed Scheme

Our proposed secret sharing scheme is discussed in this section. Our sharing algorithm is divided into two phases: Sharing phase and Reconstruction phase. Each phase is discussed in the following section with examples.

2.1 Sharing Phase

The pixels of the original secret image A is scrambled with the help of two random matrices, R_1 and R_2 to generate another image S_3. The random matrices R_1 and R_2 are considered as first two shares S_1, S_2 and the generated image is considered as the third share S_3. The pixel value of $(i, j)^{th}$ location of actual secret image is placed to (x, y) location of the third share S_3 if that location is not holding any other pixel value. The x value is the content of $(i, j)^{th}$ location of R_1 matrix whereas y is value of $(i, j)^{th}$ location of R_2 matrix. If $(x, y)^{th}$ location of S_3 is already occupied by some other pixel then we generate a random pair (x', y') and check if that location of S_3 is free. If it is not free, we choose another random pair and check it again. This process is repeated till we get (x', y') value which is a free position in S_3. When such (x', y') is found, we assign the $A[i, j]$ to $S_3[x', y']$. The corresponding location of random matrices are also updated by assigning $R_1[i, j] = x'$, and $R_2[i, j] = y'$. For an example, say the first position $(0, 0)^{th}$ of the secret image is 136 and $(0, 0)^{th}$ position of R_1 and R_2 matrices are 100 and 50 respectively. 136 is placed at $(100, 50)^{th}$ position of a new matrix, S_3 which is the third share of the proposed scheme. Sat the 2^{nd} pixel value $(0, 1)^{th}$ of the A, R_1 and R_2 are 116, 80, 100 respectively. Then 116 will be placed at $(80, 100)^{th}$ location of S_3 matrix. This process will continue for all the pixel values of A which gives three shares S_1, S_2 and S_3. The share generation algorithm is given below in pseudo-code.

Share Generation Algorithm

Input: A gray-level secret image A of size $h \times w$ and two Random matrices R_1 and R_2 containing values between 0 to 255 of size $h \times w$

Output: Three shares S_1, S_2 and S_3 of size $h \times w$

```
$S_1 = R_1$ and $S_2 = R_2$
```

```
for i=1 to h do {
   for j=1 to w do {
      $X = R_1[i, j]$
      $Y = R_2[i, j]$
      if ($S_3[X, Y] == 0$)
         $S_3[X, Y] = A[i, j]$
      else {
         $X' = rand()$
         $Y' = rand()$
         while(!$is\_blank\_position(X', Y')$){
            $X' = rand()$
            $Y' = rand()$
         }
         $S_3[X', Y'] = A[i, j]$
         $R_1[i, j] = X'$
         $R_2[i, j] = Y'$
      }
   }
}
```

2.2 Revealing Phase

In revealing phase, all the three shares, S_1, S_2 and S_3 are considered as three input images. Values at $(0, 0)^{th}$ positions of S_1 and S_2 matrices are accessed first. Say these are X' and Y' respectively. Then the value which is stored at (X, Y) position of S_3 matrix is stored as 1st pixel value of a new matrix, A'. Say the first values of location $(0, 0)^{th}$ of S_1 and S_2 are 100 and 50 respectively. The value stored at $(100, 50)^{th}$ location of S_3 matrix is 136. This value, 136 is set as 1^{st}, $(0, 0)^{th}$ pixel value of a matrix, A. Similarly the value of location $(0, 1)^{th}$ of S_1 and S_2 matrices are 80 and 100 respectively. The value stored at $(80, 100)^{th}$ location of S_3 matrix (116) set as $(0, 1)^{th}$ pixel value of A'. This process continues for all the values of S_1 and S_2 matrices. The complete reconstruction algorithm is given below in pseudo-code

Reconstruction Algorithm
Input: Three shares S_1,S_2 and S_3 of size $h \times w$
Output: A gray-level recovered image A' of size $h \times w$

```
for i=1 to h do {
   for j=1 to w do {
      $A[i, j] = S_3[S_1[i, j], S_2[i, j]]$
   }
}
```

2.3 Complexity Analysis

The time complexity of our share generation algorithm in best case is $O(n^2)$ because it just scans the original secret image once and places the pixels to another image S_3. In worst case, the time complexity turns out to be $O(n^4)$ because the while loop may run for $O(n^2)$ time for searching for a blank position. In average case it will be $O(n^2)$ only as it will find the blank position to a nearby position in finite number of steps. The time complexity of reconstruction algorithm is always $O(n^2)$.

2.4 Proposed Scheme for Binary and Colour Images

A binary image has only two possible values for each pixel: black and white which can be represented by a single bit only. To use our proposed scheme on binary images, the binary images are converted to gray scale image by combining neighbouring 8 bits to 1 byte. Now the proposed scheme for grayscale image can be applied on binary images. In revealing phase, a corresponding step should be added to split 1 byte of the revealed gray scale image into 8 bits to get the recovered secret image. A color image can be represented as three gray scale images corresponding to each color plane red, green and blue. To extend the proposed schemes for color images, decompose the color image into three components of R, G and B, each of which can be seen as grayscale images. Then perform the proposed scheme for grayscale image to each component R, G and B separately. Finally, compose R, G and B components are combined together to generate color shares.

3 Experimental Results

This section presents and analysis the experimental results by using the proposed method. Top evaluate the performance of our new scheme, we have used eight different images. The images are Lena.jpg, Airplane.jpg, Flower.jpg, Baboon.jpg, Duck.bmp, Child.bmp, Lady.jpg and Logo.tiff. The codes were developed in Matlab 7.0 running on Microsoft Windows XP system with a Pentium Dual Core Processor. Due to space limitations, the graphical result of two images are shown here. Our first secret image is a gray scale image of Duck shown Fig. 1 of size 150 × 170. Fig. 2, Fig. 3 and Fig. 4 are the three shares, S1, S2 and S3 each of size 150 × 170 generated by our algorithm. Fig. 5 is the recovered image of size 150 × 170 obtained by our algorithm.

Our second secret image is Lena image of size 512 × 512 which is shown in Fig. 6. The share generated by our proposed algorithm is shown in Fig. 7, 8 and 9. The shares are also of size 512 × 512. The reconstructed image obtained by the combining all three share images is shown in Fig. 10.

Our proposed technique is not lossless because some pixels of original image can not be written to third share because we can not get a free position in third share. But the visual quality of recovered image obtained by our algorithm is

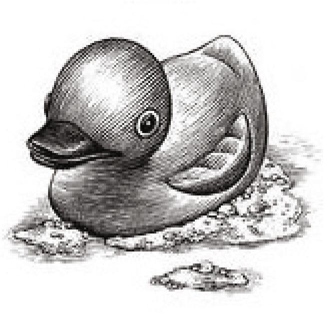


Fig. 1. The Duck image

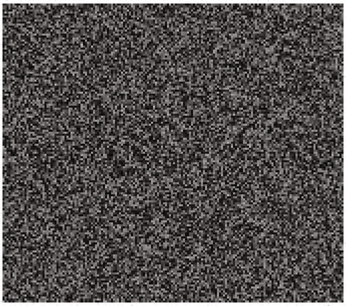


Fig. 2. First share of Duck

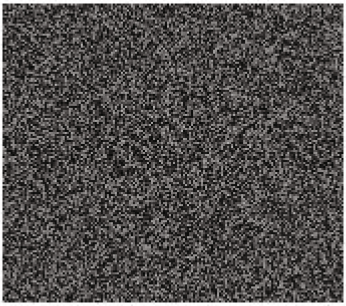


Fig. 3. Second share of Duck

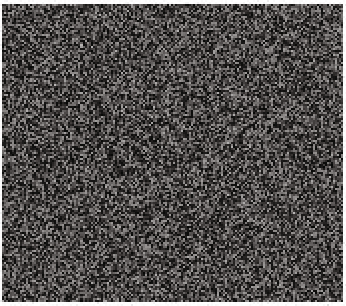


Fig. 4. Third share of Duck image

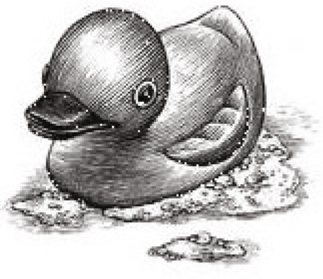


Fig. 5. The recovered Duck

good enough. To check the visual quality of the output images, we have used peak-signal-to-noise ratio (PSNR) metric which is defined as

$$PSNR = 10 \times log \frac{255^2}{MSE} \tag{1}$$

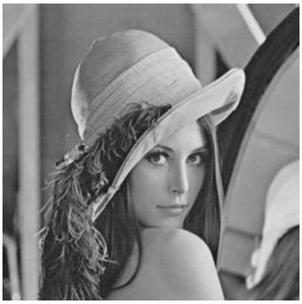


Fig. 6. The Original Lena image

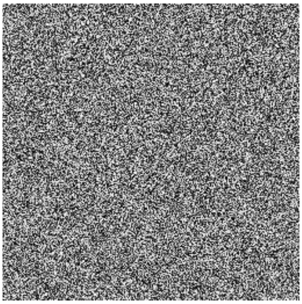


Fig. 7. First share of Lena

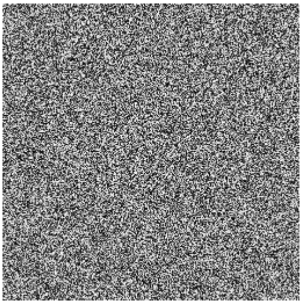


Fig. 8. Second share of Lena

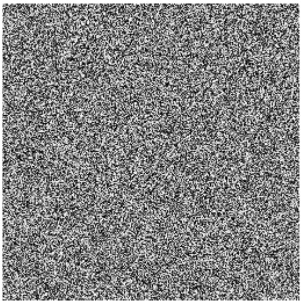


Fig. 9. Third share of Lena

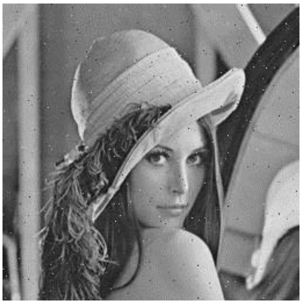


Fig. 10. The recovered Lena image

where

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (h_{i,j} - h'_{i,j})^2, \tag{2}$$

where $h_{i,j}$ is the pixel value of the original image and the $h'_{i,j}$ is the pixel value of the recovered image. MSE is the Mean Squared Error.

Table 1. The PSNR values and standard deviation of different secret images

Image Name	PSNR	Std. Deviation
Lena.jpg	31.36	58.9
Airplane.jpg	33.56	44.44
Flower.jpg	31.44	52.38
Baboon.jpg	26.52	69.92
Duck.bmp	29.76	60.39
Child.bmp	34.10	38.42
Lady.jpg	21.36	86.19
Logo.tiff	13.68	116.43

The PSNR values for various images we have used for our experimentation is given in Table 1. As one can see from Table 1 that the PSNR on different images range from 18 to 31. To find out the reason for this variation of PSNR, we studied the statistical properties of images. We found that for those images whose standard deviation is large, the PSNR is low whereas fro those images whose standard deviation is low, the PSNR is high. Our algorithm works well on those images whose standard deviation is low. The other criterion of secret sharing scheme is that none of the shares should posses any information about the original secret but the recovered image should be similar to the original secret. Structural Similarity Index Metrics (SSIM) is such a metrics for measuring the similarity between two images. SSIM compares local patterns of pixel intensities that have been normalized for luminance and contrast [17]. SSIM is defined as,

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

Where x and y denote the original and recovered image, respectively.
μ_x the average of x_{ij};
μ_y the average of y_{ij};
σ_x^2 the variance of X;
σ_y^2 the variance of Y;
σ_{xy} the covariance of X and Y;
$c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ two variables to stabilize the division with weak denominator;
L the dynamic range of the pixel-values (typically this is $2^{\#bits/pixel} - 1$);
$K_1 = 0.01$ and $k_2 = 0.03$ by default

The resultant SSIM index is a decimal value between 0 and 1, and value 1 is only reachable in the case of two identical sets of data. The SSIM values for various shares and reconstructed images with their respective images is given in Table 2. The SSIM value of each individual share of every image is very low signifying that shares do not reveal any information about the original secret. The SSIM values of reconstructed images are nearly 1 signifying that the reconstructed images are similar to original images.

Table 2. The SSIM values for different of shares and reconstructed image with secret images

Image Name	Share1	Share2	Share3	Recons Image
Lena.jpg	0.0479	0.0558	0.1179	0.9985
Airplane.jpg	0.0704	0.1015	0.0565	0.8995
Flower.jpg	0.0090	0.0047	0.0047	0.9557
Baboon.jpg	0.0626	0.0508	0.1928	0.9593
Duck.bmp	0.0737	0.2683	0.0298	0.9930
Child.bmp	0.0102	0.0113	0.0384	0.9564
Lady.jpg	0.0093	0.0030	0.0059	1
Logo.tiff	0.0508	0.0395	0.0261	0.9817

3.1 Comparison with Similar Works

We have compared our proposed scheme to other similar published works in terms of contrast, pixel expansion, and reconstruction operation. As one can see from table 3 that our scheme is better than existing schemes in terms of contrast ratio. We achieve a contrast ration of 1. In terms of reconstruction, we use pixel replacement whose algorithmic complexity is $O(n^2)$ only.

Table 3. Comparison of different secret sharing schemes

Category	Contrast	Pixel Exp	Reconstruction
Yi [9]	≤ 1	1	XOR
Cimato [5] & Li [16]	$\ll 1$	$\gg$1	OR (Stacking)
J.P.Singh [15]	$\ll 1$	$\frac{1}{3}$	Bitwise
Proposed Scheme	1	1	Pixel Replacement

4 Conclusion

In this paper we have proposed a simple secret sharing scheme based on two random matrices which can generate three shares. The proposed scheme does not increase the size of the secret image by expanding pixels. The individual shares do not reveal any information about the original secret image. This scheme can be directly used to share gray scale images. It also can be extended with binary and color images. We also find a relationship of image statistical properties with our sharing scheme. Our scheme gives better results for images whose standard deviation is low. We are in the process of extending this scheme to (k, n) secret sharing scheme with ideal contrast.

References

1. Blakley, G.R.: Safeguarding cryptographic keys. AFIPS NCC 48, 313–317 (1979)
2. Samir, A.: How to share a secret. Communications of ACM 22(11), 612–613 (1979)
3. Karnin, E.D., Greene, J.W., Hellman, M.E.: On secret sharing systems. Information Theory 29(1), 35–41 (1983)
4. Brickell, E.F.: Some ideal secret sharing schemes. J. Comb. Math. Comb. Comput. 6, 105–113 (1989)
5. Cimato, S., De Prisco, R., De Santis, A.: Optimal colored threshold visual cryptography schemes. Designs Codes and Cryptography 35(3), 311–315 (2005)
6. Thien, C.C., Lin, J.C.: Secret image sharing. Computers and Graphics 26(5), 665–670 (2002)
7. Bai, L.: A strong ramp secret sharing scheme using matrix projection. In: 2nd Intl. Workshop on Trust, Security and Privacy for Ubiquitous Computing, pp. 656–660. IEEE (2006)
8. Tuyls, P., Hollmann, H.D.L., van Lint, J.H., Tolhuizen, L.: Xor-based visual cryptography schemes. Designs Codes and Cryptography 37, 169–186 (2005)
9. Yi, F., Wang, D.S., Luo, P., Dai, Y.Q.: Two new color (n, n)-secret sharing schemes. Journal on Communications 28(5), 30–35 (2007)
10. Wang, D., Zhang, L., Ma, N., Li, X.B.: Two secret sharing schemes based on boolean operations. Pattern Recognition 40(10), 2776–2785 (2007)
11. Chang, C.C., Hsieh, Y.P., Lin, C.H.: Sharing secrets in stego images with authentication. Pattern Recognition 41(10), 3130–3137 (2008)
12. Zhao, R., Zhao, J.J., Dai, F., Zhao, F.Q.: A new image secret sharing scheme to identify cheaters. Computer Standards and Interfaces 31(1), 252–257 (2009)
13. Chao, K.Y., Lin, J.C.: Secret image sharing: a boolean-operations based approach combining benefits of polynomial-based and fast approaches. International Journal of Pattern Recognition and Artificial Intelligence 23(2), 263–285 (2009)
14. Dong, L., Ku, M.: Novel (n, n) secret image sharing scheme based on addition. In: 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 583–586 (2010)
15. Singh, J.P., Nag, A., Bhattacharjee, T.: Random matrices based image secret sharing. International Journal of Advanced Research in Computer Science 2(4), 104–108 (2011)
16. Li, P., Ma, P.-J., Su, X.-H., Yang, C.-N.: Improvements of a two-in-one image secret sharing scheme based on gray mixing model. Journal of Visual Communication and Image Representation 23(3), 441–453 (2012)
17. Wang, Z., Bovik, A.C.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)

An Application of Defeasible Logic Programming for Firewall Verification and Reconfiguration

Pritom Rajkhowa[1], Shyamanta M. Hazarika[1], and Guillermo R. Simari[2]

[1] Department of Computer Science & Engineering
Tezpur University, Tezpur 784028, India
{pritomr,smh}@tezu.ernet.in
[2] Department of Computer Science & Engineering
Universidad Nacional del Sur, Bahia Blanca, Argentina
grs@cs.uns.edu.ar

Abstract. Firewalls are the frontier defense in network security. Firewalls provide a set of rules that identify how to handle individual data packets arriving at the network. Firewall configuration is increasingly becoming difficult. Filter properties called *anomalies* hint at possible conflicts between rules. An argumentation framework could provide ways of handling such conflicts. Verification of a firewall involve finding out whether anomalies exist or not. Reconfiguration involves removing critical anomalies discovered in the verification phase. In this paper, we show how a Defeasible Logic Programming approach with an underlying argumentation based semantics could be applied for verification and reconfiguration of a firewall.

Keywords: Defeasible Logic Programming, stateless firewall, stateful firewall, anomaly, argumentation.

1 Introduction

Firewalls are the frontier defense in network security. Firewalls filter out unwanted packets coming from or going to the secured network. Firewall rules are specified in order of priority and are of the form:

<order> : if <network-conditions> take <action>

However, managing firewall configuration is increasingly becoming complex. Errors in firewall configuration includes conflict among the existing rules, failing to specify all the required rules that enforce a certain level of security, inappropriate rule ordering, invalid syntax etc. The complexity and interdependency of policy rules makes firewall policy management a challenging task; continuous evolution of networks making it even more difficult. Filter properties called *anomalies* that hint at possible misconfiguration have therefore been introduced by Network Management researchers [1]. Verification of a firewall involves finding out whether anomalies exist or not. Anomalies make a firewall do not conform to the policy specification [2]. Reconfiguration of a firewall involves removing critical anomalies discovered in the verification phase.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 527–542, 2013.

Formal methods have been used in firewall anomaly detection. Considerable work on approaches based on logic has been undertaken. A formal logic in understanding meaning of firewall rules was proposed [3]. One of the earliest approach proposed in [4] represents firewall rule sets as Binary Decision Diagrams (BDDs), allowing the rule set to be analysed as boolean expressions. Similar to this, Multi Terminal Interval Decision Diagrams were used [12] to enable efficient packet classification. [7] presented a tool based on constraint logic programming (CLP) for analyzing firewall rules. More recently, firewalls anomalies are being seen as spatial properties. Thanasegaran et al [11] explored the spatial relationships amongst rules in a bit-vector based spatial calculus, BISCAL to detect and classify the conflicts. Villemaire and Hallé [8] and Hazarika [10] showed that the condition part of a rule in a rule-based firewall can be viewed as a spatial region while their sequential application lends a temporal aspect. With this notion they introduced spatio-temporal logics for firewalls; anomalies are properties within the logic. Model-checking such properties can account for anomalies in a firewall. [9] proposed a logic based on notions related to visibility for defining firewall anomalies. [5] describes technique based on argumentation for Logic Programming with Priorities. [6] use a system of meta-level argumentation for firewall configuration and resolving conflict.

In this paper, we show how defeasible argumentation could be applied in the firewall domain to yield interesting results through representation and reasoning about conflicts and reconfiguration. Our work differs from previous approaches using argumentation [5,6], as we express firewall properties within a *defeasible* logic in order to explore defeasibility in firewall policy. Defeasible rules are defined as rules that provide a weak link liable to defeat or overrule by some rule after all has been considered; e.g. check a rule for anomaly if only it is causing conflict with some rule in the rules set. It is unnecessary to go and check each rule for anomaly. Under a consideration that each rule is conflict free unless it is defeated by some conflict; we show how Defeasible Logic Programming (DeLP) could be exploited for validation and reconfiguration of a firewall.

2 Background: DeLP and Firewall Anomalies

2.1 Defeasible Logic Programming

DeLP [15] is an alternative form of declarative programming. DeLP is a blend of Logic Programming with Defeasible Argumentation, allowing representation of tentative knowledge and leaving for the inference mechanism the task of finding the conclusions that the knowledge base warrants. Two kinds of rules considered by DeLP, makes it different from Logic Programming, from which it inherits the formal characterization of programs as sets of rules. The rules considered include strict rules and defeasible rules. Strict rules are assumed to represent sound knowledge. Defeasible rules are assumed to represent tentative knowledge which may be defeated by other information. DeLP functions by answering queries ($\mathcal{Q}$). Warranted arguments constructed using rules and facts (considered as special cases of strict rules) specify the answer to the query. An answer is yielded by the inference mechanism based on the warrant procedure that is run upon generation of all the possible arguments. The generated arguments may support or contradict the query $\mathcal{Q}$. The characteristic warrant procedure of DeLP, based on Defeasible Argumentation, enables comparison as well as selection of only one of the two contradicting arguments.

Definition 1. Defeasible Logic Program*: A Defeasible Logic Program $\mathcal{P}$ is a set* (Π, Δ)*, where* Π *stands for the union of strict rules* Π_{R} *and facts* Π_{F}*;* Δ *denotes defeasible rule. Strict Rules are rules in the classical sense, i.e., whenever the permission of the rules is given, we are allowed to apply the rule and get the conclusion. Strict rule are of the form* $p \leftarrow q_1,\ q_2,\ q_3,..,\ q_{n-1},\ q_n$*. Defeasible Rules are of the form* $p \prec q_1,\ q_2,\ q_3,..,\ q_{n-1},\ q_n$*. Defeasible Rules are contingent rules that get defeated by contrary evidence. Facts (strict rules with empty body) are known truth that are treated as ground literals.*

Construction of arguments in DeLP is a result of the literal derivation and provides a tentative support for the claims. An argument $\mathcal{A}$ for a query $\mathcal{Q}$, (denoted $\langle\mathcal{A},\mathcal{Q}\rangle$) can be considered as a proof for $\mathcal{Q}$ where $\mathcal{A}$ is a set (possibly empty) of ground defeasible rules in conjunction with a set that satisfy the additional constraints of non-contradiction (i.e.an argument s should not allow the derivation of contradictory literals) and minimality (i.e., the set of defeasible information used to derive $\mathcal{Q}$ should be minimal). Mechanism similar to the usual query-driven SLD derivation from logic programming involving backward chaining on both strict and defeasible rules is used to obtain arguments. Incomplete and tentative information of a program $\mathcal{P}$ may lead to an attack on argument $\langle\mathcal{A},\mathcal{Q}\rangle$ by other arguments which may be derived from the same program $\mathcal{P}$. An argument$\langle\mathcal{B},\mathcal{R}\rangle$ is considered to be a counter-argument for $\langle\mathcal{A},\mathcal{Q}\rangle$ if a subargument $\langle\mathcal{A}',\mathcal{Q}'\rangle$ (with $\mathcal{A}'$ belonging to $\mathcal{A}$) in $\langle\mathcal{A},\mathcal{Q}\rangle$ exists, such that $\langle\mathcal{B},\mathcal{R}\rangle$ and $\langle\mathcal{A}',\mathcal{Q}'\rangle$ cannot be accepted simultaneously as acceptance of both will allow inference of contradictory conclusions from $\Pi \cup \mathcal{A}' \cup \mathcal{B}$. The attacking argument $\langle\mathcal{B},\mathcal{R}\rangle$ is termed defeater for $\langle\mathcal{A},\mathcal{Q}\rangle$ if $\langle\mathcal{B},\mathcal{R}\rangle$ is preferred over $\langle\mathcal{A}',\mathcal{Q}'\ \rangle$. Specificity is the commonly used criterion, however, other criteria can also be adopted.

A recursive process is prompted by the search for defeaters in DeLP thus resulting in the formation of a *dialectical tree.* The original argument at issue forms the root and every defeater of the root argument forms a children node. To avoid circular situations during computation of branches in the dialectical tree additional restrictions are added which guarantee that the tree is finite. The children nodes in the tree can be marked as defeated (**D** nodes) or as undefeated(**U** nodes). Marking of the dialectical tree is similar to the AND-OR tree where leaves are always marked as undefeated nodes; inner nodes can be marked as undefeated or as defeated. An undefeated root (original argument, $\langle\mathcal{A},\mathcal{Q}\rangle$) of the tree, after being subjected to the above process, is deemed as acceptable or warranted. DeLP solving for a query $\mathcal{Q}$ with respect to a given program $\mathcal{P}$ accounts for determining whether $\mathcal{Q}$ is supported by a warranted argument. Given query $\mathcal{Q}$ there are four possible answers. 'Yes' if there is at least one warranted argument $\mathcal{Q}$ that follows from $\mathcal{P}$.; 'No' if there is one warrant argument for $\sim\mathcal{Q}$; 'Unknown' if $\mathcal{Q}$ is not present in the program; 'Undecided' if neither $\mathcal{Q}$ nor $\sim\mathcal{Q}$ are supported by warranted arguments in $\mathcal{P}$.

To bring home the point of argumentation in DeLP, let us discuss the argumentation of buying a used car. The car is to be brought only if it is in good condition. The DeLP program has five defeasible rules and three facts.
R1: goodCondition(X) $-\!\!\prec$ well_maintain(X)
R2: ~goodCondition(X) $-\!\!\prec$ longrun(X)
R3: goodCondition(X)$-\!\!\prec$ well_maintain(X),longrun(X)
R4: buy(X)$-\!\!\prec$car(X),good_condition(X)
R5: ~buy(X) $-\!\!\prec$ car(X),~good_condition(X)
F1: car(ford)
F2: well_maintain(ford)
F3: longrun(ford)
Rules R1 and R3 represent the good condition of the car, if the car is well maintained irrespective of long run. Rule R2 states that car is not in good condition if it runs long distance. Rules R4 states that we can buy when it is used but well maintained. Rule R5 represent the scenario when we should not buy the car. Facts F1, F2 and F3 represents that ford is a well maintained car though it runs quite a long distance. DeLP program help us to conclude by performing the query 'buy(ford)' that we have a warranted argument supporting that we should buy that car. The dialectical tree is shown in Figure 1. Dialectical tree shows

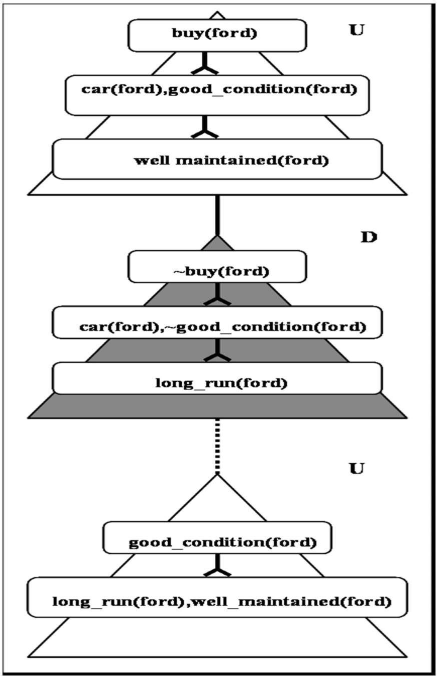

buy(ford)
U
car(ford),good_condition(ford)
well maintained(ford)
D
~buy(ford)
car(ford),~good_condition(ford)
long_run(ford)
U
good_condition(ford)
long_run(ford),well_maintained(ford)

Fig. 1. Dialectical analysis associated with the query 'buy(ford)'

how DeLP, which has support for both logical programming and argumentation, can be used for commonsense reasoning of buying a car.

2.2 Firewall Anomalies

Firewall is a set of ordered filtering rules configured primarily based on predefined security policy. Table 1 shows an example of a basic firewall policy. Each rule is formed of a *condition* and an *action*. The most common and basic firewall are based on mainly five fields: protocol, source IP address, source port, destination IP address, destination port. A rule condition is a set of fields; any fields in IP, UDP or TCP headers may be used. An action is taken for the packets matching these fields. Filtering actions are either to *accept*, which passes the packet or to *deny*, which causes the packet to be discarded.

Table 1. A basic firewall policy example.

Order	Protocol	Source IP	Source Port	Destination IP	Destination Port	Action
R1	TCP	150.172.37.20	any	*.*.*.*	80	deny
R2	TCP	150.172.37.*	any	*.*.*.*	80	accept
R3	TCP	150.172.37.[10,30]	any	171.120.32.[10,40]	21	accept
R4	TCP	150.172.37.*	any	171.120.32.40	80	deny
R5	TCP	150.172.37.30	any	*.*.*.*	21	deny
R6	TCP	150.172.37.[30,60]	any	171.120.32.[40,80]	21	accept
R7	TCP	150.172.37.[25,45]	any	171.120.32.[30,65]	21	deny
R8	TCP	*.*.*.*	any	*.*.*.*	any	deny
R9	UDP	150.172.37.*	any	171.120.32.40	53	accept
R10	UDP	*.*.*.*	any	171.120.32.40	53	accept
R11	UDP	*.*.*.*	any	*.*.*.*	any	deny
R12	TCP	150.172.37.[20,80]	any	*.*.*.*	any	deny
R13	TCP	150.172.37.[20,35]	any	171.120.32.[50,65]	21	accept
R14	TCP	192.168.37.[15,40]	any	171.120.32.[150,165]	21	accept
R15	TCP	192.168.37.[25,60]	any	171.120.32.[120,155]	21	deny

In order to build a model for intra-firewall anomaly, one need to determine all relations that may exist between two filtering rules. This has been addressed among other by [1]; and a set of five relations have been identified. Filtering policy within a firewall is dependent on the ordering of filtering rules. Note that for a set of *completely disjoint* filter rules, the ordering is insignificant. This is not usually the case and therefore ordering is important. Else, some rules may always be 'screened' by other rules producing an incorrect policy. Intra-firewall policy anomaly is the existence of such discrepancies. Anomalies are properties of filters that hint at possible misconfiguration.

3 Formal Characterization

In order to model the firewall within our framework, we derive a formal characterization of the firewall rules and anomalies. Depending on the relation that exists between two rules or a combination of rules with another rule, their orders and their actions, the nature of the anomaly may vary.

Definition 2. *A* stateless firewall rule *of order i, is given as* $R_i = \langle f_1, f_2, f_3, f_4, f_5, Action\rangle$*; where each* f_j *is a filtering field. The filtering fields are: protocol, source IP, source port, destination IP and destination port. Action* $\in$ {*allow, deny*} *describes the action to be taken if a packet matches the filtering fields in the rule.*

3.1 Anomalies

Simple Anomalies. The set-theoretic relations between the filtering fields of two rules enables one to determine all the relations that can exist between two rules. Al-Shaer et al [1] defined five possible relations that may exist between filtering rules such that there exists exactly one relation between two rules. These result in *simple* anomalies as listed below. Al-Shaer et. al. [1] give formal definitions of possible anomalies between rules in terms of rule relations.

Definition 3. Simple Shadowing Anomaly*: A rule Y is simply shadowed if there is rule X, preceding Y in firewall rule set, such that all the packets that match Y already match X and specify incompatible action.*

Definition 4. Simple Redundancy Anomaly*: A redundant rule X perform same action on same set of packets as another rule Y; therefore it the redundant rule is removed, the security policy will not be affected.*

Definition 5. Generalization*: A rule Y is generalization of a rule X, preceding Y in firewall rule set, such that all the packets of Y is a superset match of X and specify incompatible action.*

Definition 6. Correlation Anomaly*: Rule X and Y are correlated if some fields of X are subset to corresponding fields in Y.*

Second Order Anomalies. Alfaro et al [16] have shown that there could be anomalies where more than a pair of rules are involved. We refer to these as *second order* anomalies.

Definition 7. Second-order generalization*: A rule X and a group of rules Y exhibit second-oder generalization, if decision of rule X is overridden by combination of later rules* Y_1 *...* Y_n *.*

Definition 8. Second-order shadowing*: A rule X and a group of rules Y exhibit second-oder shadowing, if rule X is shadowed by a combination of later rules* Y_1 *...* Y_n *.*

Intra-state Protocol Anomalies. Apart from the above mentioned anomalies another type of anomaly viz. intra-state protocol anomalies may be observed in the stateful filters [17]. This anomaly is basically related to the inner logic of transport layer protocol states[1].

Definition 9. *A* stateful firewall rule *of order i, is given as* $R_i = (f_1, f_2, f_3, f_4, f_5, state, Action)$ *where each* f_j *is a filtering field. The filtering fields are: protocol, source IP, source port, destination IP and destination port.* $state \in \{New, Establish, Related\}$. *Action* $\in$ *{allow, deny} describes the action to be taken if a packet matches the filtering fields in the rule.*

For example in web applications TCP is used as a transport layer for three way handshake connection establishment with the server. The following operations may be distinguished in three way handshake connection scheme to establish connection between the client and the server.

-To start the connection the client sends a SYN packet to a server at LISTEN stage;
-The client waits till it receives the proper SYN-ACK packet from the server;
-As and when the client receive the ACK reply from the server, it again sends back an ACK packet to the server and goes to the ESTABLISHED state of connection.

Table 2. Firewall Rule Table

Rule	*Protocol*	*SrcIPt*	*SrcPort*	*DestIP*	*DestPort*	*State*	*Action*
R1	tcp	192.168.10.[10,20]	any	10.1.2.1	80	New, Establish	accept
R2	tcp	10.1.2.1	any	192.168.10.[10,20]	80	New	deny

The problem occurs, for instance, when the two rules R1 and R2 are present in the firewall. In this case the client will be able to send the SYN packet to the server due to the rule R1. Once this initial SYN packet from the client is received by the firewall it make an entry in the state table and wait for the ACK packet from the server till timeout. But the ACK packet sent back by the server will be filtered out by the firewall rule R2. This will always prevent the establishment of the protocol and will create an overhead in the state table by unnecessarily occupying an entry. In this scenario either both the rules should deny or accept the packets.

[1] Eventhough, a great amount of work has been done for detection of anomalies in firewall configuration, majority of the methods have been limited to stateless cases c.f. [1,13]. Few approaches [18] involve description of stateful firewall models. However, most often than not, this involved straight forward adaptation of the management processes which were previously designed for stateless firewalls. The principle of the approach described here is similar to [17], and derives its origin from the specification of a general automata which describes the different states involved in traffic packages throughout the filtering processing. In our approach,we use defeasible reasoning technique to detect and resolve intra state anomaly in case of stateful firewall

3.2 Anomaly Removal Policies

Given the anomalies stated above, we adopt the following anomaly removal policies.

Policy 1. Shadowing Anomaly Removal*: If rule Y is shadowed by X. Swap between two rules to remove the shadowing anomaly without affecting firewall policy. From Table 1,we can find that Rules R4 is shadowed by Rule R2.To remove shadow anomaly,we should swap their poistion R2 and R4 in rule set If rule Y is shadowed by set of rules* $X_i, X_{i+1}, \ldots X_j$ *such that i<j, placing the rule Y above* X_i *removes shadowing anomaly. For instance from Table 1 , combination of R3 and R6 shadows R7,therefore to remove the anomaly we have to swap the position and rule set.*

Policy 2. Redundancy Anomaly Removal*: If rule Y is redundant with X. Depending on various conditions, following actions are required*

a. If rule X exactly match with rule Y, then we can remove rule X. Removal of rule will not affect the underlying security policy.
b. If rule Y is a subset of X, then security policy will not be affected by removal of Y.
c. If rule Y is redundant with set of rules X_i, X_{i+1}, $\ldots$ X_j *such that* $i < j$, *then remove Y.*
This may be elucidated by the following example from Table 1.If Rule R13 is a subset of Rule R3 and R6 then the rule set fire wall policy will remain unaffected if we remove Rule R13.Similarly R5 can be removed from rule set without effecting firewall policy as R5 is a subset of R12.

Policy 3. Correlated Anomaly Removal*: If rule X is correlated with Y, then we needed to split the overlapping portion of the rule X and Y to construct a new rule . For Example,to remove the correlated anomaly existing between R14 and R15 we split the rule into three different rules such that the overlapping portion is transformed into a new rule Rule new and the unintersecting part of the two rules R14 and R15 are termed as* $R14_{NEW}$ *and* $R15_{NEW}$ *respectively.*

Policy 4. Protocol Anomaly*: If there is a protocol anomaly between rule X and Y. It can be resolved by the following steps:*

a. If we find protocol anomaly rule pairs in firewall, then we compare the risk value of the destination hosts; preference will be given to the action associated with the destination with greater risk value.
b. If risk value of both the host are equal then priority of both the host are considered; preference will be given to the action associated with the destination with greater priority value.

For example in above scenario hosts A and B have priority values p_i and p_j respectively in such a way that $p_i < p_j$. In that case the action of the rule where B is the destination will replace the action of the rule where A is the destination.

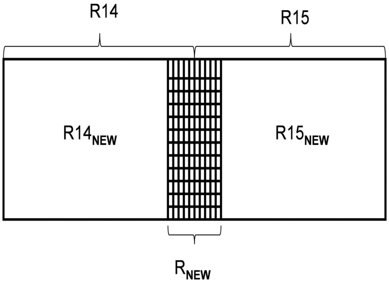


Fig. 2. Corelated Rule Spliting: R14$_{NEW}$; R15$_{NEW}$ and rule for overlap portion to replace R14 and R15

4 Using DeLP for Verification and Reconfiguration

4.1 DeLP Based Architecture

Figure 3 presents a framework for anomaly detection and removal using defeasible argumentation. A Rule Relation Analysis Engine (RRAE) determines the relation between the filtering fields of the rules to derive the ground facts. Anomaly detection and anomaly removal is through DeLP program specifically written for each. These DeLP program are interpreted over a DeLP Interpreter[2] A Rule Analysis and Reconfiguration Engine (RARE) queries the DeLP Interpreter.

Generating the Ground Facts. RRAE generates the facts either through pairwise interaction of fields of the rules OR through a nested operation when *correlated* rules exists. The basic firewall rule relation provision (Definition 10), specifies the set of ground literals for the defeasible logic program. This is for detection of simple anomalies.

Definition 10. *A* basic firewall rule relation provision *is a four tuple* $\langle \mathfrak{R}, \bar{\mathfrak{R}}, \delta, \Pi_{\mathrm{F}} \rangle$ *where,* $\mathfrak{R} = (R_1 \ldots R_n)$ *is the set of* n *rules in the firewall;* $\bar{\mathfrak{R}} = (R_{i+1} \ldots R_n)$ *for every* R_i *from* $\mathfrak{R}$*,* $i \geqq 1$*.* δ *maps set-theoretic relations between each field of* R_i *and* R_j*;* $i < j \leqq n$*;* $R_j \in \mathfrak{R}$ *(ground facts) to* Π_{F}*.*

For second order anomalies, we would generate the relation between a set of correlated rules and a third rule. This is stated by the following formal characterization of another firewall rule relation provision.

Definition 11. *A* secondary firewall rule relation provision *is a four tuple* $\langle \mathfrak{R}, \mathfrak{CR}, \mathfrak{P}, \delta, \Pi_{\mathrm{F}} \rangle$ *where,* $\mathfrak{R} = (R_1 \ldots R_n)$ *is the set of* n *rules in the firewall;* $\mathfrak{CR} = (CR_1 \ldots CR_{n-1})$*, where* CR_i *is the set of correlated rule pairs of* R_i *in*

[2] The DeLP Interpreter used here has been developed at LIDIA Universidad Nacional del Sur, Bahia Blanca, Argentina.

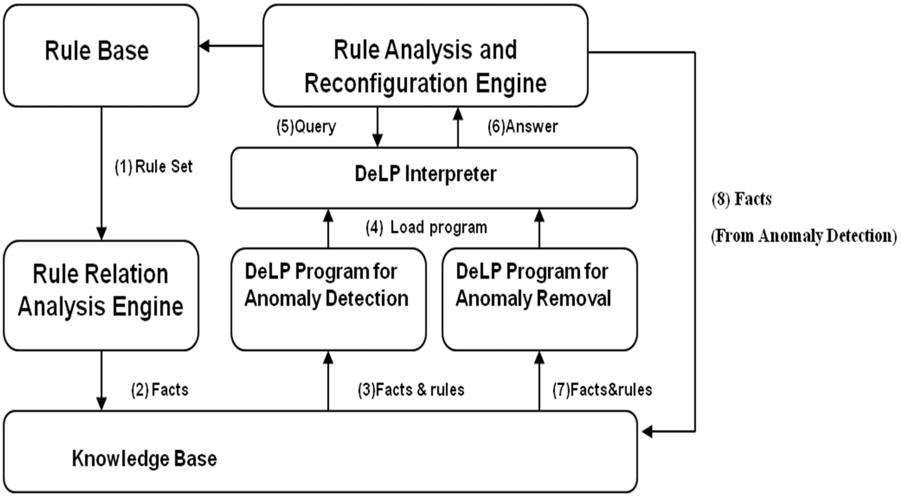


Fig. 3. Architecture of the DeLP Based Verification and Reconfiguration Framework

$\mathfrak{R}$ i.e., $CR_i = \{R_j \ldots R_m\}(j > i; m \leq n)$. $\mathfrak{P}$ is the set of linear combination of elements of each CR_i i.e, $\mathfrak{P} = \{R_i \cup R_j, R_i \cup R_j \cup R_{j+1}, \ldots, R_i \cup \cup R_j \ldots \cup R_m\}$. δ maps set-theoretic relations (ground facts) between each field of elements of $\mathfrak{P}$ and $R_k \in \mathfrak{R}$; where R_k is after R_m (i.e., $k \geq m+1$; $R_k, R_m \in \mathfrak{R}$.

Suppose rule Ri and Rj where i<j represent two rules in a firewall rule set. $\mathcal{P}$=(Π, △) represent the defeasible program in context of rule Ri and Rj where Π is the set of facts representing relation between corresponding fields. △ is the set of defeasible rules to detect anomaly between Ri and Rj. We perform $\mathcal{Q}$ = ⟨∼conflictfreerule(Ri)⟩ on $\mathcal{P}$. If answer for $\mathcal{Q}$ is ⟨yes⟩ then we run a query sequence $\mathcal{Q}$s =⟨ shadowing(Ri ,Rj), redundant(Ri,Rj), correlated (Ri,Rj), generalization (Ri,Rj), protocolanomaly(Ri,Rj) ⟩. To illustrate the working of the DeLP framework, we present below representative queries, one each for detection and reconfiguration. Without loss of generality, these are based on the example firewall in Table 1 and Table 2.

Example 1. For query ⟨∼conflictfreerule(R2)⟩; Argument $\mathcal{A}1$ is undefeated by $\mathcal{A}2$ and $\mathcal{A}3$. Argument $\mathcal{A}1$ is attacked by argument $\mathcal{A}2$ which is properly defeated by $\mathcal{A}$ 1. On the other hand $\mathcal{A}1$ is blocking defeater of $\mathcal{A}3$. As seen from the dialectical tree in Figure 4, answer for query ⟨∼conflictfreerule(R2)⟩ is yes. Answer for query ⟨shadowing(R2,R4)⟩ is 'yes'; argument $\mathcal{A}4$ defeat $\mathcal{A}5$ properly.

Table 3. Δ_1: Defeasible Rule Table for Verification

Rule
conflictfreerule(X)≺validrule(X)
∼conflictfreerule (X)—≺ redundant(X,Y),validrule(X)
∼conflictfreerule (X)—≺ shadowing(X,Y),validrule(X)
∼conflictfreerule (X)—≺ correlated(X,Y),validrule(X)
∼conflictfreerule (X)—≺ generalization(X,Y),validrule(X)
∼conflictfreerule (X)—≺ protocolanomaly(X,Y),validrule(X)
validrule(X) ← hasprotocol(X),hassourceip(X),hasdestinationip(X), hasdestinationport(X),hassourceport(X),hasaction(X)
redundant(X,Y)—≺equalprotocol(X,Y),subsetsrcip(X,Y),subsetdestip(X,Y), subsetsrcport(X,Y),subsetdestport(X,Y),equalaction(X,Y)
shadowing(X,Y) —≺ equalprotocol(X,Y),subsetsrcip(X,Y),subsetdestip(X,Y), subsetsrcport(X,Y),subsetdestport(X,Y),differentaction(X,Y)
∼shadowing(X,Y) —≺ correlated(X,Y)
correlated(X,Y)—≺ equalprotocol(X,Y),subsetsrcip(X,Y), subsetsrcport(X,Y),differentaction(X,Y)
correlated(X,Y)—≺ equalprotocol(X,Y),subsetdestip(X,Y), subsetdestport(X,Y),differentaction(X,Y)
correlated(X,Y)—≺equalprotocol(X,Y),subsetsrcip(X,Y),subsetsrcport(X,Y), subsetdestip(X,Y),differentaction(X,Y)
correlated(X,Y)—≺ equalprotocol(X,Y),subsetdestip(X,Y),subsetdestport(X,Y), subsetportip(X,Y),differentaction(X,Y)
generalization(X,Y)—≺ equalprotocol(X,Y),supersetsrcip(X,Y),supersetdestip(X,Y), supersetsrcport(X,Y),supersetdestport(X,Y)
protocolanomaly(X,Y)—≺ equalprotocol(X,Y),srcdestip(X,Y), destsrcip(X,Y),differentaction(X,Y)

$$\mathcal{A}1 = \left\{ \begin{array}{c} \sim conflictfreerule(R2) \mathrel{-\!\!\prec} shadowing(R2, R4), \\ validrule(R2) \\ shadowing(R2, R4) \mathrel{-\!\!\prec} equalprotocol(R2, R4), \\ subsetsrcip(R2, R4), subsetdestip(R2, R4), \\ subsetsrcport(R2, R4), subsetdestport(R2, R4), \\ differentaction(R2, R4) \end{array} \right\}$$

$$\mathcal{A}2 = \left\{ \begin{array}{c} \sim shadowing(R2, R4) \mathrel{-\!\!\prec} correlated(R2, R4) \\ correlated(R2, R4) \mathrel{-\!\!\prec} equalprotocol(R2, R4), \\ subsetsrcip(R2, R4), subsetsrcport(R2, R4) \\ differentaction(R2, R4) \end{array} \right\}$$

$$\mathcal{A}3 = \left\{ \begin{array}{c} conflictfreerule(R2) \prec validrule(R2) \\ validrule(R2) \mathrel{-\!\!\prec} hasprotocol(R2) \\ hassourceip(R2), hasdestinationip(R2), \\ hasdestinationport(R2), hassourceport(R2) \\ hasaction(R2) \end{array} \right\}$$

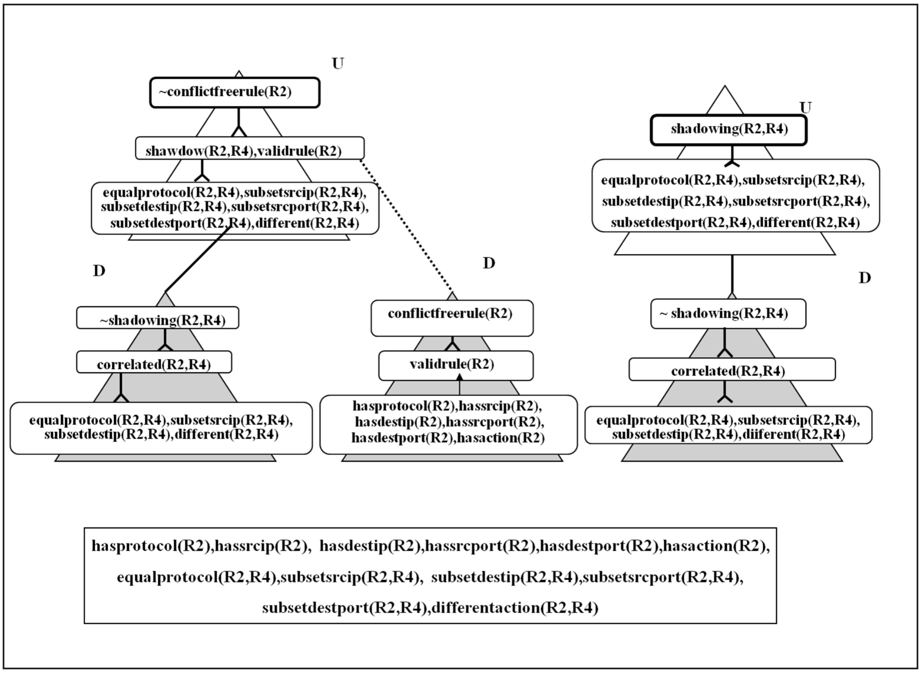

U
~conflictfreerule(R2)
shawdow(R2,R4),validrule(R2)
equalprotocol(R2,R4),subsetsrcip(R2,R4),
subsetdestip(R2,R4),subsetsrcport(R2,R4),
subsetdestport(R2,R4),different(R2,R4)
D
~shadowing(R2,R4)
correlated(R2,R4)
equalprotocol(R2,R4),subsetsrcip(R2,R4),
subsetdestip(R2,R4),different(R2,R4)
D
conflictfreerule(R2)
validrule(R2)
hasprotocol(R2),hassrcip(R2),
hasdestip(R2),hassrcport(R2),
hasdestport(R2),hasaction(R2)
U
shadowing(R2,R4)
equalprotocol(R2,R4),subsetsrcip(R2,R4),
subsetdestip(R2,R4),subsetsrcport(R2,R4),
subsetdestport(R2,R4),different(R2,R4)
D
~ shadowing(R2,R4)
correlated(R2,R4)
equalprotocol(R2,R4),subsetsrcip(R2,R4),
subsetdestip(R2,R4),diiferent(R2,R4)
hasprotocol(R2),hassrcip(R2), hasdestip(R2),hassrcport(R2),hasdestport(R2),hasaction(R2),
equalprotocol(R2,R4),subsetsrcip(R2,R4), subsetdestip(R2,R4),subsetsrcport(R2,R4),
subsetdestport(R2,R4),differentaction(R2,R4)

Fig. 4. DeLP dialectical tree on the left support the conclusion that rule R2 is not conflictfree; whereas the DeLP dialectical tree on the right support the conclusion 'shadowing(R2,R4)'.

$$\mathcal{A}4 = \left\{ \begin{array}{c} shadowing(R2, R4) \mathrel{-\!\!\prec} equalprotocol(R2, R4) \\ subsetsrcip(R2, R4), subsetdestip(R2, R4), \\ subsetsrcport(R2, R4), subsetdestport(R2, R4) \\ differentaction(R2, R4) \end{array} \right\}$$

$$\mathcal{A}5 = \left\{ \begin{array}{c} {\sim}shadowing(R2, R4) \mathrel{-\!\!\prec} correlated(R2, R4) \\ correlated(R2, R4) \mathrel{-\!\!\prec} equalprotocol(R2, R4), \\ subsetsrcip(R2, R4), subsetsrcport(R2, R4), \\ differentaction(R2, R4) \end{array} \right\}$$

4.2 DeLP for Anomaly Removal

The defeasible program $\mathcal{P}_2(\Pi, \triangle_2)$ is used to remove anomalies among rules in the firewall rule set. Whenever '∼conflictfreerule' is true, we have information about type of anomaly. Depending on the 'type' we have different reconfiguration policies which is reflected in the rule set in Table 4. Associated information with regards to 'type' of anomaly such as 'risk value' and 'priority' between rules is used here analogous to 'contextual query' of [14].

Table 4. Δ_2: Defeasible Rule Table for Reconfiguration

Rule
split (X)—≺ reconfigure(X),correlated(X,Y)
split (X)—≺ reconfigure(X),correlated(X,Y),groupofRule(X)
remove (X)—≺ reconfigure(X),redundant(X,Y)
∼remove (X)—≺ reconfigure(X),redundant(X,Y),subset(X,Y)
remove (Y)—≺ reconfigure(X),redundant(X,Y),subset(X,Y),∼remove (X)
remove (X)—≺ reconfigure(X),redundant(X,Y),subset(Y,X)
remove (X)—≺ reconfigure(X),redundant(X,Y),equal(X,Y)
∼remove (Y)—≺ reconfigure(X),redundant(X,Y),subset(X,Y)
∼remove (X)—≺ reconfigure(X),redundant(X,Y),equal(X,Y)
changeaction(X)—≺ reconfigure(X),protocolanomaly(X,Y)
∼changeaction(X)—≺ reconfigure(X),protocolanomaly(X,Y),riskvaluegreater(X,Y)
changeaction(Y)—≺ reconfigure(X),protocolanomaly(X,Y),riskvaluegreater(X,Y),∼changeaction(X)
changeaction(X)—≺ reconfigure(X),protocolanomaly(X,Y),riskvalueequal(X,Y)
changeaction(Y)—≺ reconfigure(X),protocolanomaly(X,Y), riskvalueequal(X,Y),prioritygreater(X,Y),∼changeaction(X)
reconfigure(X)—≺ ∼conflictfreerule (X)
∼reconfigure(X)—≺ ∼conflictfreerule (X),generalization(X,Y)

Example 2. Reconfiguration of rule R1 depends on whether it is involved in conflict with any other rules or the type of conflict it is involved in. For query <reconfiguration(R1)> answer is no as argument $\mathcal{A}6$ is properly defeated by argument $\mathcal{A}7$. See dialectical tree in Figure 5. Dialectical tree on the left support to conclude that eventhough '∼conflictfreerule' is true, in presence of 'generalization', '∼reconfigure' is warranted. Similarly dialectical tree on the right supports the conclusion that rule R12 need not be 'remove' eventhough 'redundant(R12,R5)' is true; whereas 'remove(R5) is warranted. Answer for query <remove(R5)> is yes as argument $\mathcal{A}8$ is defeated by $\mathcal{A}9$ which is defeated by $\mathcal{A}10$. Since defeater for $\mathcal{A}10$ is not present, $\mathcal{A}9$ is reinstated.

$$\mathcal{A}6 = \left\{ reconfigure(R1) \text{—≺} {\sim} conflictfreerule(R1) \right\}$$

$$\mathcal{A}7 = \left\{ \begin{array}{c} {\sim} reconfigure(R1) \text{—≺} {\sim} conflictfreerule(R1), \\ generalization(R1, R2) \\ generalization(R1, R2) \text{—≺} equalprotocol(R1, R2), \\ supersetsrcip(R1, R2), supersetdestip(R1, R2), \\ supersetsrcport(R1, R2), supersetdestport(R1, R2) \end{array} \right\}$$

$$\mathcal{A}8 = \left\{ \begin{array}{c} remove(R5) \text{—≺} reconfigure(R12), \\ redundant(R12, R5), subset(R12, R5), \\ {\sim} remove(R12) \\ reconfigure(R12) \text{—≺} {\sim} conflictfreerule(R12) \end{array} \right\}$$

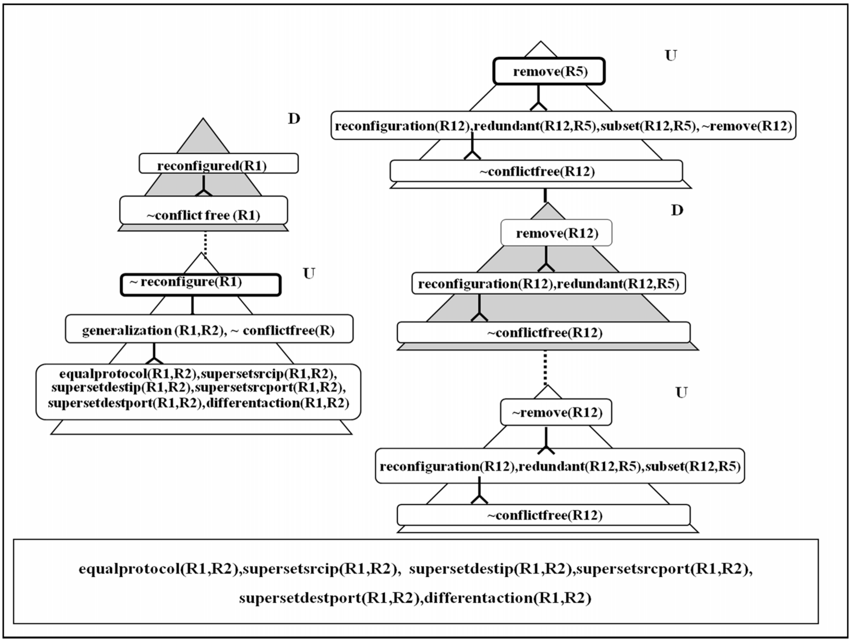

D
reconfigured(R1)
~conflict free (R1)
~ reconfigure(R1)
U
generalization (R1,R2), ~ conflictfree(R)
equalprotocol(R1,R2),supersetsrcip(R1,R2),
supersetdestip(R1,R2),supersetsrcport(R1,R2),
supersetdestport(R1,R2),differentaction(R1,R2)
remove(R5)
U
reconfiguration(R12),redundant(R12,R5),subset(R12,R5),~remove(R12)
~conflictfree(R12)
remove(R12)
D
reconfiguration(R12),redundant(R12,R5)
~conflictfree(R12)
~remove(R12)
U
reconfiguration(R12),redundant(R12,R5),subset(R12,R5)
~conflictfree(R12)
equalprotocol(R1,R2),supersetsrcip(R1,R2), supersetdestip(R1,R2),supersetsrcport(R1,R2),
supersetdestport(R1,R2),differentaction(R1,R2)

Fig. 5. DeLP dialectical tree on the left supporting the conclusion that rule R1 need not be 'reconfigure'. DeLP dialectical tree on the right showing 'remove(R5)' is warranted.

$$\mathcal{A}9 = \left\{ \begin{array}{c} remove(R12) \mathrel{-\!\!\prec} reconfigure(R12), \\ redundant(R12, R5) \\ reconfigure(R12) \mathrel{-\!\!\prec} {\sim}conflictfreerule(R12) \end{array} \right\}$$

$$\mathcal{A}10 = \left\{ \begin{array}{c} remove(R12) \mathrel{-\!\!\prec} reconfigure(R12), \\ redundant(R12, R5), subset(R12, R5) \\ reconfigure(R12) \mathrel{-\!\!\prec} {\sim}conflictfreerule(R12) \end{array} \right\}$$

5 Summary and Final Comments

We have discussed how a DeLP approach with an underlying argumentation based semantics could be applied for verification and reconfiguration of a firewall. We have demonstrated with illustrative examples that the DeLP based architecture is capable of performing firewall analysis. Firewall anomalies identified in the literature can be automatically detected. DeLP uses argumentation reasoning and exhibit rules that support a conclusion; leading to identification of source of anomalous configuration of the firewall. Further, under a given set of anomaly resolution policies, firewalls can be automatically reconfigured without violation of initial firewall policies. Defeasible argumentation in reconfiguration recommends the action regarding how to resolve conflicts by considering different facts associated with the rules. Table 5 shows the firewall of Table 1 after verification and reconfiguration.

Table 5. Firewall after verification and reconfiguration

Order	Protocol	Source IP	Source Port	Destination IP	Destination Port	Action
R1	TCP	150.172.37.20	any	*.*.*.*	80	deny
R4	TCP	150.172.37.*	any	171.120.32.40	80	deny
R7	TCP	150.172.37.[25,45]	any	171.120.32.[30,65]	21	deny
R3	TCP	150.172.37.[10,30]	any	171.120.32.[10,40]	21	accept
R2	TCP	150.172.37.*	any	*.*.*.*	80	accept
R6	TCP	150.172.37.[30,60]	any	171.120.32.[40,80]	21	accept
R8	TCP	*.*.*.*	any	*.*.*.*	any	deny
R9	UDP	150.172.37.*	any	171.120.32.40	53	accept
R10	UDP	*.*.*.*	any	171.120.32.40	53	accept
R11	UDP	*.*.*.*	any	*.*.*.*	any	deny
R12	TCP	150.172.37.[20,80]	any	*.*.*.*	any	deny
R14$_{NEW}$	TCP	192.168.37.[15,25]	any	171.120.32.[155,165]	21	accept
R15 $_{NEW}$	TCP	192.168.37.[40,60]	any	171.120.32.[120,150]	21	deny
R$_{NEW}$	TCP	192.168.37.[26,39]	any	171.120.32.[151,154]	21	deny

In context of defeasible argumentation, the work presented here follow the general idea of an argumentative framework as in [15]. As far as our knowledge this is the first adaptation of an argumentation framework for verification and reconfiguration of a firewall. This is worthwhile for the very fact that the main advantage of this approach is the ability of defeasible argumentation to deal with the exponentially increasing size of the firewall policy by reducing unnecessary checks. We have a prototype implementation. The implementation of the system was developed using primarily Java 1.6. This includes many new features of the language including enhance version of collection framework, performance, network interface and serialization which are extensively used in our implementation. Swing is used for development of user interface. Eclipse IDE is used for development of java programs. For defeasible argumentation an abstract machine called Justification Abstract Machine (JAM)[3] is used.

The present work is limited to detection and reconfiguration of intra-firewall anomalies. We envisage that a defeasible argumentative framework can be effectively used for verification and reconfiguration of inter-firewall anomalies, i.e., anomalies arising out of two or more firewalls operating together in a network. Such a framework could be along the lines of meta-level argumentation [6]. This is part of ongoing research.

[3] JAM was specially developed by LIDIA Universidad Nacional del Sur(Bahia Blanca,Argentina) for efficient implementation of DeLP. It is available online. (`http://lidia.cs.uns.edu.ar/delp_client`) [15].

Acknowledgment. This research is supported by DIT, Govt. of India through its project 'Design and Development of a Dynamic Firewall Solution' under Grant No. 13(1)/2011-IG.

References

1. Al-Shaer, E.S., Hamed, H.: Management and translation of filtering security policies. In: IEEE International Conference On Communications (ICC 2003) (2003)
2. Liu, A.X.: Formal Verification of Firewall Policies. In: Proceedings of the 2008 IEEE International Conference on Communications (ICC), Beijing, China (2008)
3. Govaerts, J., Bandara, A., Curran, K.: A formal logic approach to firewall packet filtering analysis and generation. Artificial Intelligence Review 29(3), 223–248 (2008)
4. Hazelhurst, S., Fatti, A., Henwood, A.: Binary decision diagram representations of firewall and router access lists. Technical report, Department of Computer Science, University of the Witwatersrand (1998)
5. Bandara, A.K., Kakas, A.C., Lupu, E.C., Russo, A.: Using argumentation logic for firewall configuration management. In: IFIP/IEEE International Symposium on Integrated Network Management, IM 2009, pp. 180–187. IEEE (2009)
6. Applebaum, A., Li, Z., Syed, A.R., Levitt, P.K.S., Rowe, J., Sklar, E.: Firewall configuration: An application of multiagent metalevel argumentation. In: Proceedings of the 9th Workshop on Argumentation in Multiagent Systems (2012)
7. Eronen, P., Zitting, J.: An expert system for analyzing firewall rules. In: Proc. of the 6th Nordic Workshop on Secure IT Systems, NordSec 2001 (2001)
8. Villemaire, R., Hall, S.: Strong Temporal, Weak Spatial Logic for Rule Based Filters. In: TIME 2009, pp. 115–121 (2009)
9. Khorchani, B., Villemaire, R., Hall, S.: Firewall anomaly detection with a model checker for visibility logic. In: NOMS 2012, pp. 466–469 (2012)
10. Hazarika, S.M.: Carving Rule-based Filters within a Spatio-temporal Logic. In: Proceedings of the National Workshop on Security 2010, pp. 30–35 (2010)
11. Thanasegaran, S., Yin, Y., Tateiwa, Y., Katayama, Y., Takahashi, N.: A topological approach to detect conflicts in firewall policies. In: IEEE International Parallel and Distributed Processing Symposium, pp. 1–7 (2009)
12. Christiansen, M., Emmanuel, F.: An MITDD based firewall using decision diagrams for packet filtering. Telecommun. Systems 27(2-4), 297–319 (2004)
13. Mayer, A., Wool, A., Ziskind, E.: Fang: A Firewall Analysis Engine. In: Proceedings of 21st IEEE Symposium on Security & Privacy, Oakland, CA (2000)
14. Tucat, M., Garcia, A.J., Simari, G.R.: Using Defeasible Logic Programming with Contextual Queries for Developing Recommender Servers. In: Proceedings of the AAAI Fall Symposium (2009)
15. Garca, A., Simari, G.: Defeasible Logic Programming: An Argumentative Approach. Theory and Practice of Logic Programming 4(1), 95–138 (2004)
16. Garcia-Alfaro, J., Boulahia-Cuppens, N., Cuppens, F.: Complete analysis of configuration rules to guarantee reliable network security policies. International Journal of Information Security, 1615–5262
17. Cuppens, F., Cuppens-Boulahia, N., Garcia-Alfaro, J., Moataz, T., Rimasson, X.: Handling Stateful Firewall Anomalies. In: Gritzalis, D., Furnell, S., Theoharidou, M. (eds.) SEC 2012. IFIP AICT, vol. 376, pp. 174–186. Springer, Heidelberg (2012)
18. Gouda, M., Liu, A.: A model of stateful firewalls and its properties. In: DSN, Yokohama, Japan, pp. 128–137 (2005)

Threats and Challenges to Security of Electronic Health Records

Shalini Bhartiya[1] and Deepti Mehrotra[2]

[1] IITM (GGSIPU), New Delhi, India
[2] ASCS, NOIDA, India
shalinibhartiya69@gmail.com, dmehrotra@amity.edu

Abstract. Healthcare has always been a sensitive and a complex process. Rapid strides have been made both in the field of information technology as well as health care successfully integrating both for better facilities and services offered by the health-givers. Electronic health records (EHRs) is the product of this integration and forms an integral part of the automated healthcare system. Accessing of EHR by each stakeholder complements the issues of data disclosure, confidentiality, authenticity and privacy that are likely to occur due to many reasons. This paper aims at studying and identifying security threats to EHR in the hospital information system currently prevailing in the hospitals (HIS). It further categorizes the threats based on security characteristics and rates them on the basis of impact and magnitude of loss to the patients. The paper highlights real-time scenarios with each as an important requirement of the health-givers on one hand, can also be a reason of security breaches on other hand. It concludes by listing challenges and recommendations to curb security threats commonly found in the physical setup of healthcare environment.

Keywords: Challenges to EHR, EHR, HIS, Security, Threats.

1 Introduction

The health industry is undergoing a fast transition from its conventional method of care-giving. E-health or Health Informatics is an ICT-integrated method adopted by the hospitals for providing healthcare services to the patients anytime, anywhere without any restriction of location or facility. Hospital Information System (HIS) is a customized and tailor-made application facilitating the care-providers requirements through integration of components from IT as well as medical domain. HIS accumulates entire details of patient's health in form of electronic health record (EHR).

HIS includes various processes including registration, billing, admission, prescription, investigation registration and reports, discharge summary, appointments, medico-legal requirements, to list a few. Each process is handled individually by well-organized schemas called departments in the hospital that are constantly communicating and sharing the health information for delivering the facilities to the patient. The transfer of health information between departments electronically, opens a new door for theft and disclosure of sensitive and confidential data to unauthorized and unidentified users.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 543–559, 2013.

Though data is crucial for the productivity of any business, the varsity of ensuring confidentiality and privacy of health data outcasts other industry needs.

Confidentiality is a form of informational privacy characteristic of certain relationships, such as the physician-patient relationship. Personal information obtained in the course of that relationship should not be revealed to others unless the patient is made aware of this intention and consents to disclosure [9].With all the security systems like firewalls, Intrusion Detection Systems (IDS), anti-virus software, encryption/decryption techniques and role-based access grants, there is enough room for security exploits to EHR. Currently deployed hospital information systems reveals many grey areas that accumulate to security threats and data breach or disclosure of health data and largely by the insiders.

The prime objective—'providing health care to the patients' directs the users to surpass the security rules imposed in the application. Here, in this paper, we have discussed the scenarios where the confidentiality and privacy of EHR is deliberately compromised. We are strictly limiting the discussion to security aspect of storage and retrieval of health data by the care-providers i.e. doctors and paramedical staff and also the patient.

This paper is divided into nine sections. Section I is an introduction to Health Information System (HIS) and the security concern of electronic health records in electronic setup. Section II reviews the work done by other researchers to gain better insight of security issues in health informatics. Section III details the research methodology used to collect and analyze the findings related to current practices and security threats in the real working environment of the hospital. Section IV tabulates the conveniences and constraints of health informatics as described by clinicians. Section V enlists the varied security threats and probable reasons of their occurrence and Section VI elaborates the real-time scenarios that require concentration and focus to develop more specific and suitable security solutions. Section VII pens down the possibilities of stress and challenges while accessing EHRs. Section VIII states the suggestions to overcome the challenges to security of EHRs in various scenarios. Finally, Section IX wraps up emphasizing the importance of health informatics and the need to develop suitable security solutions in this area.

2 Literature Review

The integrity, availability and confidentiality [7] of digitized health data becomes the highest priority for developers as well as end-users. Profound study on the security aspects of electronic health records (EHR) world-wide, significantly states the demand and acceptance [1] by the healthcare industry in transforming its services into digital-based environment. The confidentiality and privacy of sensitive health data of patients moves out from the physical lock and key security to uniform and standardized bits and bytes structure. EHRs do deliver greater mobility and accessibility to information [11], but can also increase security risks if the appropriate precautions are not taken. Unlike paper records, access to EHRs can be restricted, so staffs have access to records based on job function. Specific to protecting the information stored in EHRs, the HIPAA [6] Security Rule requires that health care providers set up physical, administrative, and technical safeguards to protect electronic health information. NIST [14] has developed guidelines to facilitate application of appropriate levels of information security according to kind of impact or consequences that might result from the unauthorized disclosure, modification, or use of the information or information system. The guidelines are based on review of categorization of security terms and definitions by FIPS 199. World Health Organization [12] has given guidelines about how to review the current health record systems before discussing issues and challenges to the security of EHR. As emphasized by [2], protecting the health information from unauthorized access and malicious attacks would enhance confidence and acceptance among its stakeholders. Effort is required as stated by [5] to incorporate compliance like HIPAA security guidance with own security policies in standardization of HIS at the national and international level.

The table highlights the findings of few surveys [13][15][16] conducted with an objective of identifying security issues and solutions adopted by security managers in healthcare and other industries.

Table 1. Review of Surveys conducted on Security Issues in Healthcare

Survey	Objective	Finding
CSI3 Computer Crime and Security Survey 2009	To identify the areas of data theft and breaches in various industrial setups including healthcare	60% of total financial losses resulted from "malicious insider" attacks
2011 HIMSS Security Survey, November 2011	To identify key issues surrounding the tools and policies in place to secure electronic patient data at these healthcare organizations. An important component explored in this study is the issue of risk analysis.	Three-quarter of healthcare organizations are doing risk assessment and using it to determine which security controls should be put into place. Survey revealed that there is no robust plan for securely sharing the information outside of their organizations.
CSI Computer Crime and Security Survey, 2010-2011. Survey included security Practitioners from various sectors	To determine not only what security technologies respondents used, but additionally how satisfied they are with those technologies.	Malware attacks are the most commonly seen attack. Managers are looking security solutions in virtualization, cloud computing and exploring better visibility into security status of the networks
SearchHealthIT.com, April 2011	Are hospitals meeting federal laws like HIPPA for data protection and what security policies and measures they think to implement in near future for security of healthcare data	The respondents focused on encryption and tighter access controls for HIPPA privacy compliance in networked environment. Single sign-on (SSO) is also being acknowledged as a strong authentication method in integrated EHR system

The security issues perceived by an IT manager and hospital manager do not fall in the same domain of thoughts. In addition to the security characteristics[3] as prominently talked about in Information Technology, security principles in healthcare environment include the ethics, right to information, legal rights, laws and legal implication for both—care-provider and the patient. Integration of ICT and healthcare faces tough challenges associated with its adoption especially in developing countries as illustrated by [5] taking the example of Tanzania. The conditions can be evidently analyzed in Indian perspective [8] due to the similar financial, educational and social conditions. The cross case effect matrix generated by [4] identifies the basic functionalities and its effects on security of health records. The Department of Information Technology (DIT) and Ministry of Communication & Information Technology (MCIT) have prepared a framework [10] for Information Technology Infrastructure for Health (ITIH) in India with the support of Apollo Health Street Limited (AHSL) to build a secured and robust digitized healthcare environment. Framing of strong privacy law, legal regulations, policies and acceptance of digital records as legal evidences in the court can foster an optimistic absorption by healthcare professionals.

3 Research Methodology

It is a questionnaire based study conducted at public sector, government controlled chain of hospitals under labour ministry. There are about 2026 integrated sites through centralized data center controlled by IT vendor and its team. The study was made at one of these centers to observe the usage of application and gather the opinion of clinicians and paramedical staff with an objective of identifying the security applied to health records and also where the data is vulnerable to security threats and breaches. The study covered all the areas of services i.e. OPDs, IPDs, Day-care, Casualty, OTs, ICUs and IT control room, Laboratories and Pharmacies. Stratified sampling method was used to collect the views selecting Professors, Associate Professors, Senior/Junior Residents and Interns from each department. Besides, other prime clinicians were approached that include nodal officers, operational heads, medical superintendents, recruitment heads, anesthetists etc. The feedback of 200 clinicians was collected through well-structured questionnaire distributed to the clinicians. The questionnaire focused on questions related with data availability, entry-points in the system and possible vulnerabilities, authentication controls, ease of use, increase in efficiency of health-providers, etc. One set of questionnaire was also uploaded on Surveymonkey.com for the purpose. Besides questionnaire, informal interviews and meetings were used to gather relevant information.

4 Clinician's Perspective

The clinicians acknowledged the long and short-term benefits of having E-health Records. They strongly agree that electronic health data can substantially reduce the risk of data breach but found the recording methods and interfaces as time-consuming, inappropriate and substandard. As an instance, a senior doctor selects a team of junior doctors for treating or operating a patient on case-basis, requiring sharable access of case-sheet to all the members of his/her team. The team might remain same or have modifications depending on case-to-case. This results in a multiplicative distribution of rights and privileges to the doctors as well as paramedical staff that they continuously hold even after the case is closed. Here, we present clinicians views based on their hands-on experiences of using HIS. Their conveniences and constraints with respect to usability, availability and security of EHR are represented in tabular form below.

Table 2. Clinician's Conveniences and Constraints in using Hospital Information Systems for storing and accessing EHRs

FEATURES	CONVENIENCES	CONSTRAINTS
Password-Based or Role–based Access to Data	➢ Guarantee access to only legal users ➢ Enhances trust and belief in the system ➢ Helps department heads in tracking and analyzing efficiency and errors on individual basis	➢ OPDs are handled by team of doctors, sharing the same computer disallowing simultaneous update of records in the database. ➢ Doctors on adhoc appointments do not have separate login Ids and are forced to share the Ids of their colleagues or HODs. ➢ Cannot control the disclosure of information between doctors
Ease of Use	➢ Patient details are readily available ➢ Easy maintenance of health details and reports	➢ Requires use of multiple keys and opening of various windows, increasing the effort and time of data-entry ➢ The interfaces are not satisfactory ➢ Takes a lot of time to retrieve data ➢ Improper training takes more time to understand the working of HIS
Efficiency	➢ Enables correlation with previous history and observations ➢ Past records and history of patient helps in better diagnosis	➢ Cannot be used during heavy flow of patients mainly in OPD. ➢ Criticalities like Casualties and Emergencies cannot be dealt with the HIS alone. ➢ Time consuming data entry methods ➢ Slow processing adds to delay in work

Table 2. (*Continued*)

FEATURES	CONVENIENCES	CONSTRAINTS
IT Skills and Knowledge	➢ Enables independently handling of processes and various devices like scanner, printer etc.	➢ Requires practice and training to gain better grasp on the usage of the system ➢ Recognizable constraint among doctors in the age group of 50 and above. ➢ Need the help of trusted colleagues or paramedical staff to record entries
Availability of Data	➢ Investigation and Prescription details are available in respective departments. ➢ Record can be accesses from any location ➢ Health Data of a patient can be accessed from anywhere without any location constraint.	➢ Complete details of the patient is not always available ➢ Referral patient's data is not available ➢ Investigations done outside the hospital are never recorded. ➢ Typographical errors and data mismatch needs manual verification
Security	➢ Cannot access the application without valid login Ids and passwords ➢ Access to only relevant processes ➢ Restricted and Controlled Internet usage. ➢ Suppress all e-mail services except for intranet mail-box.	➢ Uncontrolled printing of documents and reports. No logs available to identify the number of prints taken for any record. ➢ No control and track on patient's information is being sent outside the hospital. ➢ Sharing of computer or records hides the individual identity of the user. ➢ No monitoring and restrictions on modification of health data.

The figure illustrates the conveniences and constraints of the doctors using the HIS for providing healthcare services. Relative frequency percentage was calculated for each parameter and enabled to identify usability, availability and Security of health data along with improving efficiency and requiring IT skills for using the application. It was observed that 21% of respondents found password-based access as satisfactory. 20% found acquiring skills to use the application as an overhead explaining that the prime objective is providing healthcare to the patients.

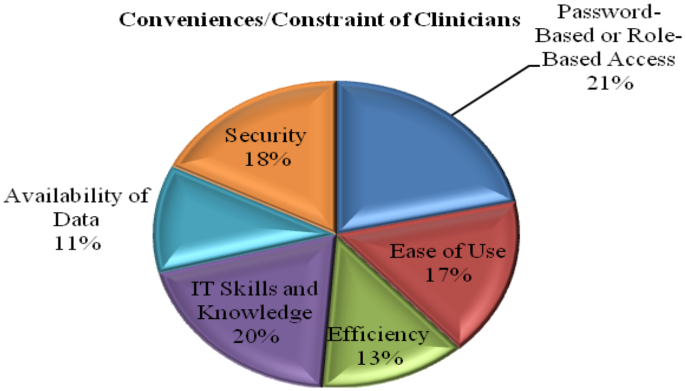


Fig. 1. Shows the relative frequency of clinician's experiences of using HIS

5 Prevalent Security Threats

As technology continues to play an increasingly important role in health care, the national movement towards electronic health record systems is leading to many improvements in the quality of patient care. Yet, as with paper record systems, there are risks. Healthcare Information systems (HIS) are exposed to numerous threats that can result in significant loss and damage to medical records. The survey identified numerous confidentiality and privacy vulnerabilities prevalent in the system with some due to ignorance of its stakeholders and others due to the current practices of healthcare.

Password Sharing: Password sharing poses threat to data integrity and repudiation of EHR. Data confidentiality and privacy is at stake as the study showed that more than 80% clinicians share their passwords with fellow colleagues.

Typographical Errors: This is a threat to accuracy and correctness of EHRs. Healthcare industry is highly error-conscious and accuracy is of prime concern here. Typographical errors are common threat mainly due to lack of technical skills found in data-entry operators.

Physical Security: Unauthorized penetration into the hospital vicinity is an additional physical threat to the availability and security hospital infrastructure including devices, equipments and medical records. Inappropriate and insufficient of physical security exposes health data to illegal and unauthorized disclosure and malfunctioning.

Storage of Backups: Storage of the backups is more crucial that taking backup. The security of stored data, especially backup data, has received less attention. A policy-based multi-backup data on a regular schedule, storage of backups off site at multiple locations and ensuring that backup media is physically secured needs to be ascertained to protect the sensitive health records.

Timely-Availability of Data: The non-availability of health records during system upgrades or power failure is not accepted by the clinicians. The healthcare needs cannot afford any downtime whatever minimum it may be.

Role-Based Access: Users in health informatics system get to access only relevant details as per their working profile in the hospital but changes to default access roles are frequently demanded due to shifting duties mostly by paramedical staff. The privileges once granted are never reverted.

Threat from Former Employees: The active IDs of former employees are another threat that provides the opportunity to disgruntled or terminated employees to perpetrate into the system. Restraining orders, password changes, and other special security measures are necessary in some situations.

Table 3. Frequency of Common Security Threats to EHRs as observed in working environment of health care

THREATS	FREQUENCY	PERCENTAGE
Sharing Passwords	177	89
Typographical Errors	104	52
Physical Security	151	76
Storage of Backups	77	39
Timely-Availability	126	63
Deficiencies in Role-Based Access Controls	138	69
Former Employees	81	41
Personal Details of Patients	179	90
Password Policies	185	93
IT Privileges	140	70

Personal Demographics of Patients-A patient becomes a victim of identity theft because his/her demographic or personal details are accessible and modifiable from any workstation.

Password Policies-Authentication through valid login Ids and passwords is the most common choice of any IT administrator. To strengthen the authentication controls, health informatics must incorporate robust password policies that demand changing of passwords periodically and encourage users to choose strong passwords.

IT Privileges-IT people hold absolute rights and privileges to manage and control the processes of health informatics. The clinicians viewed it as a threat to confidentiality and privacy to health data.

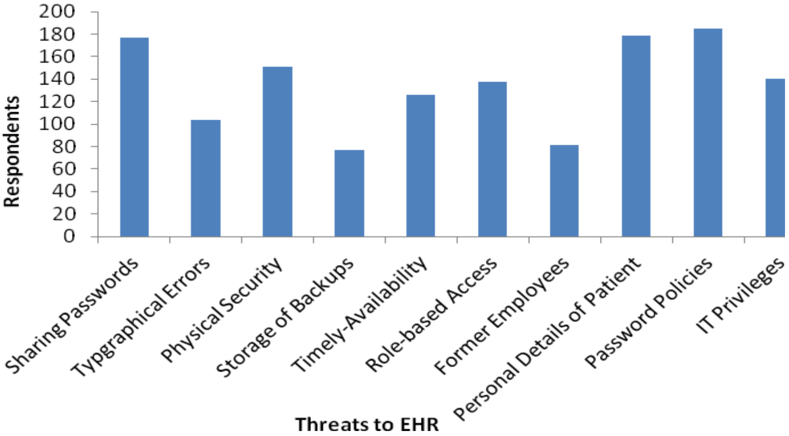


Fig. 2. Frequency of Security threats and Vulnerabilities to health data as perceived by the Clinicians

Table 3 lists the frequencies of security threats as perceived by the clinicians. Figure 2 graphically represents the clinician's perception about each of the threats exposing data to numerous security breaches in healthcare environment. The impact of weak password policies, exposed demographic details of the patient and sharing of passwords is high whereas typographical mistakes, security of storage of backups and former employees are considered as having low impact on the security of health records.

Other Factors

Training and Education—Lacking

Hospital management arranges and schedule training module for their employees once a while but it is not a continuous process. The staff demands more adequate training and hands on practice sessions to be conducted to be able to efficiently operate the HIS. Observations found the staff comfortable using their regular part of the modules or activities but lacked knowledge to operate other parts also under their control.

Compliance

Standards like HIPAA, HL7, and DICOM were either not known to majority of the staff people in the hospital or was not given much weight age and consideration for the compulsory inclusion of these standards in the system. Patients usually have no or restricted access to their health information where they can view their investigation reports and test details or book an appointment with the doctor prior coming to the hospital. The patient has no access to complete diagnostic and history of his/her case.

Patients are not aware of their rights regarding the authority of their health records and obligations of the hospital relating to storing and securing their confidential health information.

Security in Place
Role based access privileges are given to each user and the department according to their requirements. Firewalls with full scan and Proxy servers are installed to monitor and record all the network traffic. The internet access is available to only few privilege users of the hospitals like HODs, MS, Additional MS, and Nodal Officers. Other health-providers requiring internet access is given controlled and restricted access completely monitored and recorded. Logs of each transaction are maintained along with the details of the employee logged in at that point of time. Every transaction made by the user is directly recorded at the centralized data-centre. Patches and updates are incorporated whenever required. The system works in closed network where no outside access is provided to the users. The system incorporates a private mailing process whose use is restricted to high profile users only or some paramedical staff members. Multiple servers supporting every operating system are installed. Hardware Firewall and proxy servers are installed enabling security checks at each network points for any type of data communications and transmissions.

6 Current Practices

Health information is characterized by high complexity and heterogeneity in both the nature and the sensitivity levels of the different data sets included in it. Sensitive health information such as HIV status, obstetrics history and mental health history could become more easily accessible as health records become fully automated. If sensitive health information is accessible by others, this would clearly represent a breach of the patient's privacy.

Security is generally defined as the extent to which personal information can be stored and transmitted in such a manner that access to the information is limited to authorized parties. The threats described in the previous section are very common, found in any business environment. But the slightest possibility of their occurrence in health industry can cause a high magnitude of loss to exclusively patients and also the hospital. The question arises-"What forces the clinicians to compromise the security of EHRs in health informatics system? "

The answer lies in the current practices of care-giving. The weak integration and communication between departments exposes patient's identity and health data forfeiting security mechanisms and policies altogether. A deep thought and attention needs to be given to the following scenarios that justifies that abiding by the security rules and policies are out of control of doctors and hospital administrators.

Real-Time Scenarios

E-health security and privacy are challenging not only due to the difficulties of developing an error-free, complex framework, but also because of the complications of various issues addressed by all the stakeholders in the E-health industry such as the patients, vendors, and health providers.

Our observation identified that clinicians want to follow procedures of HIS but various limitations and demands of the environment of health care forces them to compromise with the confidentiality and privacy of EHR. Here, we present a few of real-time scenarios that act as a barrier to security of electronic health data of the patients.

- Frequent shift of duties and departments is very common among the doctors and paramedical staff. This forces the IT admin to modify the default access control lists (ACLs) as per the demands of a particular facility and services. The ACLs once modified remain in that state till the administrator receives requests from hospital-in-charge.
- Hospital Laboratories are not fully equipped to cater to every need of the patient. Hence, many tests are outsourced exposing the patient's sensitive health data to outside world.
- Investigation devices are not programmed to transfer images and findings to the health informatics system. The diagnostic details have to be typed either by the clinicians or support staff thereby surpassing non-repudiation and authentication controls.
- OPDs have a heavy flow of patient's every day. The HIS take about 5 to 10 minutes to view, add and update each patient's information. It becomes impossible to manage the waiting queue outside the doctor's room. Moreover, the systems are slow and hang up quite frequently.
- The wards and rooms do not have computers at the bed-side of patient, except for in ICUs. Moreover, the doctors on round can't afford the time taken to display the progress of each patient by HIS.
- The hospital administration policies permit the staff to take prints of confidential and health related documents any number of times. No written entries were found in the logs that could identify the number of print-outs taken by an employee at any given point of time.
- Emergencies and Accidental cases require immediate action. The doctor-on-duty has to provide the utmost care and treatment to the injured. Such details sometimes are never recorded due to non-availability of electronic details of the concerned patient or are forgotten over the period of time to be recorded into the system.

Impact of Real Time Scenarios on Security of EHR

Breach to EHR causes devastating damage to the healthcare industry's reputation and is a serious detriment to the goal of building a patient-centric paperless model to be used by the care-providers without any limitations or constraints. Here, we categorize the activities and services as observed in the current practices into type of security threat and further classify the impact of that threat on the security of EHR.

Table 4. Impact of Current Practices on Security Features in Health Informatics

Scenarios	Security Threat	Severity
Frequent Shift of Duties	Data Disclosure	Low
Outsourcing of Investigations	Confidentiality Loss	High
Handling Emergency and Accidental Cases	Data Integrity Loss	Medium
Non-Integration between Medical equipments and HIS	Data Availability	Medium
Reluctance of Doctors	Data Integrity Loss	High
Handling OPDs	Data Integrity Loss	High
Resource Sharing	Non-Repudiation	Medium
Ignorance of Patients	Identity and Confidentiality Loss	High

Table 4 shows the type of security threat and its severity level in each of the scenarios rampant in health informatics. The analysis brings out two important features that need to be prioritized while planning and designing any HIS. The reluctance of doctors and outsourcing of investigations are ranked with high level of severity. Reluctance is a mindset which can be changed with user-friendly interfaces and training. Outsourcing of investigations cannot be eliminated as no hospital can be 100% equipped with all the machines and devices required for varied laboratory tests, therefore, needs automatic integration with the hospital information system. Thus, as much organized the application is, without a rigorous security plan, the main objective of developing E-healthcare environment would forfeit.

7 Challenges to EHR

The logics behind the accessibility of data, as observed during the survey, were found to reflect a static behavior i.e. the rules and policies uniformly applied to every component and process of HIS. To be effective, security must be multi-layered and if it fails, measures must be in place to safeguard the health data from access.

Here, we address certain issues that openly challenge the security and usability of health informatics system.

Interoperability

Interoperability addresses issues of how to best facilitate the coding, transmission and use of meaning across seamless health services, between providers and patients. Its geographic scope ranges from local interoperability (within, e.g., hospitals or hospital networks) to regional, national and cross-border interoperability. It is required across heterogeneous EHR systems in order to gain the benefits of computerized support for decision making, workflow management and evidence based healthcare. The information captured in EHR systems provides easily retrievable data via networks and creates significant risks representing unique security challenges.

Data Availability

With little and inadequate medical knowledge, patients can misinterpret the investigation results and diagnosis. Entire information available on internet gives rise to cyberchondria, i.e. self-diagnosis by the patient without consulting the physician. The doctor might find it difficult to satisfy the patient's queries when every technical term is exposed to the patient. It might result in doctor going into restless or indifferent state due to panic and disturbances created by the patient. The display of entire health reports can cause mental and social trauma to the patient as every person has a different level of handling stress. This can result in image defamation, loss of job, emotional distress, family neglect, social stigma blocking every path to recovery.

Data Confidentiality and Privacy

The EHR is vulnerable to threats from various secondary stakeholders of the system as identified during the survey. Every single unit of data passes through networks established and controlled by the IT team giving them implicit privileges to peep into the confidential and sensitive health information of the patients. The health data accessed by the employers as per the policies of the organization can be used against the benefits of the employee. Insurance companies can deny claims and benefits to the insurer. The friends and family of the patient can easily access the investigation and other details even without his/her consent.

Internet

Internet is a source that binds the world in a close-knitted family where everyone can interact and share their thoughts and intelligence. Doctors believe that internet is a useful resource of sharing health data with the desired stakeholders. It would result in reduction of cost and time in making the information available to the people. It enables location-independent facility of not only availability of health data but also rendering of services by doctors. E-mail is one such facility that enables communication and transfer of important documents between the two parties. A typographical mistake in sender's address can deliver the reports to unintended person. E-mail spoofing is another threat and a challenge to security of EHR. Bulk-mail can result in non-delivery of some mails that are never checked or send again unless being explicitly demanded by the patient. Patients never bother to update E-mail accounts when changed, with the hospital. Not all patients have their own E-mail Ids or Internet access. Uncontrolled and unmonitored dependency on service providers for E-mail services questions the reliability and security of EHR.

Policies and Standards

How can a doctor take the consent of a patient in unconscious state before treating him/her? This is an important question that challenges the HIPAA rule of taking patient's consent before accessing his/her medical records. Moreover, if we talk about in Indian perspective, as per the Census of India's 2011, of the 121 crore Indians, 83.3

crore (68.84%) live in rural areas that still lack technological resources. Similar diversities can be acknowledged in other parts of the globe. Hence, the standards like HIPAA, DICOM needs to be restructured to reflect and accommodate these diversities. Security of EHR is challenged due to lack of uniform, multipurpose data standards that meet the needs of the diverse groups who record and use health information. There is a lack of policies and legislations to protect privacy while permitting critical analytic uses of health data. Serious intervention of government is required in formulating policies and laws for health informatics.

IT Team

The IT team enjoys full rights and privileges on every resource and data of the health informatics system. Configuration errors might lead to hacking or other attacks on health data. Ignorance of IT companies while developing HIS is a major challenge that drags users away from using the HIS.

8 Recommendations

The study and analysis discussed above reflect the security threats and challenges prevailing in a closed network of health departments in hospitals that are located in different cities. Threats to confidentiality and privacy of data are more from insiders in such an environment as the outsiders have negligible and controlled role to access these records. The stakeholders demand an efficient system that could enable the availability of accurate health record information that is truly integrated with all the databases. The system should be able to recognize the user on per-transaction basis that can be held accountable for the acts committed by him/her if any discrepancy or inconsistency is observed. The security controls and models require refinement where the following factors should also be taken into consideration:

- A vigorous plan for physical security of the hospitals units and departments especially at night would reduce the possibilities of theft of physical devices and documents.
- Providing controlled and limited powers and privileges to IT vendors responsible for developing and maintaining the hospital management systems and its databases.
- Developing and implementing dynamic access controls that could allocate and monitor the authorization of each stakeholder from time to time or as per the demand and need.
- An inbuilt or implicit check should be imposed on various internal processes of the application, such as, keeping a count of printouts taken, use of external storage devices, use of internal commands and shortcuts, accessing the records of deceased patients etc.
- Primary users of healthcare are doctors. The system generally allocates identical authorization privileges considering them at the same hierarchy. The authorization and access controls should be assigned based on qualification, experience, designation of the doctor on one hand and the

justified request on the other hand. The request can be justified by supporting the application with decision-support system that can intelligently view and analyze the request based on the attributes provided in the query.

- Security controls complemented with accountability ensures better trust and confidence among stakeholders of the system. Hence, measures and policies must be designed and framed that would strengthen non-repudiation in sharable and interoperable healthcare environment.
- The role of a doctor as well as paramedical staff administer frequent shift of duties in their everyday working. The RBACs needs to be frequently altered to accommodate their data needs. Therefore, strict actions are required that could dynamically allocate and also de-allocate the privileges and permissions on various areas of the application, probably bounded by a time-frame.
- There is huge diversification of location, education, population, funds and resources not only around the globe but also within the nation. It is required to design standards like HIPAA, HITECH Act, HL7 and DICOM around these parameters so that security controls on EHRs can be imposed irrespective of the heterogeneity and interoperability of health data stored in such diversified environments.

9 Conclusion

The study revealed many facts related to the usability, availability and security of electronic health records accessed by the care-providers in their daily practices. The hospitals are well equipped with full-fledged HIS but most of the clinicians were found reluctant to use it in their daily practices due to various reasons. Due to low and partial usability, the transformation of health records to electronic form still remains incomplete and the hospitals maintain and use the paper-based records in parallel with the EHRs. Huge rush of patients in OPDs every day (Average: 2000 patients), it is difficult to manage and provide healthcare services totally depending on the HIS. The system generally slows down at such times resulting in non-availability of patient’s information during treatment. Additionally, the IPDs have computers placed only at the nurse’s workstation, forcing the doctors to refer to case-sheets while taking rounds in the wards and rooms of the patients. The frequent shift of duties of paramedical staff forces the IT manager to assign added privileges for them, exposing the sensitive patient’s data to a larger hierarchy of users.

Security, integrity and privacy of personal medical data is of utmost importance, and whilst many research projects worldwide are investigating the application of new technologies to E-healthcare solutions, security and reliability of these technologies is an area that requires further exploration. This paper highlights security threats and their impact on the day-to-day working of health care environment in a closed network. The collection and analysis of health information demonstrated real-time scenarios that require deep and futuristically planned security mechanism or a model capable enough to enhance the confidence and trust of physicians on health informatics. The health care systems are relatively to protective measures to keep the security and privacy of their patients and their health details. Business process redesign and an understanding of the change management process are fundamental to this activity. Healthcare organizations need to analyze and assess usages of EHRs with different perspectives and reflect such perceptions in proposals, system selection, development, installation, and implementation of health informatics system in order to ensure that all needs of the organization are met.

References

1. Fisher, S.R., Creusat, J.-P., McNamara, D.A.: 2008 McKesson Corporation, Improving Physician Adoption of CPOE Systems, http://www.strategiestoperform.com/volume3_issue2/docs/ImprovingPhysicianAdoption.pdf
2. Lin, S.-C., Tsai, W.-H., Tseng, S.-S., Tzeng, W.-G., Yuan, S.-M.: A framework of high confidence e-healthcare information system. In: International Conference WWW/Internet 2003 (2003)
3. Alanazi, H.O., Jalab, H.A., Alam, G.M., Zaidan, B.B., Zaidan, A.A.: Security characteristics: Securing electronic medical records transmissions over unsecured communications: An overview for better medical governance. Journal of Medicinal Plants Research 4(19), 2059–2074 (2010)
4. Fernando, J.: Jabberwocky, The Nonsense of Clinician Ehealth Security. International Journal of Digital Society (IJDS) 1(3) (September 2010)
5. Omary, Z., Lupiana, D., Mtenzi, F., Wu, B.: School of Computing, Dublin Institute of Technology, Tanzania Case, Analysis of the Challenges Affecting E-healthcare Adoption in Developing Countries: A Case of Tanzania. International Journal of Information Studies 2(1) (2010)
6. HIPAA Compliance Review Analysis and Summary of Results, Centers for Medicare & Medicaid Services (CMS) Office of E-Health Standards and Services (OESS), Reviews, 2008 HIPAA Compliance Reviews CMS Office of E-Health Standards and Services (2008)
7. Appari, A., Johnson, M.E.: Information Security and Privacy in Healthcare: Current State of Research (August 2008)
8. Patients' and Citizens Task Force of the European Health Telematics Association (EHTEL), Angelica Frithiof, Rod Mitchell, IAPO, Nicola Bedlington, European Patients Forum, Harm Jan Roelants, Jean Luc Bernard, Le CISS, Johan Hjertqvist, David Garwood, Formulation of policies and standards: The Electronic Health Record, A Position Paper, 25 September (2006)
9. Gostin, L.O., Turek-Brezina, J., Powers, M., Kozloff, R., Faden, R., Steinauer, D.D.: Privacy and security of personal information in a new health care system. The Journal of the American Medical Association 270(20), 2487–2493 (1993)
10. The Department of Information Technology (DIT) (Ministry of Communication & Information Technology (MCIT)) with the support of the project Implementation Agency Apollo Health Street Limited (AHSL), Framework for Information Technology Infrastructure for Health, vol. I, II (2004)

11. Wainer, J., Campos, C.J.R., Salinas, M.D.U., Sigulem, D.: Security Requirements for a Lifelong Electronic Health Record System: An Opinion. Open Med. Inform. Journal 2, 160–165 (2008)
12. Electronic Health Records: Manual for Developing Countries, © World Health Organization (2006)
13. 2011 HIMSS Security Survey, © 2011 Healthcare Information and Management Systems Society, supported by The Medical Group Management Association (MGMA) (November 2011), `http://www.himss.org`
14. Guide for Mapping Types of Information and Information Systems to Security Categories, National Institute of Standards and Technology (NIST), 1 revision, vol. I, 53 pages. NIST Special Publication 800-60 (August 2008)
15. Jean DerGurahian, Data privacy and Security, SearchHealthIT.com (April 2011)
16. Robert Richardson, CSI Director, CSI Computer Crime and Security Survey (2010, 2011), `http://www.GoCSI.com`

On Second-Order Nonlinearities of Two Classes of Cubic Boolean Functions

Deep Singh and Maheshanand Bhaintwal

Department of Mathematics,
Indian Institute of Technology Roorkee, Roorkee 247667 India
deepsinghspn@gmail.com,mahesfma@iitr.ernet.in

Abstract. The higher order nonlinearity of a Boolean function is a cryptographic criterion, which plays an important role in the design of secure block ciphers and stream ciphers. In this paper, we obtain lower bounds of second-order nonlinearities of two classes of highly nonlinear cubic Boolean functions of the form $f(x) = tr_1^n(\lambda x^{2^{2r}+2^{r+1}+1})$, $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$, for $n = 3r$ and $n = 5r$ by investigating the lower bounds of the first order nonlinearity of their derivatives.

Keywords: Boolean functions, Walsh-Hadamard transform, Nonlinearity, Reed-Muller codes.

1 Introduction

The higher order nonlinearity of a Boolean function is a cryptographic criterion, which plays a role against different kinds of attacks, namely: Best Affine Approximation Attacks [9], Higher-order Approximation Attacks [4] etc., on block ciphers and stream ciphers. In addition, it plays a role in coding theory, since it is related to the covering radii of Reed-Muller codes. The rth-order nonlinearity, where $r \geq 1$, of an n-variable Boolean function f , denoted by $nl_r(f)$, is defined as the minimum Hamming distance of f from all n-variable Boolean functions of degrees at most r. The first order nonlinearity (or simply nonlinearity) $nl(f)$ of a Boolean functions on n variables can be computed by using fast Walsh-Hadamard transform in time $o(n2^n)$. Unlike first order nonlinearity, there is no efficient algorithm to compute second-order nonlinearities for $n > 11$. Second-order nonlinearity is known only for some particular functions. Fourquet and Tavernier [5] provided a nice algorithm which works efficiently for $r = 2$ and $n \leq 11$, and in some cases for $n \leq 13$. However, it is in general a difficult problem to obtain an exact lower bound for second order nonlinearities; few attempts have been made in this direction [2,3,10]. Iwata and Kurosawa [10] have given a construction of Boolean functions whose rth-order nonlinearity is lower bounded by

$$nl_r(f) \geq \begin{cases} 2^{n-r-3}(r+4), & \text{if } r = 0\text{mod}2, \\ 2^{n-r-3}(r+5), & \text{if } r = 1\text{mod}2. \end{cases}$$

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 560–567, 2013.

for $0 \leq r \leq n-3$. They have termed the functions satisfying this bound as rth order bent functions. For odd $n \geq 9$, a tight upper bound of nonlinearities of Boolean functions is not known. Therefore, there is a need to construct Boolean functions with controlled nonlinearity profile. The best known asymptotic upper bound on nl_r, obtained by Carlet and Mesnager [3], is

$$nl_r(f) \leq 2^{n-1} - \frac{\sqrt{15}}{2} \cdot (1+\sqrt{2})^{r-2} \cdot 2^{\frac{n}{2}} + O(n^{r-2}).$$

Carlet [2] has developed a recursive approach to compute the lower bounds on rth-order nonlinearities of a function f using the $(r-1)$th-order nonlinearities of the derivatives of the f. Recently, Sun and Wu [15] have obtained a lower bound on second order nonlinearities of a class of cubic Boolean functions of the form $f(x) = tr_1^n(\lambda x^{2^{2r}+2^r+1})$, where $n = 4r$ and $\lambda \in \mathbb{F}_{2^r} \setminus \{0\}$. For more results on lower bounds on second-order nonlinearities of several classes of Boolean functions we refer to [2,6,7,8,11,14,15].

In this paper, we study the lower bounds of second order nonlinearities of two classes of highly nonlinear cubic Boolean functions of the form $f(x) = tr_1^n(\lambda x^{2^{2r}+2^{r+1}+1})$, $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$ for $n = 3r$ and for $n = 5r$. The bounds obtained in this paper are compared with the bounds obtained by Sun and Wu [15] and those in [2,10] and in Theorem 1 of [6].

The remainder of the paper is organized as follows: Section 2 provide preliminary results. Section 3 covers the main results. In Section 4 we compare our results with the existing results. Section 5 concludes the paper.

2 Preliminaries

Boolean functions on n-variables are mappings from $\mathbb{F}_{2^n}$ to $\mathbb{F}_2$. Let $\mathcal{B}_n$ denote the set of all Boolean functions on n-variables. The algebraic normal form of $f \in \mathcal{B}_n$ is $f(x_1, x_2, \ldots, x_n) = \sum_{a=(a_1,\ldots,a_n)\in\mathbb{F}_2^n} \mu_a(\prod_{i=1}^n x_i^{a_i})$, where $\mu_a \in \mathbb{F}_2$. The algebraic degree $\deg(f)$ of f is defined as $\max\{wt(a) : \mu_a \neq 0, a \in \mathbb{F}_{2^n}\}$. For any two functions $f, g \in \mathcal{B}_n$, $d(f,g) = |\{x : f(x) \neq g(x), x \in \mathbb{F}_{2^n}\}|$ is said to be the Hamming distance between f and g. The *derivative* of $f \in \mathcal{B}_n$ with respect to $a \in \mathbb{F}_{2^n}$ is defined by $D_a f(x) = f(x) + f(x+a)$.

The trace function $tr_1^n : \mathbb{F}_{2^n} \to \mathbb{F}_2$ is defined by

$$tr_1^n(x) = x + x^2 + x^{2^2} + \cdots + x^{2^{n-1}}, \text{ for all } x \in \mathbb{F}_{2^n}.$$

The Walsh-Hadamard transform of $f \in \mathcal{B}_n$ at $\lambda \in \mathbb{F}_{2^n}$ is defined as

$$W_f(\lambda) = \sum_{x\in\mathbb{F}_{2^n}} (-1)^{f(x)+tr_1^n(\lambda x)}.$$

Let $\mathcal{A}_n$ be the set of all affine functions on n variables. The nonlinearity of $f \in \mathcal{B}_n$ is defined as $nl(f) = \min_{l\in\mathcal{A}_n}\{d(f,l)\}$. The nonlinearity of $f \in \mathcal{B}_n$ in terms of Walsh-Hadamard transform is defined as

$$nl(f) = 2^{n-1} - \frac{1}{2} \max_{\lambda\in\mathbb{F}_{2^n}} |W_f(\lambda)|.$$

By using Parseval's identity

$$\sum_{\lambda \in \mathbb{F}_{2^n}} (W_f(\lambda))^2 = 2^{2n},$$

it can be shown that $max\{|W_f(\lambda)| : \lambda \in \mathbb{F}_{2^n}\} \geq 2^{n/2}$, i.e., $nl(f) \leq 2^{n-1} - 2^{\frac{n}{2}-1}$. For n even, $f \in \mathcal{B}_n$ is called a bent [13] function if $nl(f) = 2^{n-1} - 2^{\frac{n}{2}-1}$.

Suppose $f \in \mathcal{B}_n$ is a quadratic function and $B(x,y) = f(0) + f(x) + f(y) + f(x+y)$ is the bilinear form associated with f. The kernel $\mathcal{E}_f$ of $B(x,y)$ is the subspace of $\mathbb{F}_{2^n}$ defined by

$$\mathcal{E}_f = \{x \in \mathbb{F}_{2^n} : B(x,y) = 0 \text{ for all } y \in \mathbb{F}_{2^n}\}.$$

Lemma 1 ([1], Proposition 1). *Let V be a vector space over a field $\mathbb{F}_q$ of characteristic 2 and $P : V \longrightarrow \mathbb{F}_q$ be a quadratic form. Then the dimension of V and the dimension of the kernel of P have the same parity.*

Lemma 2 ([1], Lemma 1). *Let f be any quadratic Boolean function. The kernel $\mathcal{E}_f$ is the subspace of $\mathbb{F}_{2^n}$ consisting of those a such that the derivative D_af is constant. That is,*

$$\mathcal{E}_f = \{a \in \mathbb{F}_{2^n} : D_af = \text{ constant }\}.$$

If $f \in \mathcal{B}_n$ be a quadratic Boolean function and $B(x,y)$ be the associated bilinear form, then the Walsh spectrum of f depends only on the dimension of the kernel $\mathcal{E}_f$ of $B(x,y)$ [1,12].

Carlet [2] has obtained the following recursive lower bounds on the nonlinearity profile of Boolean functions.

Proposition 1 ([2], Proposition 2). *Let $f \in \mathcal{B}_n$ and r be a positive integer $(r < n)$, then we have*

$$nl_r(f) \geq \frac{1}{2} \max_{a \in \mathbb{F}_{2^n}} nl_{r-1}(D_af).$$

Proposition 2 ([2], Proposition 3). *Let $f \in \mathcal{B}_n$ and r be a positive integer $(r < n)$, then we have*

$$nl_r(f) \geq 2^{n-1} - \frac{1}{2}\sqrt{2^{2n} - 2\sum_{a \in \mathbb{F}_{2^n}} nl_{r-1}(D_af)}.$$

Table 1. Weight distribution of the Walsh spectrum of a quadratic function f

$W_f(\alpha)$	number of α
0	$2^n - 2^{n-k}$
$2^{(n+k)/2}$	$2^{n-k-1} + (-1)^{f(0)}2^{(n-k-2)/2}$
$-2^{(n+k)/2}$	$2^{n-k-1} - (-1)^{f(0)}2^{(n-k-2)/2}$

The following corollary is due to Proposition 2.

Corollary 1 ([2], Corollary 2). *Let $f \in \mathcal{B}_n$ and r be a positive integer $(r < n)$. Assume that, for some nonnegative integers M and m, we have $nl_{r-1}(D_a f) \geq 2^{n-1} - M2^m$ for every nonzero $a \in \mathbb{F}_{2^n}$. Then*

$$nl_r(f) \geq 2^{n-1} - \tfrac{1}{2}\sqrt{(2^n-1)M2^{m+1} + 2^n}.$$

Carlet [2] further remarked that in general the lower bound obtained in Proposition 2 is better than the bound obtained in Proposition 1.

The following result is used to improve the bounds of the rth-order nonlinearities of a Boolean function and known as McEliece's theorem.

Proposition 3 ([12], Chap. 15, Cor. 13). *The rth-order nonlinearities of $f \in \mathcal{B}_n$ with algebraic degree d is divisible by $2^{\lceil \frac{n}{d} \rceil - 1}$.*

3 Main Results

In this section, we deduce the lower bounds of the second order nonlinearity of two classes of cubic Boolean functions by investigating the lower bounds of the first order nonlinearity of their derivatives.

Theorem 1. *Suppose $f \in \mathcal{B}_n$ such that $f(x) = tr(\lambda x^{2^{2r}+2^{r+1}+1})$ $\forall$ $x \in \mathbb{F}_{2^n}$, with $n = 3r$ and $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$. Then*

$$nl_2(f) \geq 2^{n-1} - \frac{1}{2}\sqrt{(2^n-1)2^{\frac{n+r+2}{2}} + 2^n}.$$

Proof. The derivative of $f(x) = tr(\lambda x^{2^{2r}+2^{r+1}+1})$ with $n = 3r$ and $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$, with respect to $a \in \mathbb{F}_{2^n} \setminus \{0\}$ is

$$\begin{aligned} D_a f(x) = f(x) + f(x+a) &= tr\left(\lambda (x+a)^{2^{2r}+2^{r+1}+1}\right) + tr(\lambda x^{2^{2r}+2^{r+1}+1}) \\ &= tr\left(\lambda\left(x^{2^{2r}+2^{r+1}} + x^{2^{2r}}a^{2^{r+1}} + x^{2^{r+1}}a^{2^{2r}} + a^{2^{2r}+2^{r+1}}\right)(x+a)\right. \\ &\qquad \left. + \lambda x^{2^{2r}+2^{r+1}+1}\right) \\ &= tr\left(\lambda\left(x^{2^{2r}+2^{r+1}}a + x^{2^{2r}+1}a^{2^{r+1}} + x^{2^{r+1}+1}a^{2^{2r}}\right)\right) + l(x), \end{aligned}$$

where $l(x)$ is an affine function. Let $b \in \mathbb{F}_{2^n} \setminus \{0\}$ be such that $a \neq b$, then

$$\begin{aligned} D_b D_a f(x) &= tr\left(\lambda\left((x+b)^{2^{2r}+2^{r+1}}a + (x+b)^{2^{2r}+1}a^{2^{r+1}} + (x+b)^{2^{r+1}+1}a^{2^{2r}}\right)\right) \\ &\quad + tr\left(\lambda(x^{2^{2r}+2^{r+1}}a + x^{2^{2r}+1}a^{2^{r+1}} + x^{2^{r+1}+1}a^{2^{2r}})\right) + constant \\ &= tr\left(x^{2^{2r}}(\lambda b a^{2^{r+1}} + \lambda a b^{2^{r+1}}) + x^{2^{r+1}}(\lambda b a^{2^{2r}} + \lambda a b^{2^{2r}})\right. \\ &\qquad \left. + x(\lambda a^{2^{r+1}} b^{2^{2r}} + \lambda a^{2^{2r}} b^{2^{r+1}})\right) + constant \end{aligned}$$

$$\begin{aligned}
&= tr\left(x\left((\lambda b a^{2^{r+1}} + \lambda a b^{2^{r+1}})^{2^r} + (\lambda b a^{2^{2r}} + \lambda a b^{2^{2r}})^{2^{2r-1}}\right.\right. \\
&\qquad \left.\left. +(\lambda a^{2^{r+1}} b^{2^{2r}} + \lambda a^{2^{2r}} b^{2^{r+1}})\right)\right) + constant \\
&= tr\left(x\left(\lambda^{2^r} b^{2^r} a^{2^{2r+1}} + \lambda^{2^r} a^{2^r} b^{2^{2r+1}} + \lambda^{2^{2r-1}} b^{2^{2r-1}} a^{2^{r-1}}\right.\right. \\
&\quad \left.\left. +\lambda^{2^{2r-1}} a^{2^{2r-1}} b^{2^{r-1}} + \lambda a^{2^{r+1}} b^{2^{2r}} + \lambda a^{2^{2r}} b^{2^{r+1}}\right)\right) + constant \\
&= tr\left(x P_{\lambda,a}(b)\right) + constant.
\end{aligned}$$

Clearly, $D_b D_a f(x)$ is constant if and only if $P_{\lambda,a}(b) = 0$. Therefore,

$$\mathcal{E}_{D_a f} = \{b \in \mathbb{F}_{2^n} :\ P_{\lambda,a}(b) = 0\}.$$

The kernel $\mathcal{E}_{D_a f}$ of $D_a f$ is the set of zeroes of $P_{\lambda,a}(b)$, or equivalently the set of zeroes of $(P_{\lambda,a}(b))^{2^{2r+1}}$. Therefore,

$$\begin{aligned}
L_{\lambda,a}(b) = (P_{\lambda,a}(b))^{2^{2r+1}} &= \left(\lambda^{2^r} b^{2^r} a^{2^{2r+1}} + \lambda^{2^r} a^{2^r} b^{2^{2r+1}} + \lambda^{2^{2r-1}} b^{2^{2r-1}} a^{2^{r-1}}\right. \\
&\quad \left. +\lambda^{2^{2r-1}} a^{2^{2r-1}} b^{2^{r-1}} + \lambda a^{2^{r+1}} b^{2^{2r}} + \lambda a^{2^{2r}} b^{2^{r+1}}\right)^{2^{2r+1}} \\
&= \lambda^2 a^2 b^{2^{r+2}} + \lambda^{2^{2r+1}} a^4 b^{2^{r+1}} + \lambda^{2^r} a b^{2^r} \\
&\quad +\lambda^{2^{2r+1}} a^{2^{r+1}} b^4 + \lambda^2 b^2 a^{2^{r+2}} + \lambda^{2^r} a^{2^r} b.
\end{aligned}$$

Since $L_{\lambda,a}(b)$ is a linearized polynomial in b of degree at most 2^{r+2}. This implies that the dimension k of $\mathcal{E}_{D_a f}$ is at most $r+2$. Thus, for all $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$, Walsh-Hadamard transform is

$$W_{D_a f}(\alpha) = 2^{\frac{n+k}{2}} \leq 2^{\frac{n+r+2}{2}}.$$

Since the nonlinearity of the derivative $D_a f$ of $f \in \mathcal{B}_n$ is

$$nl(D_a f) = 2^{n-1} - \frac{1}{2} \max_{\lambda \in \mathbb{F}_{2^n}} |W_{D_a f}(\lambda)|,$$

therefore, we have

$$nl(D_a f) \geq 2^{n-1} - 2^{\frac{n+r}{2}}. \tag{1}$$

Using Proposition 1, we have

$$nl_2(f) \geq \frac{1}{2}\left(2^{n-1} - 2^{\frac{n+r}{2}}\right) = 2^{n-2} - 2^{\frac{n+r-2}{2}}. \tag{2}$$

On comparing 1 and Corollary 1, we get $M = 1, m = \frac{n+r}{2}$. Now, using the result of Corollary 1, we get

$$\begin{aligned}
nl_2(f) &\geq 2^{n-1} - \frac{1}{2}\sqrt{(2^n - 1)M 2^{m+1} + 2^n} \\
&= 2^{n-1} - \frac{1}{2}\sqrt{(2^n - 1) 2^{\frac{n+r+2}{2}} + 2^n}.
\end{aligned} \tag{3}$$

On subtracting the lower bound obtained in (2) from the lower bound obtained in (3) and using $n = 3r$, we get

$$2^{n-1} - \tfrac{1}{2}\sqrt{(2^n-1)2^{\frac{n+r+2}{2}} + 2^n} - \tfrac{1}{2}\left(2^{n-1} - 2^{\frac{n+r}{2}}\right)$$
$$= 2^{n-2} + 2^{\frac{4n-6}{6}} - \tfrac{1}{2}\sqrt{(2^n-1)2^{\frac{4n+6}{6}} + 2^n} > 0.$$

for all values of n. Therefore, the bound given in (3) is better than the lower bound obtained in (2). Thus, we have

$$nl_2(f) \geq 2^{n-1} - \frac{1}{2}\sqrt{(2^n-1)2^{\frac{n+r+2}{2}} + 2^n}. \qquad \square$$

In the following result we provide lower bounds on second order nonlinearities of the function $f(x) = tr_1^n(\lambda x^{2^{2r}+2^{r+1}+1})$ for all $x \in \mathbb{F}_{2^n}$, where $n = 5r$ and $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$.

Theorem 2. *Suppose $f \in \mathcal{B}_n$ such that $f(x) = tr_1^n(\lambda x^{2^{2r}+2^{r+1}+1})$ $\forall$ $x \in \mathbb{F}_{2^n}$, with $n = 5r$ and $\lambda \in \mathbb{F}_{2^n} \setminus \{0\}$. Then*

$$nl_2(f) \geq 2^{n-1} - \frac{1}{2}\sqrt{(2^n-1)2^{\frac{n+3r+2}{2}} + 2^n}.$$

Proof. The proof is almost similar to the proof of Theorem 1.

Remark 1. The general lower bounds on second order nonlinearities of Boolean functions due to Carlet [2] and Iwata-Kurosawa [10] are $2^{n-1} - 2^{n-\frac{3}{2}}$ and $2^{n-2} - 2^{n-4}$, respectively. Clearly,

$$(2^{n-2} - 2^{n-4}) - (2^{n-1} - 2^{n-\frac{3}{2}}) = 2^{n-4}(4\sqrt{2} - 5) \geq 0.$$

Hence, the bounds obtained by Iwata-Kurosawa [10] are better than Carlet's general bounds [2]. Now, on subtracting Iwata-Kurosawa's bounds from the bounds obtained in Theorem 2 and using $n = 5r$, we have

$$\left(2^{n-1} - \tfrac{1}{2}\sqrt{(2^n-1)2^{\frac{n+3r+2}{2}} + 2^n}\right) - \left(2^{n-2} - 2^{n-4}\right)$$
$$= 5\ 2^{n-4} - \tfrac{1}{2}\sqrt{(2^n-1)2^{\frac{8n+10}{10}} + 2^n} > 0$$

if and only if $n \geq 12$. Therefore, in Theorem 2 we investigate a class of highly nonlinear cubic Boolean functions whose lower bounds of second order nonlinearities (for all $n \geq 12$) are better than the lower bounds obtained by Iwata-Kurosawa [10].

4 Comparison

In Table 2, we present the numerical comparison between the lower bounds obtained in Theorem 1 with the lower bounds those obtained by Iwata-Kurosawa

Table 2. Numerical comparison of the lower bounds on second-order nonlinearities obtained by Theorem 1 using McEliece's Theorem with the bounds obtained in [2,6,10,15]

n	6	9	12	15	18	21	24
Bounds in Theorem 1	10	128	1328	12288	107904	917504	7.647232×10^6
Bounds due to Iwata-Kurosawa [10]	12	96	764	6144	49152	393216	3.145728×10^6
Bounds in [6, Theorem 1]	10	–	1024	–	84732	–	6.291456×10^6
Bounds due to Sun-Wu [15]	–	–	1318	–	–	–	7.339910×10^6
Carlet's general bounds [2]	10	76	600	4800	38392	307122	2.456968×10^6

Table 3. Numerical comparison of the lower bounds on second-order nonlinearities obtained by Theorem 2 using McEliece's Theorem with the bounds obtained in [2,10]

n	15	20	25	30	35
Bounds in Theorem 2	8192	338944	12582912	441965056	$1.5032385536 \times 10^{10}$
Iwata-Kurosawa's bounds [10]	6144	196608	6291456	201326592	6.442450944×10^9
Carlet's general bounds [2]	4800	153562	4913934	157245850	5.031867186×10^9

[10], Gangopadhyay et al. [6], Sun-Wu [15] and the general bounds obtained by Carlet in [2]. It is clear from Table 2 that for $n \geq 9$, the bounds obtained in Theorem 1 are better than the bounds obtained in [2,10,6,15]. Further, it is observed from remark 1 and Table 3 that the bounds obtained in Theorem 2 are larger than the bounds obtained by Iwata-Kurosawa in [10] and Carlet's general bounds in [2].

5 Conclusion

In this paper, we have obtained the lower bounds on second-order nonlinearities of two classes of highly nonlinear cubic Boolean functions. Here it should be noted that, high first order nonlinearity of a Boolean function does not implies the high second order nonlinearity. For example: bent function $tr(x^{2^i+1})$ possess maximum possible first order nonlinearity but their second order nonlinearity is zero. The results in this paper show that the lower bounds of second order nonlinearity of two classes of Boolean functions are also high. In particular, the lower bounds of second order nonlinearities obtained in Theorem 1 for $n \geq 9$ (when $n = 3r$) are better than the lower bounds obtained in [2,6,10,15]. Further, the lower bounds of second order nonlinearities obtained in Theorem 2 for $n \geq 12$ (when $n = 3r$) are better than those of Iwata-Kurosawa [10]. Since higher order nonlinearity of Boolean functions is useful in providing protection against different kinds of attacks on stream and block ciphers therefore, we expect that the results in this paper may be useful in choosing cryptographically significant Boolean functions.

Acknowledgments. The work of the first author is supported by NBHM (DAE), INDIA.

References

1. Canteaut, A., Charpin, P., Kyureghyan, G.M.: A New Class of Monomial Bent Functions. Finite Fields and Appli. 14, 221–241 (2008)
2. Carlet, C.: Recursive Lower Bounds on the Nonlinearity Profile of Boolean Functions and Their Applications. IEEE Trans. Inform. Theory 54(3), 1262–1272 (2008)
3. Carlet, C., Mesnager, S.: Improving the Upper Bounds on the Covering Radii of Binary Reed-Muller Codes. IEEE Trans. Inform. Theory 53(1), 162–173 (2007)
4. Courtois, N.T.: Higher Order Correlation Attacks, XL Algorithm and Cryptanalysis of Toyocrypt. In: Lee, P.J., Lim, C.H. (eds.) ICISC 2002. LNCS, vol. 2587, pp. 182–199. Springer, Heidelberg (2003)
5. Fourquet, R., Tavernier, C.: An Improved List Decoding Algorithm for the Second Order Reed-Muller Codes and its Applications. Designs Codes and Crypto. 49, 323–340 (2008)
6. Gangopadhyay, S., Sarkar, S., Telang, R.: On the Lower Bounds of the Second Order Nonlinearities of Some Boolean Functions. Information Sciences 180, 266–273 (2010)
7. Gangopadhyay, S., Singh, B.K.: On Second-Order Nonlinearities of Some Type Bent Functions, `http://eprint.iacr.org/2010/286.pdf`
8. Gode, R., Gangopadhyay, S.: On Lower Bounds of Second-Order Nonlinearities of Cubic Bent Functions Constructed by Concatenating Gold Functions. Int. J. Comput. Math. 88(15), 3125–3135 (2011)
9. Golić, J.: Fast Low Order Approximation of Cryptographic Functions. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 268–282. Springer, Heidelberg (1996)
10. Iwata, T., Kurosawa, K.: Probabilistic Higher Order Differential Attack and Higher Order Bent Functions. In: Lam, K.-Y., Okamoto, E., Xing, C. (eds.) ASIACRYPT 1999. LNCS, vol. 1716, pp. 62–74. Springer, Heidelberg (1999)
11. Kolokotronis, N., Limniotis, K.: Maiorana-McFarland Functions with High Second Order Nonlinearity, `http://eprint.iacr.org/2011/212.pdf`
12. MacWilliams, F.J., Sloane, N.J.A.: The Theory of Error Correcting Codes. North-Holland, Amsterdam (1977)
13. Rothaus, O.S.: On "Bent" Functions. Journal of Combinatorial Theory, Series A 20, 300–305 (1976)
14. Sun, G., Wu, C.: The Lower Bounds on the Second-Order Nonlinearity of Three Classes of Boolean Functions with High Nonlinearity. Information Sciences 179(3), 267–278 (2009)
15. Sun, G., Wu, C.: The Lower Bounds on the Second-Order Nonlinearity of a Class of Boolean Functions with High Nonlinearity. Appli. Alg. in Eng. Commu. and Comp. 22, 37–45 (2011)

A Certificateless Authenticated Key Agreement Protocol for Digital Rights Management System

Dheerendra Mishra and Sourav Mukhopadhyay

Department of Mathematics
Indian Institute of Technology
Kharagpur–721302, India
{dheerendra,sourav}@maths.iitkgp.ernet.in

Abstract. Digital rights management (DRM) is the system which tries to ensure authorized content consumption. Current DRM systems either adopt public key cryptography (PKC) or identity based public key cryptography (ID-PKC). PKC associates certificate management which includes revocation, storage, distribution and verification of certificate, as a result, certificate authority becomes the bottleneck for the large network. While, ID-PKC has drawback of key escrow. However, for secure and authorized content distribution, evacuation from these problems is needed. In this paper, we present a certificateless authenticated key agreement protocol for DRM system, which ensures flawless mutual authentication and establishes a session key between user and license server. Furthermore, we analyzed proposed scheme to show that proposed scheme is secured.

Keywords: Digital rights management, Certificateless public key cryptography, Bilinear pairing, Authentication.

1 Introduction

With the advancement of Internet technology, e-commerce industry achieves a scalable infrastructure for digital content distribution at low cost. Internet facilitates an online trade of digital content (text, music, movies, and software). However, the digital contents can be easily copied and redistributed over the network without any drop in the quality of contents. As a result, illegal copies of content are available over the network which causes a huge loss of revenue to the right holders. Therefore, rights holders are demanding a mechanism which can regulate the authorized content consumption so that copyright protection would be achieved. One such a mechanism is digital rights management (DRM)system which is developed to ensure the copyright protection [11].

DRM system also tries to maintain flexible and secure content distribution. For flexible and secure communication, an efficient mutual authentication and key agreement protocol is needed where the involved parties can authenticate each other and establish a secure session key. The session key is generated with the information shares of involve parties which used to achieve its goal of confidentiality and data integrity. Current DRM systems [9,12,16,6,7,8,15] basically

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 568–577, 2013.

apply two approaches, namely, public key cryptography (PKC) [14] and identity based public key cryptography (ID-PKC) [13]. Schemes [9,12] use PKC to authenticate public key where PKC maintains a certificate authority (CA). CA proofs the relation between entity and its public key. Moreover, it manages certificate management including storage, distribution and revocation. However, CA becomes bottleneck for large network. Therefore, computational cost of certificate verification become infeasible. While, schemes [6,7,15] apply identity based infrastructure where involve parties achieve their private key from private key generator (PKG) and public key is derived from their public identity such as email address. Yen et al. [15] also presented an ID based authenticated key agreement protocol which manage secure communication. However, in ID-PKC, PKG knows the private key of each user, that means PKG could generates forge signature of any entity. This causes the key escrow problem.

In this paper, we will apply the pairing based certificateless authenticated key agreement protocol for DRM system which is introduced by Al-Riyami and Paterson [1]. In this scheme, license server and user achieve their private keys using PKG generated partial private key share and self generates secret value. Further, both parties establish authenticated session key to communicate securely. Even more, They can establish different session keys for different sessions to achieve security. The proposed protocol eliminated the use of trusted certificate authority and solve key escrow problem. Moreover, user adopts symmetric key encryption to achieve content license which requires less computation compare to public key encryption.

The rest of the paper is organized as follows: In section 2, we discuss a typical DRM model, recall the concept of public key cryptography, identity based cryptography and certificateless cryptography, and define some notation that we will use throughout the paper. We present our content distribution scheme in section 3. We present security analysis in section 4. Finally, in section 5, we draw a conclusion.

2 Preliminaries

2.1 Basic DRM System

A general DRM architecture involves four core component: content provider (owner), license server, distributor and user [11].

Content Provider. Content provider holds the digital rights of the content and wants to protect the content. It works as a packager. To protect the content from unauthorized user, it encrypts the content. It provides protected content with content information to the distributor and content key with usage rules to the license server.

Distributor. Distributor works as a service provider. It associates a media server and sets up a website. It keeps protected content over the media server and display content information over the website. The distributor provides encrypted content and content detailed information (file size, file type, player, etc.) to the users.

License Server. License server generates the license by using key seeds where license comprises of content key, usage rules and constraints. It authenticates the user by using standard authentication mechanisms such as password based, smart card based, etc. It issues the license only for authorized users.

User. A DRM user downloads the protected content from the media server and acquires the license from the license server. A user always wants that content should be easy to play and easy to download besides secure payment mechanism and privacy.

2.2 Bilinear Pairings

Let $(G_1, +)$ and $(G_2, .)$ be the additive and multiplicative cyclic groups of order q respectively where q is a k-bit large prime. The bilinear pairing $e : G_1 \times G_1 \rightarrow G_2$ defined by $e(\cdot, \cdot)$ has the following properties as discussed in [2,4]:

- **Bilinear:** $e(aP, bQ) = e(P, Q)^{ab}$, for all $P, Q \in G_1$ and $a, b \in Z_q^*$;
- **Non-degenerate:** There exist $P, Q \in G_1$ such that $e(P, Q) \neq 1$;
- **Computable:** There exists an efficient algorithm to compute $e(P, Q)$, for all $P, Q \in G_1$.

The security of CL-PKC in $e(\cdot, \cdot)$ based on the hardness of following computational problems:

Discrete Logarithm Problem: For a given generator P of G_1 and $Q \in G_1$, find an element $a \in Z_q^*$ such that $aP = Q$.
Computational Diffie-Hellman (CDH) Problem: Let P be a generator of G_1. Given $\langle P, aP, bP \rangle \in G_1$ compute abP for $a, b \in Z_q^*$.
Bilinear diffie-Hellman (BDH) Problem: Let P be a generator of G_1. Given $\langle P, aP, bP, cP \rangle \in G_1$ compute $e(P, P)^a bc$, for $a, b, c \in Z_q^*$.

2.3 Public Key Cryptography

The public key cryptography (PKC) is introduced by Diffie and Hellman [5]. PKC involves two different keys for encryption and decryption instead of single key as symmetric key system. Since, public key is random string in PKC. Therefore, To prove the relation between entity and its public key, PKC adopts certificate mechanism where certificate-based protocols work by considering that each entity has a public and private key pair. These public keys are authenticated via certificate authority (CA) which issue a certificate. When two entities wish to establish a session key, a pair of ephemeral (short term) public keys are exchanged between them. The ephemeral and static keys are then combined in a way so as to obtain the agreed session key. The authenticity of the static keys provided by signature of CA assures that only the entities who posses the static keys are able to compute the session key.

2.4 Identity Based Public Key Cryptography (ID-PKI)

Identity-based cryptosystem eliminates the generation and distribution of entities public key problem by making each entity public key derivable from some known aspects of his identity, such as email address. Here, entities achieve their private key from a trusted third party called a Private Key Generator (PKG), after their authenticity verification. Shamir [13] introduced the concept of identity-based encryption (IBE) to simplify public key management procedures (or the public key distribution problem) by eliminating certificate-based public key infrastructure. However, the first fully functional pairing-based IBE scheme was proposed in [2]. Shortly after this, many pairing based cryptographic protocols were proposed. A survey over pairing based cryptography is presented in [4]. The identity-based PKI can be an efficient alternative of certificate-based PKI, especially when efficient key management and moderate security are required for large networks.

In an ID-based encryption scheme consists of four algorithms (*i*) Setup, (*ii*) Extract (Key generation), (*iii*) Encryption, and (*iv*) Decryption. For more details, one can refer [3].

2.5 Certificateless Public Key Cryptography (CL-PKI)

Certificateless cryptography is introduced in 2003 by Al-Riyami and Paterson [1]. It eliminates the necessity of certificate authority (CA) and removes key escrow problem in the system. It comprises of seven algorithms which are as follows:

Setup: It is a probabilistic algorithm run by the private Key Generator (PKG) which takes security parameter, randomly chosen master key and a list of public parameters such as description of message space and ciphertext space.
Partial Private Key-Extract: It is a probabilistic algorithm which run by the PKG. It takes input as user's identity $ID \in \{0,1\}^*$ and the master key. It returns partial private key.
Set Secret Value: It is a probabilistic algorithm which is perform by the entity. It takes list of public parameters and produces a random secret value for entity.
Set Private Key: It is a deterministic private key generation algorithm which is run by the entity. It takes input as entity partial private key and secret value, then outputs a private key.
Set Public Key: It is a deterministic public key generation algorithm which run by entity. It takes parameter and entity secret value, then computes entity public key.
Encrypt: It is a probabilistic algorithm which takes input as message, receiveridentity and public key. It outputs ciphertext.
Decrypt: It is a deterministic algorithm which takes a ciphertext and receiver private key. It returns original message.

3 Proposed Protocol

The basic architecture of proposed DRM system is similar to Liu et al. [10] system. Here, the content provider handles the content packing (encryption) work. Once the content encryption is over, it provides the content key with usage rules to the license server and protected content with content information to the distributor. License server authenticates the user, receives the payment, and generates the license. While, Distributor works as a service provider and facilitates the protected content distribution in the system. Parties involved in our DRM model are:

- Private key generator (PKG)
- Content provider (C)
- Distributor (D)
- License server (L)
- DRM User (U)

Content provider keeps the original unprotected digital contents and provides these contents for business use after their encryption. If it has r contents, namely, $M_1, M_2, \ldots, M_r$ with their unique identity $\mathsf{id}_{M_1}, \mathsf{id}_{M_2}, \ldots, \mathsf{id}_{M_r}$. Then, he generates r symmetric keys K_1, K_2, $K_3, \ldots, K_r$ and encrypts each content with an unique symmetric key and gets

$$E_{\mathsf{sym}}(M_i|K_i),\ i = 1, 2, 3, \ldots, r.$$

Content provider provides content decryption keys (key seeds) with usage rules and permissions to the license server through a secure channel. Distributor achieves encrypted contents $\{E_{\mathsf{sym}}(M_i|K_i),\ \text{for all } i = 1, 2, 3, \ldots, r\}$ with content information from the Packager. Distributors keep protected contents over the media server and display content details over the website. To communicate securely in the system, entities achieve their secret partial keys with the help of packager and generates their public and private keys. In this process system usages five algorithms: Setup, Partial private key extract, Set secret value, Set private key and Set public key. Description of key generation process is as follows:

Setup: Private key generator (PKG) chooses an arbitrary generator $P \in G_1$, selects a master key $\mathsf{mk} \in Z_q^*$ and sets $\mathsf{PK} = \mathsf{mk}P$. It chooses hash functions $H_1 : \{0,1\}^* \to G_1^*$, $H_2 : \{0,1\}^k \times \{0,1\}^* \times \{0,1\}^* \to \{0,1\}^n$, and $H : \{0,1\}^* \times \{0,1\}^* \times G_1 \times G_1 \times G_2 \to \{0,1\}^k$. Then, PKG publishes system parameters $\langle G_1, G_2, e(.,.), k, P, \mathsf{PK}, H_1, H_2, H\rangle$ and Keep master key mk secret.

Partial Private key extraction: License server (L) and user U submit their public identities ID_L and ID_U to the PKG. Then, PKG verifies the proof of identities. If verification succeeds, then generates the partial private keys in the following way:

- Compute $Q_L = H_1(ID_L)$ and$Q_U = H_1(ID_U) \in G_1^*$.
- By using its master key mk, PKG generates the partial private keys $W_L = \mathsf{mk}Q_L$ and $W_U = \mathsf{mk}Q_U$ and delivers these partial keys W_L and W_U to L and U respectively through a secure channel.

On receiving their partial private keys L and U can verify their partial keys respectively as follows:

$$e(W_L, P) = e(\mathsf{mk}Q_L, P) = e(Q_L, \mathsf{mk}P) = e(Q_L, \mathsf{PK})$$

$$e(W_U, P) = e(\mathsf{mk}Q_U, P) = e(Q_U, \mathsf{mk}P) = e(Q_U, \mathsf{PK}).$$

Private and public key extraction: L and U achieve their private and public keys as follows:

- L selects a secret value $x_L \in Z_q^*$ at random and keeps x_L secret. Then, L generates its private key SK_L by computing $SK_L = x_L W_L = x_L \mathsf{mk} Q_L$. L constructs its public key $PK_L = \langle X_L, Y_L \rangle$ where $X_L = x_L P$ and $Y_L = x_L \mathsf{PK} = x_L \mathsf{mk} P$.
- U selects a secret value $x_U \in Z_q^*$ at random and keeps x_U secret. Then, U generates its private key SK_U by computing $SK_U = x_U W_U = x_U \mathsf{mk} Q_U$. U constructs its public key $PK_U = \langle X_U, Y_U \rangle$ where $X_U = x_U P$ and $Y_U = x_U \mathsf{PK} = x_U \mathsf{mk} P$.

3.1 License Acquisition

User visits the distributor's website and selects some content with identity id_{M_t} and downloads encrypted content $E_{\mathsf{sym}}(M_t|K_t)$ from media server where media server provides free download of encrypted content. However, the encrypted content can not be played without the valid license where license server issues the license for authorized users. To acquire the license, user first establishes an authenticated key agreement protocol with license server, then achieves the license by using established secure communication from the license server. The detailed process is as follows:

Step 1. U chooses a random value $u \in Z_q^*$ and computes $T_U = uP$. Then, sends $\langle ID_U, T_U, PK_U \rangle$ to L.
Step 2. On receiving the user message, L selects a random value $l \in Z_q^*$ and gets $T_L = lP$. Then, L computes $Q_U = H_1(ID_U)$ and achieves S_L as follows:

$$S_L = e(Q_U, Y_U)^l \cdot e(SK_L, T_U) = e(Q_U, P)^{x_U \mathsf{mk} l} \cdot e(Q_L, P)^{x_L \mathsf{mk} u}.$$

Step 3. L computes $lT_U = luP$ and $x_L X_U = x_L x_U P$, then gets the session key as:

$$sk = H(ID_U || ID_L || luP || x_L x_U P || S_L).$$

Then, L sends $\langle ID_L, P_L, T_L, \mathsf{mac} \rangle$ to U where $\mathsf{mac} = H_2(sk || ID_U || ID_L)$.
Step 4. On receiving the message, U computes $Q_L = H_1(ID_L)$ and achieve S_U as follows:

$$S_U = e(Q_L, Y_L)^u \cdot e(SK_U, T_L) = e(Q_L, P)^{x_L \mathsf{mk} u} \cdot e(Q_U, P)^{x_U \mathsf{mk} l}.$$

Step 5. U computes $uT_L = ulP$ and $x_U X_L = x_U x_L P$, then gets the session key as

$$sk^* = H(ID_U||ID_L||ulP||x_U x_L P||S_U).$$

Then, U compute $\mathsf{mac}^* = H_2(sk^*||ID_L||ID_U)$ and checks the condition $\mathsf{mac}^* =? \mathsf{mac}$. If the condition hold, U selects his required content with identity id_{M_t}. Then, U encrypts id_M using session key sk and sends encrypted id_M with mac^* to L.

Step 6. On receiving the message, L verifies $\mathsf{mac} =? \mathsf{mac}^*$. If verification success, L decrypts the encrypted identity using sk and gets id_{M_t}. Then, L receives the payment. L allows two types of payment system which are as follows:

- *Prepayment*: user deposits an initial amount to the license server and gets a membership. A member can engage in a virtual finite number of interactions with the license server, to get the license at the total cost, which does not exceed the initial deposit amount.
- *Pays per item*: User need not to deposit any initial amount as an advance. In this case, the user pays at the time of license acquisition.

Step 7. On receiving the payment, L generates the license $license_{id_{M_t}}$ where license includes serial number, content key, usages rules and user's identity.

Step 8. L encrypts the license using symmetric session key sk and sends encrypted license to U. In addition, license server also maintain the record of usages license statistic for future business use.

Step 9. On receiving the message, U decrypts the message using session key sk and achieve the desired license $\mathsf{license}_{id_M}$. With the help of license, user can play the content.

User needs to established a session key only once. Once the session has established, a user can achieve any number of license in that session. To enhance the security, user can establish independent session keys for each session. An overview of pairing based authenticated key agreement protocol is given in figure 1.

4 Security Analysis

In this section, we will justify that proposed scheme provides authorized and secure communication between license server and user.

Passive attack: Eavesdropper can collect the information $\langle P, uP, lP, x_U\mathsf{mk}P, x_L\mathsf{mk}P, Q_U, Q_L, x_U P, x_L P\rangle$ which transmits via public channel. However, to compute $e(Q_L, P)^{x_L\mathsf{mk}u}$ and $e(Q_U, P)^{x_U\mathsf{mk}l}$ from given $\langle Q_L, uP, x_L\mathsf{mk}P\rangle$ and $\langle Q_U, lP, x_U\mathsf{mk}P\rangle$ respectively is equivalent to BDH problem. Where, BDH problem is hard to compute. Moreover, to achieve session key $sk = H(ID_U||ID_L||ulP||S_U)$, the values ulP is needed. But, to compute ulP from given $\langle P, uP, lP\rangle$ is equivalent to CDH problem which is hard.

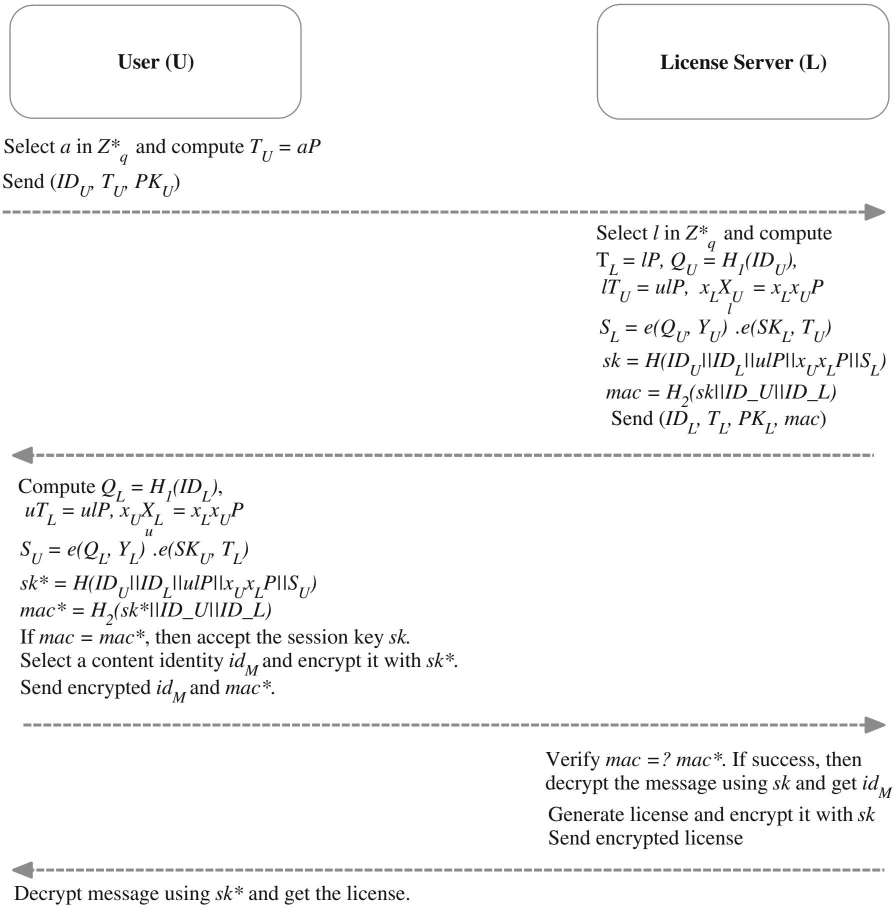

User (U)
License Server (L)
Select a in Z^*_q and compute $T_U = aP$
Send (ID_U, T_U, PK_U)
Select l in Z^*_q and compute
$T_L = lP, Q_U = H_1(ID_U),$
$lT_U = ulP, \ x_L X_U^{\ l} = x_L x_U P$
$S_L = e(Q_U, Y_U).e(SK_L, T_U)$
$sk = H(ID_U||ID_L||ulP||x_U x_L P||S_L)$
$mac = H_2(sk||ID\_U||ID\_L)$
Send (ID_L, T_L, PK_L, mac)
Compute $Q_L = H_1(ID_L),$
$uT_L = ulP, x_U X_L^{\ u} = x_L x_U P$
$S_U = e(Q_L, Y_L).e(SK_U, T_L)$
$sk^* = H(ID_U||ID_L||ulP||x_U x_L P||S_U)$
$mac^* = H_2(sk^*||ID\_U||ID\_L)$
If $mac = mac^*$, then accept the session key sk.
Select a content identity id_M and encrypt it with sk^*.
Send encrypted id_M and mac^*.
Verify $mac =? \ mac^*$. If success, then
decrypt the message using sk and get id_M
Generate license and encrypt it with sk
Send encrypted license
Decrypt message using sk^* and get the license.

Fig. 1. Proposed license distribution mechanism

Man in the middle attack: User and license server authenticate each other without knowing each other. An adversary or malicious PKG can try man in the middle attack by sending the forge message. However, to authenticate each other message, license server and user exchange mac and mac^* to each other. Where, to compute mac $= H_2(sk||ID_U||ID_L)$ requires to compute secret session key sk. To compute sk an adversary needs to compute $x_U x_L P$ and S_L or S_U, where to compute S_S or S_U require the secret share x_U and x_L and session secret values u and l information which are not known to An adversary or malicious PKG.

Known key attack: If an adversary achieves a session key, where session key $sk = H(ID_U||ID_L||ulP||S_U)$. It does not mean that other session keys can compromise. Because, each session key involves independent short-term secret values u and l which are different for each session.

Forward secrecy If an adversary achieves entities private keys then,

- Half forward Secrecy: Compromise of the private key (x_U, SK_U) of user does not reveal previously established session keys because to achieve a session key, short time secret keys information is needed. Moreover, for given $\langle P, uP, lP\rangle$ to computation ulP is equivalent to CDH problem.
- *PKG forward secrecy:* Compromise of PKG master key does not reveal any information about session key because to achieve season key, the value (x_U, x_L) and (u, l) are needed, which can not be computed by using master key. Because, secret values (u, x_U) and (l, x_L) are randomly generate by L and U respectively. Moreover, for given $\langle P, x_UP, x_LP\rangle$ and $\langle P, uP, lP\rangle$ computation of x_Lx_UP and ulP are equivalent to CDH problem respectively.

Known session-specific temporary information attack: If short term secret keys u and l are compromised, then session keys does not reveals. Because, with short term secret keys u and l and given information $\langle Y_U = x_U\mathsf{mk}P, Y_L = x_L\mathsf{mk}P, Q_U, Q_L\rangle$ one can achieve S_U or S_L as:

$$e(uQ_L, Y_L)\cdot e(lQ_U, Y_U) = e(Q_L, P)^{x_L\mathsf{mk}u}\cdot e(Q_U, P)^{x_U\mathsf{mk}l}$$

However, to achieve $sk = H(ID_U||ID_L||luP||x_Lx_UP||S_L)$, the value x_Lx_UP is needed, where for given $\langle P, x_UP, x_LP\rangle$ computation of x_Lx_UP is equivalent to CDH problem.

Key off-set attack: When user send a message $\langle ID_U, T_U, PK_U\rangle$ to L. An adversary can replace it by $\langle ID_U, T_U^*, PK_U\rangle$ where $T_U^* = a^*T_U$. When, L computes

$$S_L^* = e(Q_U, P)^{lx_U\mathsf{mk}}\cdot e(Q_L, P)^{a^*ux_L\mathsf{mk}} \tag{1}$$

and achieves sk_1 and mac_1. L sends the message $\langle ID_L, T_L, P_L, mac_1\rangle$ to U, adversary again change the message and sends $\langle ID_L, T_{L_1}, P_L, mac_1\rangle$ where $T_L^* = a^*T_L = a^*lP$. U computes

$$S_U^* = e(Q_L, P)^{ux_L\mathsf{mk}}\cdot e(Q_U, P)^{a^*lx_U\mathsf{mk}} \tag{2}$$

Then, gets sk_1^* and mac_1^*. U, and concludes that $mac_1 \neq mac_1^*$ as by eq.(1) and eq. (2), $S_L^* \neq S_U^*$.

5 Conclusion

In this paper, we proposed a flexible and secure license distribution mechanism. In proposed mechanism, license server and user mutually authenticate each other and establish a session key, which ensures secure communication between them. Moreover, DRM principals communicate to each other using establish symmetric session key instead of public key, which is computationally feasible. Therefore, the proposed scheme is efficient and scalable for DRM system.

References

1. Al-Riyami, S.S., Paterson, K.G.: Certificateless public key cryptography. In: Laih, C.-S. (ed.) ASIACRYPT 2003. LNCS, vol. 2894, pp. 452–473. Springer, Heidelberg (2003)
2. Boneh, D., Franklin, M.: Identity-Based Encryption from the Weil Pairing. In: Kilian, J. (ed.) CRYPTO 2001. LNCS, vol. 2139, pp. 213–229. Springer, Heidelberg (2001)
3. Chatterjee, S., Sarkar, P.: Identity-based encryption. Springer-Verlag New York Inc. (2011)
4. Dutta, R., Barua, R., Sarkar, P.: Pairing Based Cryptographic Protocols: A Survey. Manuscript (2004), http://eprint.iacr.org/2004/064
5. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Transactions on Information Theory 22(6), 644–654 (1976)
6. Dutta, R., Mukhopadhyay, S., Dowling, T.: Key management in multi-distributor based DRM system with mobile clients using IBE. In: Second International Conference on the Applications of Digital Information and Web Technologies, pp. 597–602 (2009)
7. Dutta, R., Mishra, D., Mukhopadhyay, S.: Vector Space Access Structure and ID Based Distributed DRM Key Management. In: Abraham, A., Mauri, J.L., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) ACC 2011, Part IV. CCIS, vol. 193, pp. 223–232. Springer, Heidelberg (2011)
8. Dutta, R., Mishra, D., Mukhopadhyay, S.: Access policy based key management in multi-level multi-distributor DRM architecture. In: Joye, M., Mukhopadhyay, D., Tunstall, M. (eds.) InfoSecHiComNet 2011. LNCS, vol. 7011, pp. 57–71. Springer, Heidelberg (2011)
9. Hwang, S.O., Yoon, K.S., Jun, K.P., Lee, K.H.: Modeling and implementation of digital rights. Journal of Systems and Software 73(3), 533–549 (2004)
10. Liu, Q., Safavi-Naini, R., Sheppard, N.P.: Digital Rights Management for Content Distribution. In: Proceedings of Australasian Information Security Workshop Conference on ACSW Frontiers 2003, vol. 21 (January 2003)
11. Ku, W., Chi, C.H.: Survey on the technological aspects of digital rights management. Information Security, 391–403 (2004)
12. Sachan, A., Emmanuel, S., Das, A., Kankanhalli, M.S.: Privacy Preserving Multiparty Multilevel DRM Architecture. IEEE Consumer Communications and Networking Conference (CCNC 2009), 1–5 (2009)
13. Shamir, A.: Identity-Based Cryptosystems and Signature Schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)
14. Imai, H., Zheng, Y. (eds.): PKC 1998. LNCS, vol. 1431. Springer, Heidelberg (1998)
15. Yen, C.T., Liaw, H.T., Lo, N.W.: Digital rights management system with user privacy, usage transparency, and superdistribution support. Int. J. Commun. Syst., doi:10.1002/dac.2431
16. Zhang, Z.Y., Pei, Q.Q., Ma, J., Yang, L.: Security and trust in digital rights management: a survey. International Journal of Network Security 9(3), 247–263 (2009)

GAS: A Novel Grid Based Authentication System

Narayan Gowraj[1], Srinivas Avireddy[1], and Sruthi Prabhu[2]

[1] Department of Information Technology,
[2] Department of Computer Technology,
Anna University, Chennai-600044, India
{ngowraj,tholi1033,shruthe92}@yahoo.com

Abstract. With the evolving trends in technology, providing security for the users is an essential goal of the application. Authentication is one such important aspect of security which provides access control for the users of an application. The common method to provide authentication is by using a username/password pair .Graphical password authentication has proved to be more powerful and useful when compared to traditional textual password authentication. In this paper we propose a novel graphical password based authentication system called GAS(Grid based Authentication System). We focus our attention on the epigram, “It is easy to remember what we see rather than what we hear”. The methodology involves choosing a pattern called Auth Pattern which is formed by placing images in a given grid. We have chosen an optimal size for the grid as 8*8.Our proposed system considers very important parameters such as user memory and length of the password. Considering these parameters we compare our system with the state of the art authentication systems to prove our methodology’s efficiency.

Keywords: GAS, Grid system, authentication, dynamic, auth-pattern (AP).

1 Introduction

The increase in the usage of web applications has explicitly called for a secure authentication system which provides security for the users’ credentials against unauthorized access. Current authentication methods can be divided into three main areas: Knowledge based authentication, token based authentication and biometric based authentication [1] [2]. Knowledge based authentication techniques are most widely used and include text-based passwords either in the clear text format or cipher text format. Token based authentication involves issuing of a token to authorize a user. Credit card is an example for token based authentication technique. Fingerprints, iris scan, or facial recognition are examples of biometric based authentication. The major drawback of token based and the biometric based authentication methods are that these systems are expensive and require special devices which make them unfeasible for web application.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 578–590, 2013.

Textual passwords are the first choice for authentication by humans for any web application and hence it is essential to have strong and secure authentication processes. The traditional password schemes are vulnerable to many attacks. So in this paper we present a grid based authentication system which makes use of a pattern called Auth-Pattern (AP) to authenticate a user. This Auth-Pattern is created by using images from the 64 images provided to the user. In this case we use an 8*8 grid and hence 64 images. The size of the grid can be varying and hence the application manager can decide the size of the grid depending on the application and its purposes. The number of images in an AP is a variable and not a constant. The user can use a password which has a minimum length of 6 and maximum length of 32. If a user wants to register into the GAS (Grid Bases Authentication) system then the user has to enter a unique username which is in the plain text format and select an Auth-Pattern (AP) as the password. The Auth-Pattern may consist of images placed in various grid boxes thus forming a pattern for the user. The order of arrangement of the images in the Auth-Pattern formed by the user is considered as a metric so as to increase the security.

2 Related Work

Many related graphical authentication processes have been developed till date. Any graphical authentication system falls under these three broad categories: 1. Pure Recall based authentication, 2. Cued Recall based authentication and the 3. Recognition based authentication system. Pure Recall based authentication system requires the user to replicate their user credentials without any help or any reminder. Some of the pure recall based systems are DAS, Passdoodle and Qualitative DAS [3]. The Cued Recall based authentication system is very similar to the pure recall based authentication system except that it provides the user with a framework of hints, reminders and challenge/response questions to reproduce the password. Some of the cued recall based authentication systems are PassPoint [4], Pass-Go [5]. The Recognition based authentication system requires the user to reciprocate the pattern, select the images, spot the symbols etc which they used while registration and some of the powerful unique recognition based systems are WIW, Story etc [6] [7] [8].

Pure-Recall methods such as DAS (Draw A Secret) first proposed by Jermin et al, requires a user to draw the password on a 2D grid. The coordinates of this drawing on the grid are stored in order. During authentication user must redraw the picture to declare as an authenticated user. The user is authenticated if the drawing touches the grid in the same order. The major drawback of DAS is that diagonal lines are difficult to draw and difficulties might arise when the user chooses a drawing that contains strokes that pass too close to a grid-line. Users have to draw their input sufficiently away from the grid lines and intersections in order to enter the password correctly. Another important method in this model is the Passdoodle [9] [10] which is also a pattern and a doodle is considered as a full match if it is drawn in exactly the same order as when the user initially drew the passdoodle, and is considered as a visual match if it is not a full match due to stroke order, stroke direction, or number of strokes. They found that the order in which a password is drawn introduced much complexity to graphical passwords and suggested to neglect the order. Cued-Recall based authentication systems such as PassPoints proposed by Wiedenbeck et al in which passwords could be composed of several points on an image and the user is required to reciprocate the point correctly on the image given to authenticate as a valid user. Some of the other Cued-Recall methods include Pass-Go and Passmap etc.

The Snake and Ladder authentication system [8] is very similar to our proposed system. Any authentication system consists of three steps: Password registration, password entry and password verification. In the Snake and Ladder authentication method the user registers by remembering a sequence of grid cells or ladders or snakes for their password but the password being generated for the sequence is always static and not dynamic and the order of arrangement does not matter in the snake and ladder system which makes it less secure than the GAS system.

The paper is organized as follows. Section 3 deals with the proposed system architecture. Section 4 deals with the Methodology of the proposed system. Section 5 deals with the mathematical analysis. Section 6 and 7 deals with the evaluation of attacks and experimental results respectively. Finally section 8 has the conclusion.

3 System Architecture

The architecture of the authentication system consists of three tiers as shown in Fig.1. The three logical tiers are the user interface tier, application server tier

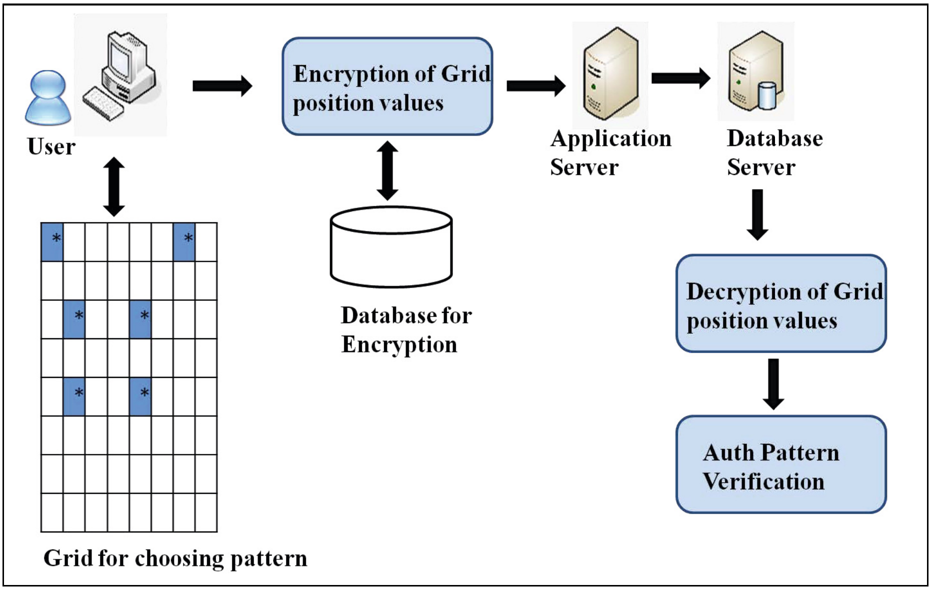
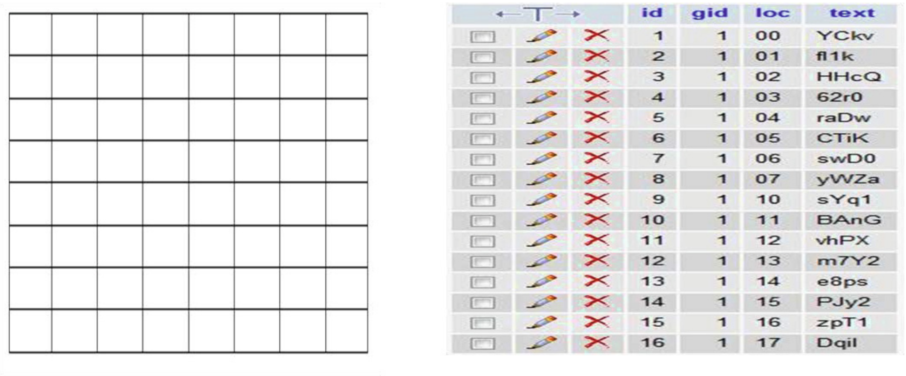

User
Encryption of Grid position values
Application Server
Database Server
Database for Encryption
Decryption of Grid position values
Auth Pattern Verification
Grid for choosing pattern

Fig. 1. System Architecture

and the database server tier. The user interface tier provides the user with the Graphical user interface (GUI) where the user interacts with the web browser to register or to validate his credentials to authenticate as a valid user which is the first step of the authentication system as discussed earlier. The user interface is the layer which allows the user to provide their user credentials. In this case, the users' username is in the plain text format and the Auth-Pattern is the password.

The user interface in the GAS system is a grid consisting of 64 boxes which is shown in Fig.3 and the user is provided with 64 images to create an Auth-Pattern as the password which is shown in Fig.2. The user has to insert the images into random locations in the grid and create a unique Auth-Pattern (AP) with the arrangement of the images as a metric in his password. The user interface is mainly created using web development languages like html, xml etc and the validation is provided using JavaScript. The user interface then communicates with the application server tier. This is accomplished by means

Fig. 2. Images for the Grid



←T→			id	gid	loc	text
□	✎	✗	1	1	00	YCkv
□	✎	✗	2	1	01	fl1k
□	✎	✗	3	1	02	HHcQ
□	✎	✗	4	1	03	62r0
□	✎	✗	5	1	04	raDw
□	✎	✗	6	1	05	CTiK
□	✎	✗	7	1	06	swD0
□	✎	✗	8	1	07	yWZa
□	✎	✗	9	1	10	sYq1
□	✎	✗	10	1	11	BAnG
□	✎	✗	11	1	12	vhPX
□	✎	✗	12	1	13	m7Y2
□	✎	✗	13	1	14	e8ps
□	✎	✗	14	1	15	PJy2
□	✎	✗	15	1	16	zpT1
□	✎	✗	16	1	17	Dqil

Fig. 3. Grid on the left and Cipher text for locations in the grid on the right

of ordinary WWW technologies, HTTP redirects [16], URL query strings, and cookies [17].

The next layer is the database server tier which has the application server which performs the authentication of a user. Based on the result of the authentication the application server grants or rejects the user to have privileges to access the application. The database server tier consists of multiple databases, each database consists of the encrypted grid table which is shown in the Fig.3 and the user credentials stored as per the Auth-Pattern (AP) chose by the user at the time of registration. Fig.3 show the cipher text of 16 boxes with id as the primary key, gid as the grid id, loc as the location of the grid location. The grid location consists of two numbers where the first number represents the row and the second number represents the column. The text represents the cipher text of each location. Fig.3 is just a sample of the database showing only 16 locations out of the 64 locations in the case of an 8*8 grid.

4 Methodology

When the user tries to login into an application, the first step involved in the authentication procedure is the dynamic generation of a 8x8 grid with random face values for each grid position which is not seen by the user and each grid position is given a ID .The generated grid is sent to the client GUI .The client provides his username which is in the plain text format and has to place the images in the grid positions which he chose while registration. Once the images are placed in the respective positions, the encrypted random face values of each grid position in the pattern are concatenated in the order the user placed. This random string along with the grid ID is sent to the Application Server .The application server then forwards it to the Database Server for further verification to provide authentication of the user.

The database server performs two functions.The first function is to decrypt the encrypted random face values and the second function is to perform the verification of the Auth Pattern.During the first step, random string gets mapped to the respective positions and the sequence of positions separated by a delimiter is concatenated so as to form the Auth-Pattern. During the second step,this auth-pattern is checked for correctness with the Auth-Pattern stored in the persistent database during registration. If the auth pattern, is matched the user is authenticated. The grid is now removed from the database. When the user tries to login again, a new dynamic grid gets generated at the server and stored in the database. So at any point in time only one grid is being stored in the database and when the authentication is over it is removed. This minimizes the spatial overhead. Fig.4 explains our novelty using a simple 3x3 grid to authenticate users who are already registered. We have considered a 3x3 grid as an example. The application manager can decide upon the size of the grid depending on the application.

The grid is then given to the client GUI for the user to select the correct pattern to login. Let us consider the correct pattern to be the positions 00, 01

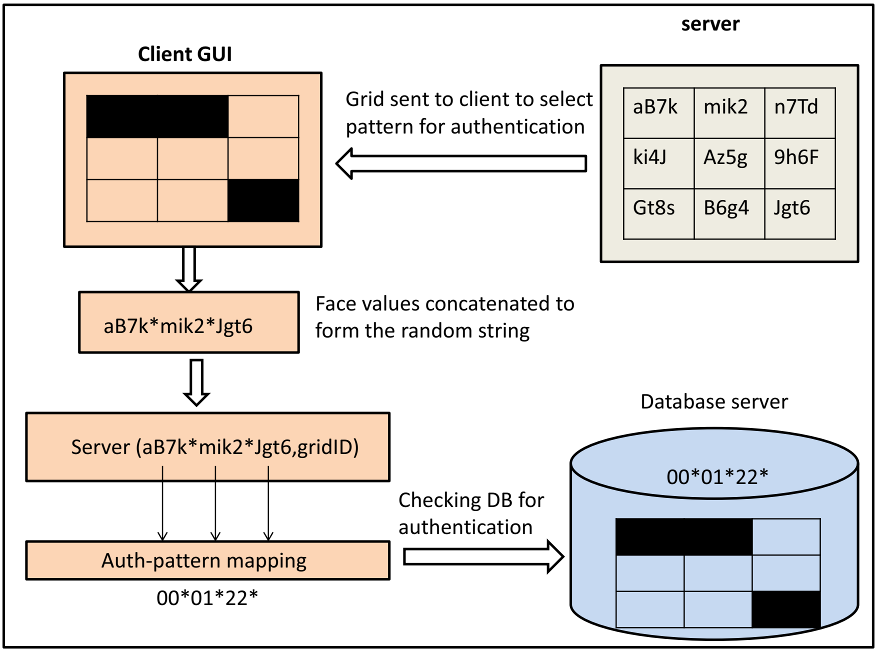
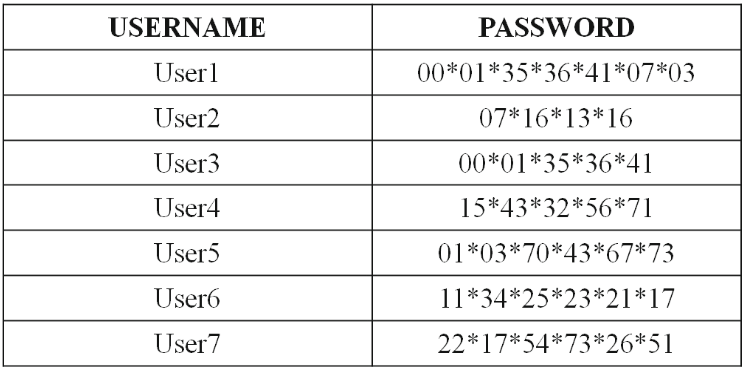

server
Client GUI
Grid sent to client to select pattern for authentication
aB7k
mik2
n7Td
ki4J
Az5g
9h6F
Gt8s
B6g4
Jgt6
aB7k*mik2*Jgt6
Face values concatenated to form the random string
Database server
Server (aB7k*mik2*Jgt6,gridID)
00*01*22*
Checking DB for authentication
Auth-pattern mapping
00*01*22*

Fig. 4. Methodology of the GAS System

and 22 chosen in order. Now, when the user selects these positions 00, 01 and 22 the face value of these positions aB7k, mik2 and Jgt6 gets concatenated as aB7k* mik2* Jgt6 .Here each grid position is associated with a randomly generated four character string. So even when someone in the middle between the client and the server gets access to the random string he may not be able to crack the password which is a sequence of positions and not the random string associated with it. In the server side, the random string sent gets mapped to the respective grid positions and the Auth-Pattern 00*01*22* is generated. The auth-pattern is checked with the database. If the pattern is correct the user logs into the database, else he needs to retry. During the retry, a new grid gets generated and the old grid used will be deleted from the server, thereby improving the overhead in space complexity. The user credentials are stored in the database where the username exists in the clear text format and the password is based on the pattern the user chooses which is understandable only by the database administrator. This form of cryptic representation is useful for password recovery in case the user forgets his or her password. In systems like DAS, Passpoint and Passdoodles password recovery is not an option whereas in GAS password recovery is possible. Fig.5 shows the username and password for a few users which are stored in the database.



USERNAME	PASSWORD
User1	00*01*35*36*41*07*03
User2	07*16*13*16
User3	00*01*35*36*41
User4	15*43*32*56*71
User5	01*03*70*43*67*73
User6	11*34*25*23*21*17
User7	22*17*54*73*26*51

Fig. 5. Encryption of the Password in the Database Server

The database contents are accessed and are understood only by the database administrator and the contents such as the field names are encrypted using the encryption algorithm MD5 [18] for security concerns. Once the user is registered the user can get access into the GAS system using his legal credentials. The credentials submitted by the user at the time of login are transferred to the application tier which has the application server through http requests. The password transferred is a dynamic password which gets changed everytime the user logins into the GUI. Even if a malicious user gets illegal access to the password by using the man in the middle attacks [19] [20] or the replay attacks [21] he will get access only to the dynamic password and not the Auth-Pattern (AP).

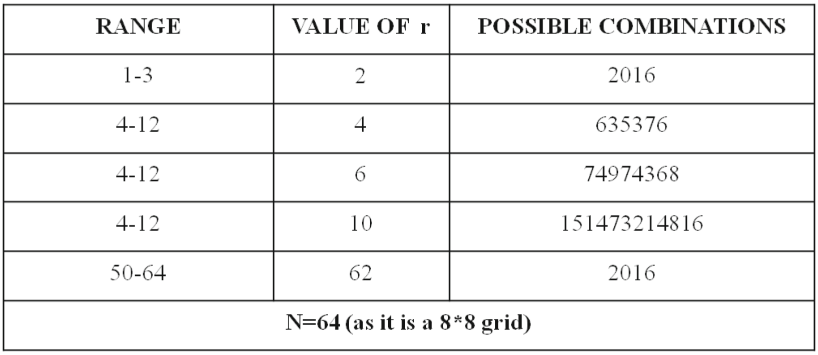
5 Mathematical Analysis

The mathematical analysis is based on 2 parameters. They are the length of the Auth Pattern and size of the grid. The Auth pattern length is constrained by user memory. Users cant remember lengthy pattern. On the other hand, short patterns compromises security.The size of the grid is constrained by the complexity involved in the storage and retrieval. To solve this issue, we formulate a optimization problem.

Optimization problems are of three categories. They are continuous, combinatorial and NP optimization problems [22]. We choose the continuous optimization problem because we have a basic objective function followed by a set of inequality constraints. We formulate the objective function F(x), so as to maximize the security of the proposed system considering the parameters discussed above. The 2 parameters are formulated as constraints.

$$i \leq G(x) \geq j \tag{1}$$

$$k \leq H(x) \geq l \tag{2}$$



RANGE	VALUE OF r	POSSIBLE COMBINATIONS
1-3	2	2016
4-12	4	635376
4-12	6	74974368
4-12	10	151473214816
50-64	62	2016
N=64 (as it is a 8*8 grid)		

Fig. 6. Possible Combinations of Various Ranges of Auth-Pattern

The constraints G(x) and H(x) represent the length of Auth Pattern and Size of the Grid. The values i and j are determined keeping user memory and security in mind. Hence we choose optimal length of the auth pattern between 4 and 12. So the value of i is 4 and j is 12. This is based on the following analysis. Using 50 images for your Auth-Pattern is an overhead and on the other hand using very few images like 2-3 compromises security. So, it would be feasible to choose 4 to 12 images for your Auth-Pattern. Using less number of images (lesser than 4) will make the Auth-Pattern less secure and using large number of images (ranging from 30 to 64) can make the Auth-Pattern difficult to memorize. This can be mathematically proved using the combinations formula quoted in the equation 3.

$$C(n,r) = n!/r!(n-r)! \tag{3}$$

Now Fig.6 represents the various combinations for an 8*8 grid pattern where n has a constant value of 64. It is very evident from the result given in the Fig.6 that Auth-Pattern with number of images ranging from 4 to 12 is very secure that those which has ranges between 1-3 and 50-64. Fig. 7 shows the graph

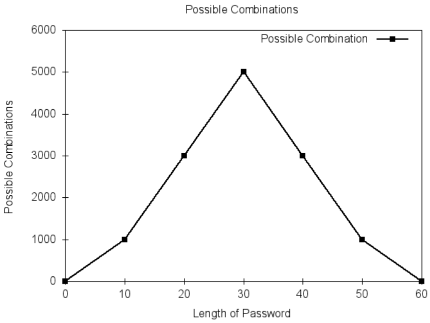


Fig. 7. Possible Combinations

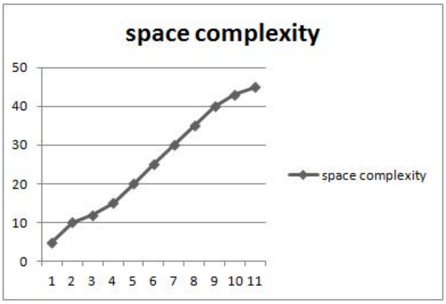


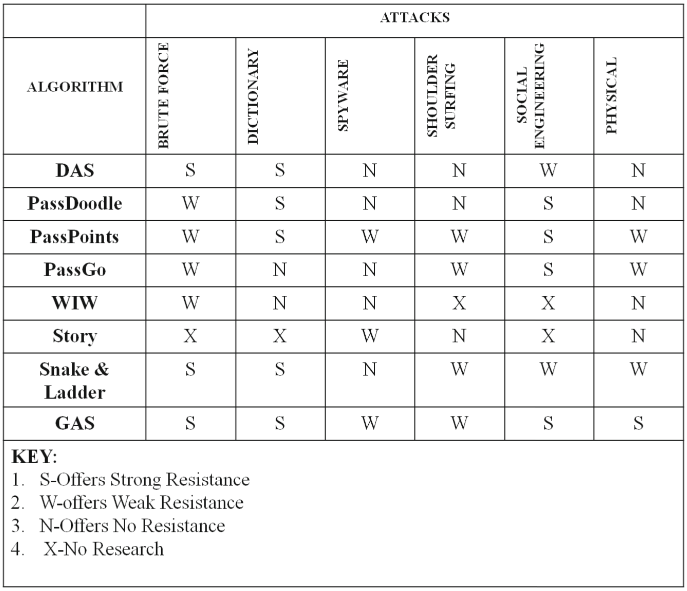
Fig. 8. Space Complexity of Various Grid Sizes

obtained from the results shown in Fig.6 which clearly indicates that any Auth-Pattern with the number of images ranging from 4 to 12 is very secure, easy to remember and incontrovertible towards password breaching. Auth-Patterns which has its length in the range of 20-40 are also very secure but very difficult to memorize, hence we go for small Auth-Patterns which are very secure and easy to remember.Since the complexity of storing retrieving the grid becomes difficult as we increase the grid size we choose values of k as 8 and l as 10 for improved security and less complexity. Fig.8 depicts how the space complexity as the size of the grid increases.

6 Evaluation against Attacks

To prove the efficiency of the proposed system under graphical password attacks, we tested our system under six most common graphical password attacks and compared it with the other state of the art authentication systems. The attacks which we consider are brute force attack, dictionary attack, spyware attack, shoulder surfing attack, social engineering and physical attack [11] [12] [13] [14] [15].

Brute force attack uses an algorithm which makes use of all possible combinations to breach the password. This is used in situations when the malicious user has no clue or hints to breach the password. Generally time consuming. Dictionary attack method uses all the words in the dictionary to check if the passwords used by the user matches any of the words in the dictionary. It is



	ATTACKS					
ALGORITHM	BRUTE FORCE	DICTIONARY	SPYWARE	SHOULDER SURFING	SOCIAL ENGINEERING	PHYSICAL
DAS	S	S	N	N	W	N
PassDoodle	W	S	N	N	S	N
PassPoints	W	S	W	W	S	W
PassGo	W	N	N	W	S	W
WIW	W	N	N	X	X	N
Story	X	X	W	N	X	N
Snake & Ladder	S	S	N	W	W	W
GAS	S	S	W	W	S	S

KEY:
1. S-Offers Strong Resistance
2. W-offers Weak Resistance
3. N-Offers No Resistance
4. X-No Research

Fig. 9. Resistance Provided by different GUA techniques

not feasible for passwords with alphanumeric keys and using dictionary attack on GUA (Graphical User Authentication) is just a waste of time. Spyware attack uses applications and tools to record sensitive data movement such as mouse movement, mouse clicks and key press incurred on the keyboard. Shoulder attack majorly depends on the window size and resolution. Key Loggers are included in this type of attack. This attack usually involves surreptitiously looking at the users credentials without the users notice. As the name suggests, it involves looking over a persons shoulder. Social engineering attack happens when a non authorized personal (malicious user) manages to impersonate sensitive data such as user credentials, codes etc from authorized employees. The attacker interacts with unsuspecting employees and gathers as much information they can to gain access to the protected data. Physical attack happens when hackers or malicious users get direct access to the data present in the server or to the database contents. This type of attack usually involves bypassing of user credentials to get unauthorized access.

Thus these are the major attacks which can compromise the authentication system present in a web application. Fig.9 shows a comparative study of the various attacks on the various methodologies already present and our novel algorithm which is the GAS system.

The results thus obtained are plotted in the form of graph as shown in the Fig.10. Out of six attacks considered for evaluation of our system it is found that our system is efficient enough to strongly resist four of them. The x-axis of the graph refers to the algorithm considered and y-axis refers to the number of attacks resisted categorically such as strongly resisted or weakly resisted or not

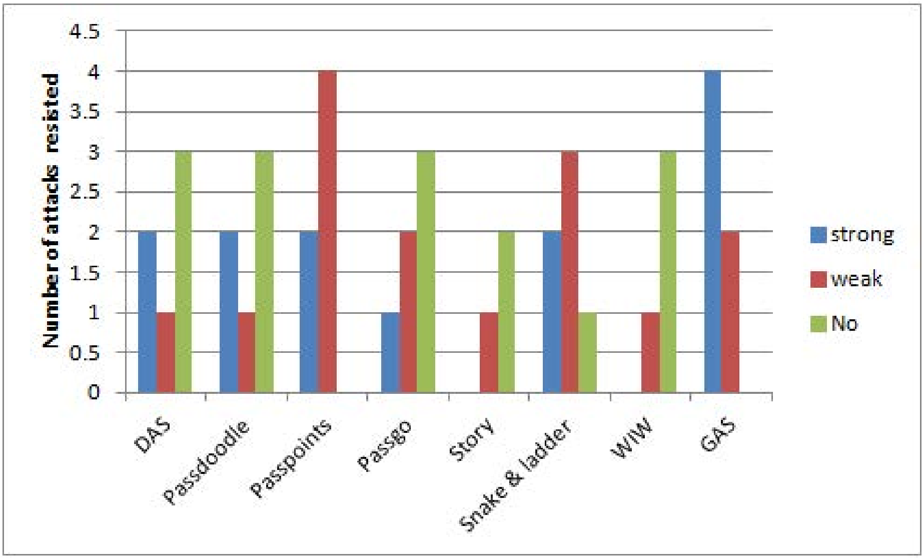


Fig. 10. Resistance Againt various Attacks

resisted. It is clearly found that our system has resisted better than the other existing algorithms.

7 Experimental Results

The experiment to test the usability of the proposed system was done by selecting 100 students. The students were exposed to use the Snake and ladder system and our proposed GAS system. They were asked to set passwords of length 6,8 and 10. The memorability of these passwords were tested in both the systems after the first week of setting the password and tabulated in Fig.12 and plotted in Fig.11. It is found that, it is easy to remember passwords set using our system than the snake and ladder system. Another important finding is that as the length of the password increases the memorability of the password decreases in case of both the systems.

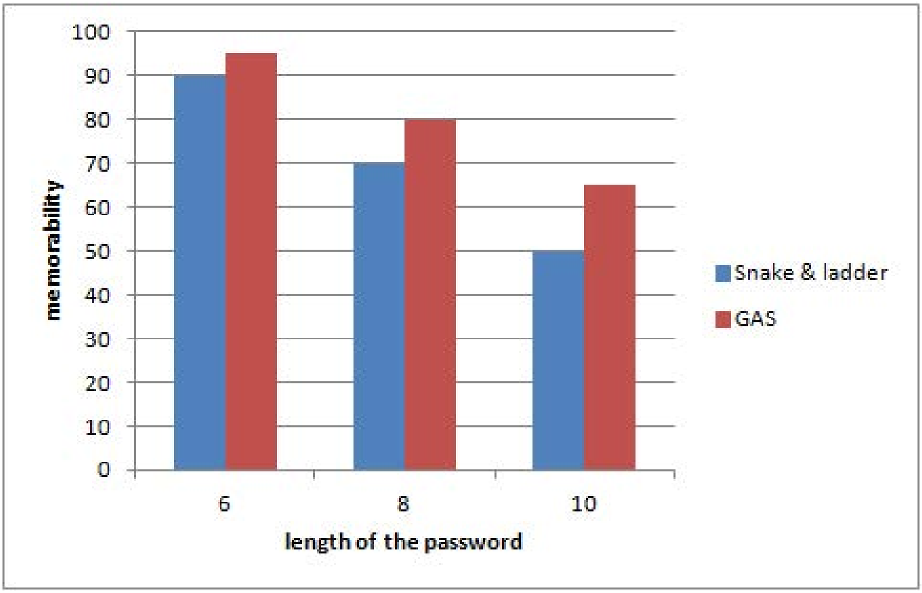
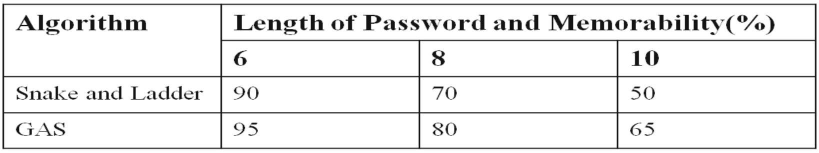


Fig. 11. Comparison of Memorability



Algorithm	Length of Password and Memorability(%)		
	6	8	10
Snake and Ladder	90	70	50
GAS	95	80	65

Fig. 12. Table for Comparison of Memorability

8 Conclusion

The research problem was to design and implement an authentication system which satisfies three goals. 1. Easy memorization of passwords: It has extremely become very difficult to memorize large number of usernames and passwords, so we decided to use a system where the password is given in the form of an image pattern which we call as Auth-Pattern. The system is primarily based on the Mnemonic principle [15] which states that "It is easy to remember what we see rather than what we hear". 2. Secure transmission of passwords: For this purpose we have designed the GAS authentication system is such a way that it produces random and encrypted passwords for every login the user makes and this random passwords gets transferred through the http requests . 3. Password Recovery in graphical authentication methods to make the authentication process more efficient and flexible to the users using it.

References

1. Chiasson, S., Stobert, E., Forget, A., Biddle, R., van Oorschot, P.C.: Persuasive Cued Click-Points: Design,Implementation, and Evaluation of Knowledge-Based Authentication Mechanism. IEEE Transactions on Dependable and Secure Computing 9(2) (March-April 2012)
2. Dass, S.C., Zhu, Y., Jain, A.K.: Validating a Biometric Authentication System: Sample Size Requirements. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(12) (December 2006)
3. Almuairfi, S.: IPAS: Implicit Password Authentication System. In: 2011 Workshops of International Conference on Advanced Information Networking and Applications
4. Salehi-Abari, A., Thorpe, J., van Oorschot, P.C.: On Purely Automated Attacks and Click-Based Graphical Passwords. In: 2008 Annual Computer Security Applications Conference
5. Tao, H.: Pass-Go, a New Graphical Password Scheme
6. Man, S., Hong, D., Matthews, M.: A Shoulder-Surfing Resistant Graphical Password Scheme WIW
7. Martinez-Diaz, M., Martin-Diaz, C., Galbally, J., Fierrez, J.: A Comparative Evaluation of Finger-Drawn Graphical Password Verification Methods. In: 2010 12th International Conference on Frontiers in Handwriting Recognition (2010)
8. Ma, Y., Feng, J.: Evaluating Usability of Three Authentication Methods in Web-Based Application. In: 2011 Ninth International Conference on Software Engineering Research, Management and Applications
9. Bicaki, K.: Towards Usable Solutions to Graphical Password Hotspot Problem. In: 2009 33rd Annual IEEE International Computer Software and Applications Conference
10. Malempati, S., Mogalla, S.: A Well Known Tool Based Graphical Authentication Technique. In: CCSEA 2011, pp. 97–104 (2011)
11. Doja, M.N., Kumar, N.: Image Authentication Schemesagainst Key-Logger Spyware. In: Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing
12. Hu, W., Wu, X., Wei, G.: The Security Analysis of Graphical Passwords. In: 2010 International Conference on Communications and Intelligence Information Security

13. Zhao, H., Li, X.: S3PAS:A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme. In: 21st International Conference on Advanced Information Networking and Applications Workshops, AINAW 2007 (2007)
14. Gao, H., Ren, Z., Chang, X., Liu, X.: A New Graphical Password Scheme Resistant to Shoulder-Surfing. In: 2010 International Conference on Cyberworlds
15. Zhao, S., Aggarwal, A., Kent, R.D.: PKI-Based Authentication Mechanisms in Grid Systems. In: International Conference on Networking, Architecture, and Storage, NAS 2007 (2007)
16. Tappenden, A., Miller, J.: A Three-Tiered Testing Strategy for Cookies. In: 2008 International Conference on Software Testing, Verification, and Validation
17. Juels, A., Jakobsson, M., Jakobsson, M.: Cache Cookies for Browser Authentication. In: Proceedings of the 2006 IEEE Symposium on Security and Privacy, S and P 2006 (2006)
18. Jrvinen, K., Tommiska, M., Skytt, J.: Hardware Implementation Analysis of the MD5 Hash Algorithm. In: Proceedings of the 38th Hawaii International Conference on System Sciences (2005)
19. Chen, Z., Guo, S., Duan, R., Wang, S.: Security Analysis on Mutual Authentication against Man-in-the-Middle Attack. In: The 1st International Conference on Information Science and Engineering, ICISE 2009 (2009)
20. Alicherry, M., Keromytis, A.D.: DoubleCheck: Multi-path Verification Against Man-in-the-Middle Attacks. IEEE (2009)
21. Rahman, K.A., Balagani, K.S., Phoha, V.V.: Making Impostor Pass Rates Meaningless: A Case of Snoop-Forge-Replay Attack on Continuous Cyber-behavioral Verification with Keystrokes. IEEE (2009)
22. http://en.wikipedia.org/wiki/Optimizationproblem

Multiplicative Watermarking of Audio in Spectral Domain

Jyotsna Singh, Parul Garg, and Alok Nath De

Division of Electronics and Communication Engineering,
Netaji Subhas Institute of Technology,
Sector 3, Dwarka, New Delhi 110075, India

Abstract. In this paper, watermark is multiplicatively embedded in discrete fourier transform magnitude of audio signal using spread spectrum based technique. A new perceptual model for magnitude of discrete fourier transform coefficients is developed which finds the regions of highest watermark embedding capacity with least perceptual distortion. Theoretical evaluation of detector performance using correlation detector and likelihood ratio detector is undertaken under the assumption that host feature follows Weibull distribution. Also, experimental results are presented in order to show the performance of the proposed scheme under various attacks such as presence of multiple watermarks, additive white gaussian noise and audio compression.

Keywords: audio, correlation detector, discrete fourier transform, log-likelihood detection, watermarking.

1 Introduction

Various watermarking embedding techniques have been proposed which embed watermark additively or multiplicatively in audio signal using the imperfections of human auditory system (HAS). These techniques explore the fact that the HAS is insensitive to small amplitude changes, either in the time [1]-[3] or frequency [4], [5] domains. Boney [1] generated the watermark by filtering a PN-sequence with a filter approximating the frequency masking characteristics of the human auditory system (HAS). This filtered watermark was then weighted in the time domain to account for temporal masking. Swanson [2] proposed audio dependent watermarking procedure which directly exploited temporal and frequency masking properties to guarantee that the embedded watermark is inaudible and robust. The shaping of watermark is performed using a masking curve computed on the original signal. This masking curve is obtained by psychoacoustic modeling of host audio signal. Bassia [3] presented an audio watermarking algorithm by adding a perceptually shaped spread-spectrum (SS) sequence in time domain.

In the other category watermark is embedded in frequency domain. Cox [4] suggested that a watermark should be constructed as an independent and identically distributed (i.i.d.) gaussian random vector that can be imperceptibly

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 591–605, 2013.

inserted in the perceptually most significant spectral components of the data. An audio watermarking scheme based on frequency-selective spread spectrum (FSSS) technique in combination with the subband decomposition of the audio signal was presented by Malik et. al. [5]. Megias [6], Fujimoto [7] and Fallahpour [8] developed a high bit-rate audio watermarking technique with robustness against common attacks and good transparency. These algorithms are based on spline interpolation technique.

The embedding techniques in [4], [5], exploit psychoacoustic characteristics of HAS while embedding the watermark additively or multiplicatively in spectral domain. These techniques explored the fact that HAS is insensitive to small amplitude changes in spectral domain. Whereas, phase discontinuity of an audio signal causes perceptible distortion when the phase relation between each frequency component of the signal is changed. Hence Discrete Fourier Transform (DFT) magnitude would be a better option for inserting watermark. However, in literature no perceptual model is defined for DFT magnitude which can decide the location and strength of watermark to be embedded in audio spectrum. Also, these techniques have two major drawbacks. First, the psychoacoustic modeling used by existing techniques require rigorous complex computations. Second, the watermark embedding capacity of these schemes is low i.e. there is not much space to accommodate watermark in the host feature within the defined perceptual limits.

To overcome these two problems, a new method of evaluating masking threshold for DFT magnitude is proposed which requires lesser computations as compared to traditional psychoacoustic model based thresholds. The technique finds best possible locations in spectra for watermark embedding and finds scaling factor of watermark to achieve high watermark embedding capacity.

In this paper, we present a blind robust watermarking system based on pseudorandom signals embedded in the magnitude of the DFT coefficients of an audio signal. Blind detection systems requires only the secret key for data extraction or detection. The cover data or watermark is not required during the extraction process. The scheme obviates the use of complex HAS calculations. Also, it allows us to build a model which can decide the location and strength of watermark in DFT spectra. The paper is organized as follows.

The watermarking system model is presented in section II. In the next section, the signal model is presented and the distribution of DFT magnitude coefficients is shown. Then, in section IV, the construction of the optimal detector is depicted. In sections V and VI, the experimental results and the conclusions are presented.

2 Description of Watermarking Model

A watermarking system encompasses three major functionalities, namely, watermark generation, watermark embedding, and watermark detection. The aim of watermark generation is to construct a sequence $\mathbf{W}$ using an appropriate function f. Hence the watermark vector $\mathbf{W} = [W(0), W(1), \cdots, W(N-1)]$, such that $W(i) \in \mathcal{R}$, where $\mathcal{R}$ is real number, is given as

$$\mathbf{W} = f(\mathbf{K}, N) \tag{1}$$

here $\mathbf{K}$ is the watermark key, N is the length of watermark. Watermarked feature $\mathbf{F}'$ is obtained by multiplicatively embedding watermark $\mathbf{W}$ in host feature $\mathbf{F}$ given as

$$\mathbf{F}' = \mathbf{F}(1 + a\mathbf{W}) \tag{2}$$

here $\mathbf{F}' = [F'(0), F'(1), \cdots, F'(N-1)]$ and a is the scaling factor lying between 0 and 1. The scaling factor is introduced to maintain imperceptibility of the distortions caused to the host signal due to watermarking.

Watermark detector is used to examine whether the signal under test $\mathbf{F}_t$ contains a watermark $\mathbf{W}$ or not under a binary-decision hypothesis test framework. Each module is now discussed in detail, in the following subsections.

2.1 Watermark Generation

The steps required for generation of watermark are as follows:

- To construct watermark $\mathbf{W}$, a white PN or pseudo-random noise sequence $\mathbf{W}_0$ is generated such that $\mathbf{W}_0 = [W_0(0), W_0(1), \cdots, W_0(N_w - 1)]$, where $W_0(i) \in (-1, 1)$. The sequence is generated using secret key $\mathbf{K}$ such that they are mutually independent with respect to the host signal.
- The magnitude nature of host feature needs to be preserved implying that $\mathbf{F}'$ given in (2), should always be greater then zero. Such condition is obtained when $aW(i)'s \; \forall \; 0 \le i \le N-1$ take the value in the finite interval [-1,1] keeping scaling factor $a \le 1$.
- The N point DFT region hosting the watermark is usually split in number of subregions, which in our case are the critical bands. The start location (m) and end location (n) of watermark embedded in these critical bands is decided by a pre-defined masking threshold. Hence the length of watermark N_w is evaluated as
$$N_w = \lceil (n-m)N \rceil \tag{3}$$
- To maintain the symmetry of DFT magnitude a reflected version of $\mathbf{W}_0$ is required to be generated as
$$W_0'(i) = W_0(N_w - i - 1), \qquad 0 \le i \le N_w - 1 \tag{4}$$
The reflected chip $\mathbf{W}_0'$ is embedded in the frequency components around coefficient $N-1$. This is essential to obtain real valued audio in time domain.

2.2 Masking Threshold for DFT Magnitude

In this paper, the magnitude of DFT coefficients of host audio signal are modified by adding watermark, such that the modified spectra is always below the predefined masking threshold, termed as maximum amplitude spread (MAS).

The MAS is defined as the maximum of all amplitude spreads (AS) of DFT components at a particular frequency location within a frame. Following steps are involved to find MAS.

STEP-I: **Finding Amplitude Spread (AS)**
The AS of DFT components is evaluated from the energy spreading function given by Schroeder [9] and its effect is seen at all the N frequency locations of a frame. Schroeder presented a real nonnegative energy spreading function which approximated the basilar spreading as a triangular spreading function and is given as

$$\begin{aligned} SF_{dB}(i,j) &= 15.81 + 7.5(\varDelta_z + 0.474) \\ &\quad -17.5\sqrt{1 + (\varDelta_z + 0.474)^2} \end{aligned} \tag{5}$$

here $SF_{dB}(i,j)$ is the energy spread in decibels (dB) from i^{th} to j^{th} frequency location. The bark separation between these two points is given as $\varDelta_z = z_j - z_i$, where z_i and z_j denote the bark frequencies of i^{th} and j^{th} frequency locations respectively.
Let the audio signal $\mathbf{s}$, given by (6), is sampled at frequency, f_s Hz, is given as

$$\mathbf{s} = [s(0), s(1) \cdots, s(N-1)] \tag{6}$$

The k^{th} component of DFT, $S(k)$, of signal $s(n)$ is given as

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi kn/N} \tag{7}$$

The samples of discrete time signal $s(n)$ is recovered using the $IDFT$ of $S(k)$ as,

$$s(n) = \frac{1}{N} \sum_{k=0}^{N-1} S(k) e^{j2\pi nk/N} \tag{8}$$

Since audio is real valued signal, its DFT will satisfy the symmetry property i.e. $S(k) = S(N-k)^*$, where $k = 1, ..., N/2 - 1$. The DFT coefficients $S(k)$ corresponds to frequencies f_k given as

$$f_k = f_s \times k/N\,, \tag{9}$$

here $0 \le k \le N-1$, N being a power of 2. Considering the duplication in the spectra for $k \ge N/2$, we evaluate the masking spread $A_1(i,j)$ for amplitude of $N/2$ components only, given as

$$A_1(i,j) = \sqrt{SF(i,j)}\,, \qquad 0 \le i \le N/2 - 1 \tag{10}$$

where $SF(i,j)$ is the inverse decibel of $SF_{dB}(i,j)$. The square root is to convert the masking spread from energy scale to amplitude scale. Now respecting the symmetry property of DFT components, we define $A(i,j)$ as,

$$A(i,j) = \begin{cases} A_1(i,j), & 0 \le j \le N/2 \\ A_1(i,N-j), & N/2 + 1 \le j \le N-1 \end{cases} \tag{11}$$

The amplitude spread of i^{th} DFT component is then defined as,

$$A'(i,j) = |A(i,j)S(i)|,\ 0 \le i \le N/2-1\,, 0 \le j \le N-1\,, \tag{12}$$

where $S(i)$ is given by (7). This gives $N/2 \times N$ matrix showing amplitude spread of each of the $N/2$ DFT components at N frequency locations. Figure 1 shows a plot of Amplitude spread $A'(i,j)$ of $i = 17^{th}$ and 20^{th} frequency components at all the frequency location f_j for $0 \le j \le N-1$ given in (9) where $N = 512$ and $fs = 44.1kHz$.

STEP II: Evaluation of Maximum Amplitude Spread

The amplitude spreads of neighboring DFT components overlap each other. Maximum amplitude spread (MAS) is the maximum of all the overlapping amplitude spreads at f_i frequency due to DFT coefficients $S(j)$, $\forall\ 0 \le j \le N/2-1$ and $j \ne i$. MAS, $Y(i)$, at location i can therefore be evaluated as

$$Y(i) = \max(A'(i,j)) \quad for \quad 0 \le j \le N/2-1 \tag{13}$$

Now maximum amplitude spread MAS for a critical bands z will be the minimum of all $Y(i)$ in that critical band. From (13), we evaluate the Maximum Amplitude Spread $Y(z)$ for critical bands $z = 1, 2, \cdots, z_t$ as

$$Y(z) = \min(|Y(i)|) \quad for \quad LB_z \le i \le HB_z\,, \tag{14}$$

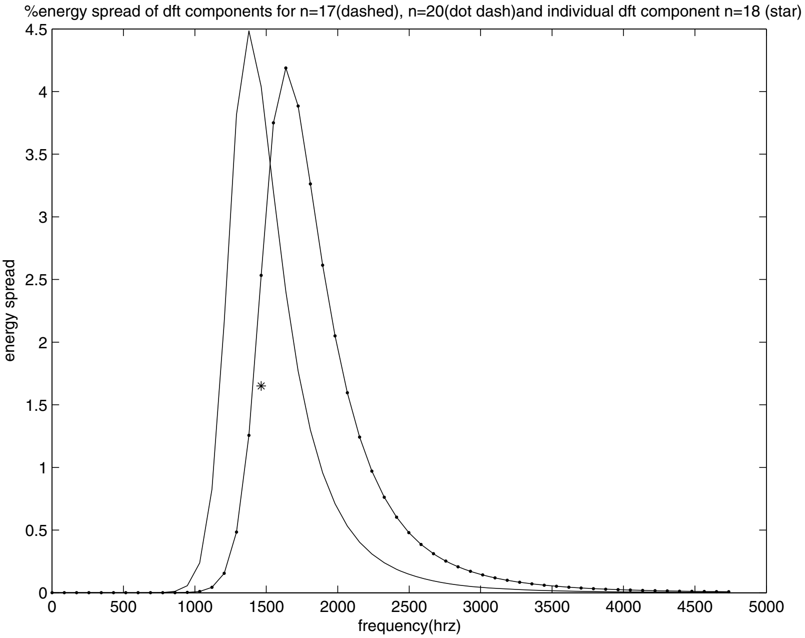


Fig. 1. Overlapping of Amplitude Spread of 17^{th} and 20^{th} DFT components and magnitude of 19^{th} DFT component

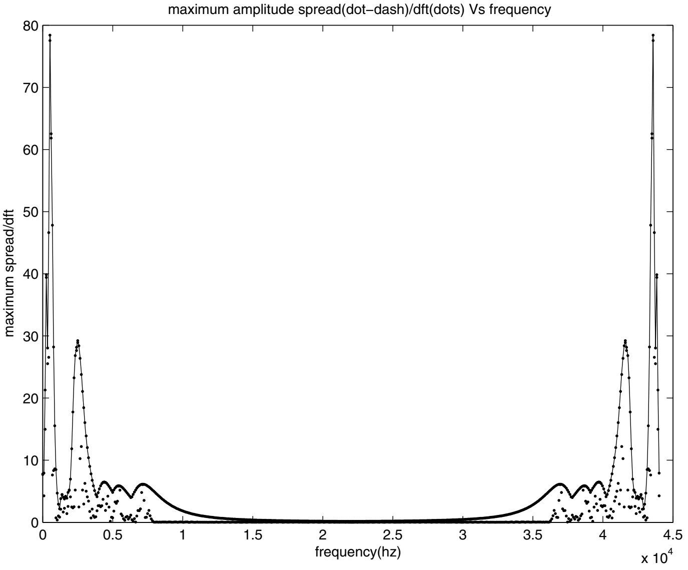


Fig. 2. Maximum amplitude spread of DFT magnitude for a given audio of frame length $N = 512$

where LB_z and HB_z are lower and upper frequency components of z^{th} critical band. Figure 2 shows the plot between maximum amplitude spread $Y(i)$ and the magnitude of DFT coefficients $F(i)$ at all the frequency locations f_i for $i = 0, 1, ..., N - 1$.

2.3 Watermark Embedding

In watermark embedding the watermark **W** is added to host signal **F** in a way that the symmetry of **F** is not disturbed. Also, the DC component and Nyquist component of DFT spectrum should remain unchanged. This is essential in order to retrieve real valued audio signal after watermarking process. The magnitude of DFT coefficients of host audio signal are modified by multiplicative watermarking, such that the modified spectra is always below the maximum amplitude spread of original signal. Hence, the DFT magnitudes are modified only in certain critical bands to maintain the transparency of audio signal. The embedding steps are described as follows

– The magnitude $F(k) = |S(k)|$ and phase $\phi(k) = \angle S(k)$ of the spectral coefficients are evaluated for $k = 0, 1, \cdots, N - 1$, where $S(k)$ is given by (7).
– The distribution of magnitude of DFT coefficients per critical band $F_z(k)$, for $LB_z \leq k \leq HB_z$ is found by translating frequency into bark scale. Here $z = 1, 2, \cdots, z_t$ are the critical bands, z_t is total number of critical bands and LB_z and HB_z are the respective lower and higher frequencies in the critical band z.

- The watermark is embedded in critical bands in which magnitude of DFT coefficients is less than the defined masking threshold, $Y(z)$.
- The final watermark is now generated as

$$W(k) = \begin{cases} W_0(i), & if \quad mN \leq k \leq nN \\ W_0'(i), & if \quad (1-n)N \leq k \leq (1-m)N \\ 0, & otherwise \end{cases} \tag{15}$$

  here $0 \leq i \leq N_w$ and $0 \leq k \leq N-1$ and $(0 < m < n < 0.5)$ to maintain symmetry of final watermark.
- Once location of embedding is decided, the watermark scaling factor a has to be calculated for each critical band to ensure inaudibility of the embedded watermark. The scale factor a_z of z^{th} critical band is obtained by dividing masking threshold $Y(z)$ by the maximum magnitude component of the DFT coefficient in each critical band as

$$a_z = A\frac{Y(z)}{max(|F(k)|)}, \qquad for \quad z = 1, 2, \cdots, z_t \tag{16}$$

  Here A is the gain factor that controls the overall magnitude of the watermarked signal $F'(k)$ given in (2). The value of A varies from 0 to 1. The scaling factor a_z decides how much the amplitude of watermark is to be suppressed in the selected critical band before adding it to the spectrum of host signal.
- The scaled watermark is now added according to rule

$$\begin{aligned} F'(k) &= F(k), & if\, F(k) \geq Y(z) \\ &= F(k)(1 + a_z W(k)), & if\, F(k) < Y(z) \end{aligned} \tag{17}$$

  here $0 \leq k \leq N-1$.
- The modified amplitude of DFT coefficient $F'(k)$ is now combined with their corresponding phases $\phi(k)$, to get watermarked DFT coefficients $S'(k)$.
- The corresponding time domain watermarked signal $s'(n)$ is obtained by calculating $IDFT$ of $S'(k)$ given by (8).

2.4 Optimal Watermark Detection

The aim of watermark detection is to verify, whether or not the given watermark $\mathbf{W}_d$ at receiver end resides in the test signal $\mathbf{F}_t$. The detection is blind i.e. secret key is the only information that detector has at the receiver end. The detector uses salient points for synchronizing the embedded information, so that audio can be analyzed for salient point extraction. Watermark detection can be considered as a binary hypothesis test, solved by means of a correlation detector [10] and log-likelihood ratio detector [11], [12]. However, few assumptions are done before performing the detector tests.

The Host signal $\mathbf{F}$ and the watermark $\mathbf{W}$ are independent and identically distributed i.i.d random variables, hence the detector is optimum. DFT magnitude is wide sense stationary process and for large number of samples likelihood ratio and correlation coefficient attain Gaussian distribution due to central limit theorem.

Likelihood Ratio Detector. The watermarked signal $\mathbf{F}'$ given in (2), may undergo various signal processing or noisy channel attacks before reaching the receiver end. The received signal $\mathbf{F}_t$ is now used for watermark detection, by using log-likelihood ratio test. The best suited distribution for magnitude of DFT coefficients $F = [f(1), \cdots, f(N)]$ is two parameter Weibull distribution [13] which is defined for positive real axis only. The parameter estimation problem consists of finding the underlying distribution parameters by observing samples of random variable described in [14]. Given N sample values $[f(1), \cdots, f(N)]$, from the random variable F, which can be modeled by a two parameter Weibull distribution, the maximum likelihood estimators are utilized to find the values of shape and scale parameters respectively [15].

Since decoding is done without resorting to the original audio, the decoder has no access to the original coefficients. Hence the distribution of the non-watermarked coefficients f_i needs to be approximated by the distribution of the watermarked coefficients f_i'. As long as the embedding strength and thus the watermark power is kept small, the difference between the two distributions will be negligible.

Having identified a suitable model for host feature, we now find the likelihood ratio, as given in [16]. Also, the performance of a log-likelihood based technique is shown by Receiver Operating Characteristic (ROC) curve drawn between probability of false alarm P_f and probability of misdetection P_m.

Correlation Detector. The correlation detector, which is the Maximum Likelihood (ML) optimal detector, is applied to additive or multiplicative watermarking system. These detectors give optimal results while considering Gaussian distribution for the host signals. The correlation detection can be performed by computing the correlation c between pseudo-random sequence $\mathbf{W}$ and watermarked signal $\mathbf{F}_t$ in time or frequency domain given as

$$c = \mathbf{F}_t\mathbf{W} = [\mathbf{F}(1 + \alpha\mathbf{W})]\mathbf{W} = \mathbf{F}\mathbf{W} + \alpha\mathbf{F}\mathbf{W}\mathbf{W} \qquad (18)$$

The correlation of watermark (pseudo-random sequence) is compared to a predefined threshold to determine whether watermark is present in the signal or not. The received signal $\mathbf{F}_t$ is used for watermark detection, by using correlation test.

3 Experimental Results

To generate experimental results, a total of 9 standard audio test sequences are taken which are listed in table I. These test sequences are adopted to analyze the performance of the proposed watermarking algorithm. Each signal was sampled at 44.1 kHz, represented by 16 bits per sample, and eight seconds in length. The DFT magnitude of audio signal was assumed to follow Weibull distribution. The shape parameter, $\beta = 0.6833$ and scale parameter, $\alpha = 2.9369$ of Weibull distribution was evaluated using maximum likelihood estimation method.

Table 1. Audio test sequences (44.1 kHz, 16 bit)

Drums	Clarinet	Flute
speech(mono)	speech(stereo)	waltz
Synth	jazz	violin

3.1 Experimental Performance Evaluation

The value of scaling factor a is changed and its effect is seen on performance of likelihood ratio detector and correlation detector respectively. For this the values of a for various critical bands are obtained using (16).

– **Effect on detection threshold**
In case of LLR detector, the effect of scale factor a is observed on detection threshold Λ. First we have shown the curves between Λ and P_f keeping the value of a fixed. The upper and lower portion of figure 3 shows the variations of Λ with respect to P_f for two values of a, 0.0024 and 0.8 respectively. The first value, $a = 0.0024$, is obtained from MAS threshold and second value, 0.8 is selected close to the maximum limit of a to show the effects clearly visible. As can be seen from figure, for the same range of P_f the variations in Λ is only $0 < |\Lambda| \leq 0.08$ when $a = 0.0024$. Whereas the variations are quite high $(0 \leq |\Lambda| \leq 20)$ for $a = 0.8$. Next we analyse the variations of Λ with respect to

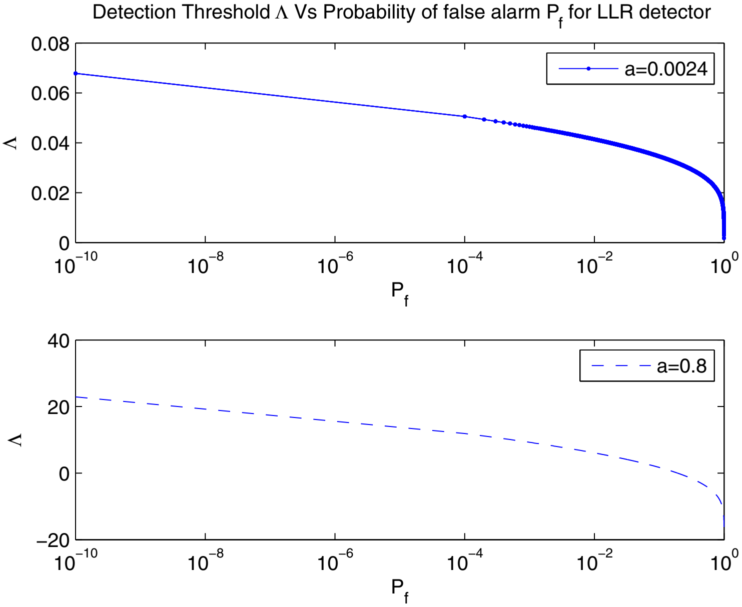


Fig. 3. Threshold versus probability of false detection for LLR detector for two values of scaling factor a

a for all the values in the range of $0 < a \leq 1$. Figure 4 shows the variation of Λ with respect to a for a fixed value of P_f ($\simeq 10^{-6}$). From the plot we observe that the value of Λ remains constant for $a \leq 0.04$. However, as the value of a is increased beyond 0.04 a steep rise in Λ is obtained. Another observation from figure is that with decreasing a, the value of detection threshold Λ also decreases which in turn degrades the detector response.

In case of correlation detector the effect of a is observed on detection threshold T_c. A plot between T_c and P_f for two different values of a (i.e. 0.0024 and 0.8) is shown in upper and lower portion of figure 5. From the figure we observe that $T_c = 0.27$ when $P_f = 10^{-3}$ for both the values of a. Also the value of threshold lies within the range of $0 \leq T_c \leq 0.5$ for a wide variation of a (i.e. $0 \leq a \leq 1$). Hence it can be inferred from the curves that the output of correlation detector is not much effected by scaling factor a.

- **Effect on ROC**

  The Receiver Operating Characteristic (ROC) curve is obtained from likelihood ratio and correlation watermark detectors, as shown in figure 6 respectively. The results are compared with actual experimental curve for both detectors with two different values of a, i.e. 0.0024 and 0.8. It is observed that for a=0.0024 the three curves nearly coincide with each other, whereas the same is not true for the case a=0.8. For the proposed value of a, the statistical detectors give optimum results which are close to actual experimental value. Further we observe that LLR detector gives better approximation to experimental results as compared to correlation detector, for all values of scaling factor.

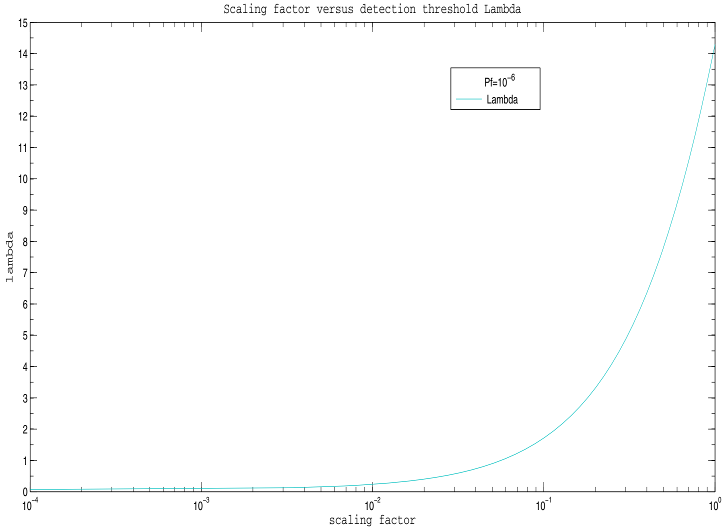


Fig. 4. Threshold versus scaling factor for Log-LLR detector for $P_f = 10^{-6}$

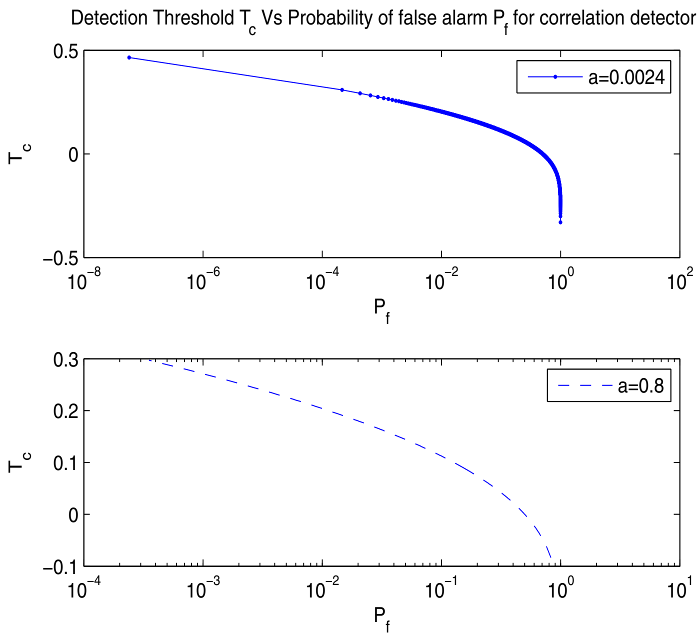


Fig. 5. Threshold versus probability of false detection for correlation detector for two values of scaling factor a

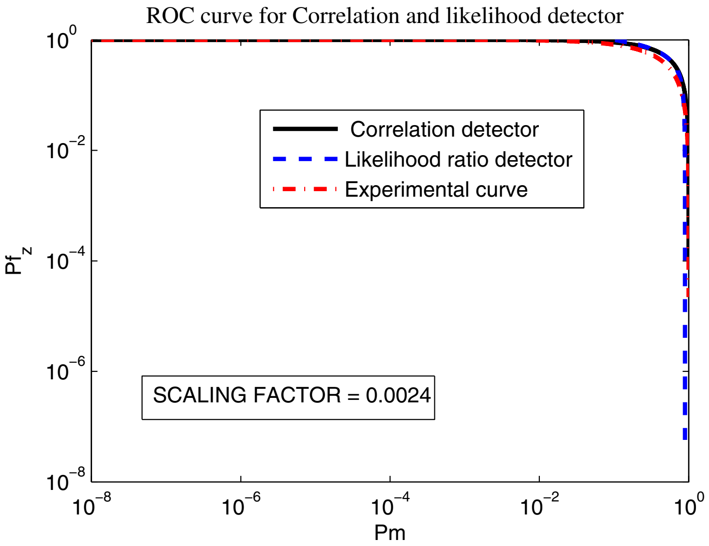


Fig. 6. Receiver operating characteristic curve for scaling factor α=0.0024

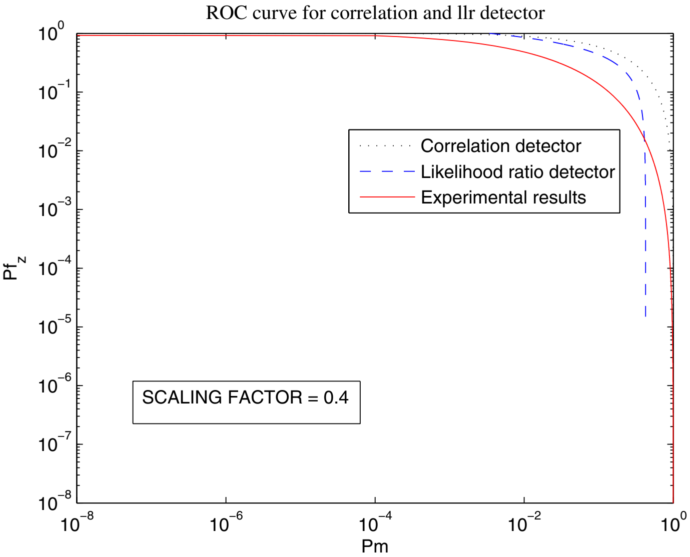


Fig. 7. Receiver operating characteristic curve for scaling factor α=0.0024 (Upper curve) and α=0.4 (lower curve) respectively

3.2 Objective and Subjective Quality Evaluation

Subjective and objective quality tests are performed to evaluate the quality of watermarked audio signal [17]. It is observed from the above results that the quality degradation of the proposed watermarking scheme is very small for the vast majority of the test items, given in table II. For all test items the Subjective difference grade (SDG) is within -0.7 to -0.065 which indicates that there is no significant distortion introduced by this scheme. For objective quality measure, software $'PQevalAudio'$ for perceptual evaluation of audio quality ($PEAQ$) is utilized to evaluate an objective difference grade (ODG), which is an objective measurement of SDG. Table 2 lists the average value of PEAQ/ODG with the give test items for varying value of a. It shows that as value of scaling factor a decreases, perceptual quality of watermarked audio becomes better. However, if the value of scaling factor is lowered below the value obtained from MAS (0.0024), ODG obtained is positive which is not acceptable, as per ITU recommendations. The value of ODG obtained from watermarked audio is -0.065 for the optimum value of $a = 0.0024$. From ROCs plotted in figure 6 and the objective quality given in table II we observe that for small values of a the detector response is poor, but perceptual quality is within acceptable limits. On the contrary, for larger values of scale factor ($a \simeq 0.04$) the detector response improves, but then the perceptual transparency is deteriorated. It can be inferred from these results that proposed technique gives a good tradeoff between perceptual transparency and detector performance.

Table 2. ODG for varying scaling factor and threshold

S.No.	*ODG*	*a*	*Λ*
1	-1.624	0.8	11.8741
2	-1.436	0.4	6.2374
3	-0.999	0.1	1.9353
4	-0.710	0.05	1.0094
5	-0.641	0.005	0.1050
6	-0.065	0.0024	0.0505
7	+0.045	0.001	0.0211

3.3 Watermark Embedding Capacity

The proposed scheme provides high watermark embedding capacity with least perceptual distortions. The embedding capacity of proposed scheme was found to be 1.4kbps, with ($ODG = -0.065$), when embedding was done in only one critical band. Table 3 compares the embedding capacity and perceptual quality of proposed scheme with other schemes present in literature. The average watermark capacity increased to 4kbps, with $ODG = -0.7$), when embedding was performed in more then one critical bands (*i.e.* 3).

Table 3. Comparison of ODG and watermark embedding capacity between available literature schemes

Technique	*ODG*	*EmbeddingCapacity*
Megias [6]	-0.5 to -2	61bps
Fujimoto [7]	–	1 Kbps
Fallahpour [8]	-0.5	3kbps
Proposed	-0.065 to -0.7	1.42 to 4kbps

3.4 Robustness to Attacks

The other major issue in watermarking is robustness to various attacks. We will now present the robustness of watermark against additive white gaussian noise (AWGN) noise, presence of multiple watermarks and MP3 compression.

Addition of AWGN Noise. More then 99 percent of watermark recovery is achieved for SNR value of $6dB$ and above. This implies high robustness of watermark against AWGN noise.

Presence of Multiple Watermark. To see the effect of presence of multiple watermark both types of detectors i.e. likelihood ratio and correlation detectors, are used. In case of LLR detector the value of Λ obtained is 9.8 for $P_f = 10^{-6}$ with scaling factor $a = 0.8$. The LLR detector output is shown for high value of

a, as the response of this detector is poor for small values of a, as can be seen from figure 6. Log-likelihood ratio Λ of correct watermark obtained is above 9.8 whereas the LLR ratio of other watermarks is well below the threshold. Similarly in case of correlation detector the value of threshold T_c obtained statistically was 0.271.

4 Conclusion

The proposed multiplicative spread spectrum based blind audio watermarking technique embeds watermark in DFT magnitude of audio signal. In order to improve two parameters, the embedding capacity and the computational complexity, a new perceptual model for magnitude of DFT coefficients is developed. This model finds the regions of highest watermark embedding capacity with least perceptual distortion. Also the proposed method reduces computations by bypassing the complex psychoacoustic modeling, required for fulfilling the condition of transparency. Theoretical evaluation of detector performance using correlation detector and likelihood ratio detector is undertaken under the assumption that host feature (DFT magnitude) follows Weibull distribution. The experimental and statistical results shown that proposed scheme gives higher embedding capacity as compared to existing watermarking techniques keeping the perceptual quality well within limits. Also, it was observed from experimental results that proposed scheme is robust to various signal processing attacks like presence of multiple watermarks, AWGN and MP3 compression.

References

1. Boney, L., Tewfik, A.H., Hamdy, K.N.: Digital watermarks for audio signal. In: Proc. IEEE Int. Conf. Multimedia Comput. Syst (ICMCS), Hiroshima, Japan, pp. 473–490 (June 1996)
2. Swanson, M.D., Zhu, B., Tewfik, A.H., Boney, L.: Robust audio watermarking using perceptual masking. Signal Process 66(3), 337–355 (1998)
3. Paraskevi Bassia, P., Ioannis Pitas, I., Nikos Nikolaidis, N.: Robust audio watermarking in the time domain. IEEE Trans. Multimedia 3, 232–241 (2001)
4. Cox, I.J., Kilian, J., Leighton, T., Shamoon, T.: Secure spread spectrum watermarking for multimedia. IEEE Transactions on Image Proc. 6(12), 1673–1687 (1997)
5. Malik, H., Ansari, R., Khokhar, A.: Robust audio watermarking using frequency-selective spread spectrum. IET Information Security 2(4), 129–150 (2008)
6. Megías, D., Herrera-Joancomartí, J., Minguillón, J.: Total Disclosure of the Embedding and Detection Algorithms for a Secure Digital Watermarking Scheme for Audio. In: Qing, S., Mao, W., López, J., Wang, G. (eds.) ICICS 2005. LNCS, vol. 3783, pp. 427–440. Springer, Heidelberg (2005)
7. Fujimoto, R., Iwaki, M., Kiryu, T.: A Method of High Bit Rate Data Hiding in Music Using Spline Interpolation. In: Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2006), pp. 11–14 (2006)

8. Fallahpour, M., Megias, D.: High capacity audio watermarking using FFT amplitude interpolation. IEICE Electronics Express 6(14), 1057–1063 (2009)
9. Schroeder, M.R., Atal, B.S., Hall, J.L.: Optimizing digital speech coders by exploiting properties of the human ear. Journal Acoust. Soc. America 66(6), 1647–1652 (1979)
10. Hyun, K.W., Dooseop, C., Hyuk, C., Taejeong, K.: Selective correlation detector for additive spread spectrum watermarking in transform domain. Signal Processing 90(8), 2605–2610 (2010)
11. Barni, M., Bartolini, F., De Rosa, A., Piva, A.: A new decoder for the optimum recovery of nonadditive watermarks. IEEE Trans. Image Processing 10(5), 755–766 (2001)
12. Cheng, Q., Huang, T.S.: Robust optimum detection of transform domain multiplicative watermarks. IEEE Trans. Signal Processing 51(4), 906–924 (2003)
13. Weibull, W.: A statistical distribution function of wide applicability. Journal of Applied Mechanics 18(3), 293–297 (1951)
14. van Trees, H.L.: Detection, Estimation and Modulation Theory, Part I. Wiley, New York (1968)
15. Stone, G.C., Van, H.G.: Parameter estimation for the Weibull distribution. IEEE Transactions On Electrical Insulation EI-12(4) (August 1977)
16. Barni, M., Bartolini, F.: Watermarking systems Engineering: Enabling Digital Assets Security and Other Applications. Marcel Dekker, New York (2004)
17. Neubauer, C., Herre, J.: Digital watermarking and its influence on audio quality. In: Proceedings of 105th Audio Engineering Society Convention, San Francisco, CA (September 1998)

A Comparative Analysis of Various Deployment Based DDoS Defense Schemes

Karanbir Singh[1], Navdeep Kaur[2], and Deepa Nehra[3]

[1] Seth Jai Parkash Polytechnic, Damla, Yamuna Nagar, Haryana, India
karan_nehra@yahoo.co.in
[2] Chandigarh Engineering College, Landran, Mohali, Punjab, India
nrwsingh@yahoo.com
[3] Tilak Raj Chadha Institute of Management & Technology, Yamuna Nagar, Haryana, India
deepa.nehra@gmail.com

Abstract. Distributed denial of service attack is a major threat to the availability of internet services and resources. The current internet infrastructure is vulnerable to DDoS attacks and has no built in mechanism to defend against them. The main task of the defense system is to accurately detect and respond against DDoS attacks. A variety of DDoS defense solutions are available but having difficulties to choose among them. There are different places in the internet, where a defense system can be deployed. The various points in the internet where defense systems can be deployed are identified and discussed here. A comparative analysis of different defense schemes corresponding to deployment points are also carried out. The main aim of this review paper is to provide an individual or academia to insight into various possible deployments locations suitable for DDoS defense system. It helps them to choose an appropriate defense method and suitable deployment location.

Keywords: DDoS, Network Security, Distributed Defense, Deployment.

1 Introduction

A Denial of Service (DoS) attack is an attack with the aim of preventing normal users from using some network resource such as a website, web service, or computer system [1]. A Distributed Denial of Service (DDoS) attack is a large scale, coordinated attack launched through many compromised system on the internet to the availability of services of a given target system or network. In DDoS attack a large number of packets are sent by the attacker through multiple machines to a victim. The packets arrive on the victim are huge in quantity and it will quickly exhaust its resources like bandwidth, cpu time and buffers. The victim devotes its most of the time in handling the attack packets and cannot switch to the legitimate clients. Thus legitimate clients are dispossessed of victim resources as long as the attack last. These attacks are widely imposed a serious threat to the internet services.

One of the first major DDoS attacks was waged against Yahoo.com in February 2000, keeping it off from the internet for nearly 2 hours, costing it significant loss of

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 606–616, 2013.

advertising revenue [2]. Recently, attackers perform a variety of DDoS attacks against many companies providing anti-spam services [3]. These attacks force them to shut down their services. Reports by law enforcement agencies indicates that the percentage of organizations that experienced virus disasters has grown geometrically every year over the last decade, with 92 percent of organizations reporting such incidents during 2003. DDoS attacks are one of the overall costly security incident for organizations. A lot of works have been done to combat against DDoS attacks. An excellent review of existing DDoS defense techniques is available in [6-12].

The organization of rest of the paper is as follows. Section 2 provides an insight into various deployment locations where defense system can be implemented. In section 3 we review and characterize some existing DDoS defense system on the basis of deployment. Section 4 compares them to evaluate their performance on the basis of some key points like detection, response, deployment, robustness and implementation. Section 5 suggests the best way to defend against DDoS Attacks by using performance evaluation. This work provides an individual to select an appropriate deployment for the defense method and point out the various deficiencies in the existing methods.

2 DDoS Defense Locations

DDoS attacks can be originated from any machine on the internet, which is geographically placed at any location. The network from where the attack stream originates is called source network. Then this attack stream is forwarded by many routers through the intermediate network and then later converged to a single machine in victim network. So, the networks responsible for the happening of DDoS attacks are source network (from where the attack originates), Intermediate network (responsible for forwarding the attack traffic to the target) and the victim network (where victim machine receives the burnt of attacks). So the three different locations are source network, intermediate network or victim network, which can host the DDoS defense system [13]. The figure 1 represents a simplified network divided into three parts as mentioned above. The edge router connects source/victim networks to the intermediate network (normally operated by ISP's). The intermediate network contains many ISP's, interconnected with each other through core routers. This figure will be used to illustrate various defensive locations. Here, the nodes at the left side are attackers and the node at the right side is the target of the DDoS attack. The figure given below illustrates that the defense system can be deployed at source network, intermediate network, victim network or distributed at all three locations.

The different deployment based DDoS defense schemes are effective in their respective area but suffers from some disadvantages. The table 1 gives their relative advantages and disadvantages.

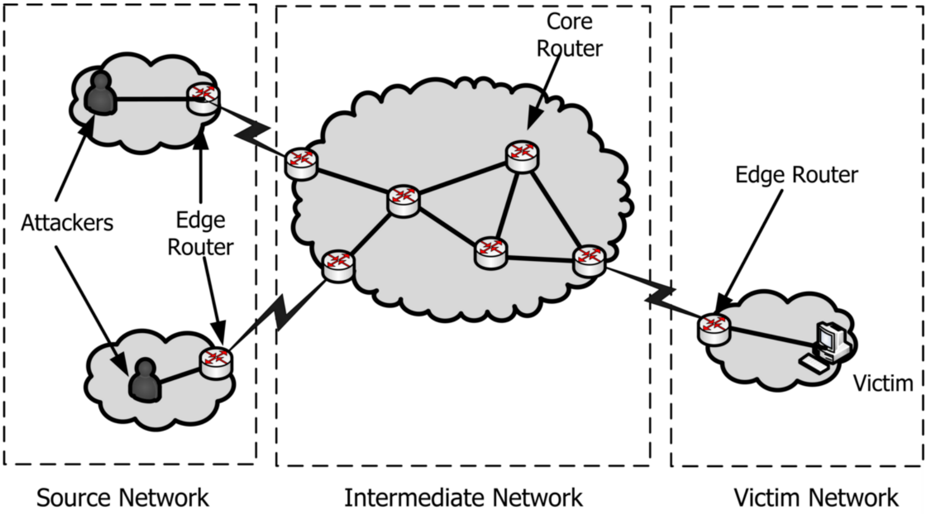


Fig. 1. Network illustrating different locations for DDoS deployment

Table 1. Comparison Between Different Deployment Locations

Deployment Point	Advantages	Disadvantages
Source Network	Good point to detect when attack happen	• Overwhelm with large no of attack packets. • Only protect individual target not the other edges.
Victim Network	Low volume of attack packets will flow in outgoing network	• Difficult to determine whether traffic is legitimate or attack. • Requires widespread deployment to all source networks.
Intermediate Network	Core router defenses are effective as all the traffic goes through them.	• Core router cannot devote sufficient resources for analyze individual packets. • Core routers could inflict massive collateral damage.
Distributed Network	More robust as the defense components are deployed on all above three locations	• To be effective it needs large scale deployment which is expensive and difficult

3 Review of Existing Defense Mechanisms

In this section, we review some existing DDoS defense methods. The methods are identified from the literature based upon their deployment in the network.

3.1 Source Based Defense

D-WARD [14] is a source-end DDoS defense system whose goal is to detect and constrain outgoing attacks at the source network. The system is installed at the source router and monitors the traffic passing through the router in both directions and correlates this observation to detect anomalies that can be a sign of DDoS attack. Upon detection, it selectively imposes a rate limit on the outgoing flow to the victim, attempting to detect and forward legitimate packets regardless of the limit. However, a major drawback is that it is only effective in actually stopping attacks if deployed at most attacking network.

Ingress Filtering [15] is a technique used to ensure that weather the incoming packets are coming from their original locations. Egress filtering is an outbound filter, which monitors and restricts the flow of outgoing traffic. A key requirement for ingress or egress filtering is information of the expected IP addresses at a particular port. For some networks with complex topologies, it is not always easy to obtain this information.

3.2 Victim Based Defense

Preferential filtering [16] is basically an IP traceback schemes to obtain the information concerning whether a network edge is infected (i.e. on the attacking path of an attacker) or clean (not on the attacking path). We observe that all the edges on the path of attacker are marked as “infected”, edges on the path of a legitimate client will normally be “clean”. This scheme filter out packets that are inscribed with the marks of “infected” edges, the scheme removes most of the DDoS traffic coming from attackers while putting little effect on legitimate traffic.

NetBouncer [23] is a client-legitimacy-based DDoS filtering. It tries to detect legitimate clients and only serve their packets. NetBouncer is deployed near the victim side and it is an inline defense device deployed in front of the possible choke point. A NetBouncer device maintains a large list of legitimate clients. If packets are received from a client (source) not on the legitimacy list, NetBouncer device will proceed to administer a variety of legitimacy tests to challenge the client to prove its legitimacy. The legitimacy of a client expires after a certain interval. NetBouncer can ensure good service to legitimate clients in the most of the cases and does not require modifications to clients or servers. However, some of the legitimate clients will not be validated. The use of only legitimate IP list has the potential problem that legitimate client identity can be misused for attacks.

3.3 Core Router Based Defense

Perimeter-based defense mechanisms [17] are used by Internet service providers (ISPs) to provide the anti-DDoS service to their customers. These methods completely depend on the edge routers of ISP to co-operatively identify the flooding sources and start rate-limit filters to block the attack traffic. This system does not need any support from ISP routers (outside or inside of the ISP), which not only makes it locally deployable, but also put less stress on the ISP core routers. This method requires widespread deployment and does not perform well in noncontiguous deployment.

Distributed Change-point Detection (DCD) [18] scheme detects DDoS flooding attacks by observing the propagation patterns of unexpected traffic changes at distributed network points. Once a sufficiently large CAT tree is constructed to exceed a preset threshold, an attack is declared. The system is deployed over multiple Autonomous Systems domains. The system detects traffic changes, checks flow propagation patterns, aggregates suspicious alerts, and merge CAT subtrees from collaborative servers into a global CAT tree. The system is built upon attack-transit routers, which works cooperatively together. Each ISP domain has a CAT server which aggregate the flooding alerts reported by the routers. CAT domain servers collaborate with each other to make the final decision.

Controller agent model [19] counteracts DDoS attacks within one ISP domain. In this model, agents represent the edge routers and controllers represent trusted entities owned by the ISP. Once a target detects an attack, it sends a request to the controller, asking all agents to mark all packets to the target after checking the marking field, the target can find out which agent (edge router) is the entry point for the attack traffic. The target then sends a refined request to the controller, asking some particular agents to filter attack traffic according to the attack signature provided by the target. The main limitation of this model is that it uses third party detection for detecting and characterizing attack traffic.

3.4 Distributed Defense

Local Aggregate-Based Congestion Control (Local ACC) [20] provides a self-contained solution in which detection and rate-limiting of DDoS attacks are done on a single router. Routers identify high-bandwidth traffic aggregates in their queue which are responsible for the majority of packet drops and responds by imposing a rate limit on each traffic aggregate. Pushback [21] extends local ACC with communication and coordination capabilities. If it is difficult for the congested router to control the aggregate, then it issues a rate limit request to its immediate upstream neighbors.

DefCOM [22] is a distributed cooperative system for DDoS defense. DefCOM builds a distributed peer-to-peer network of cooperative defense nodes which are scattered throughout the Internet. Defense nodes exchange information and control messages to detect attacks, and collectively respond to them while ensuring good service to legitimate traffic. DefCOM nodes can be classified into three categories, based on their functionality: Alert generator nodes that detect the attack and deliver an alarm to the rest of the peer network, Rate limiter nodes that rate limit a high volume of traffic destined for the victim, and Classifier nodes that perform selective rate-limiting.

Active Security System (ASSYST) [5] supports distributed response with non-contiguous deployment. All ASSYST nodes are essentially the equivalent of classifier nodes and are deployed only at edge networks. Active Security Protocol, which allows a set of active routers to interact in order to isolate the sources of a DDoS attack even in the case of address spoofing. Tuining and deployment of the Active Security System are perfectly suited to a Programmable Network environment.

4 Comparative Analysis of Different DDoS Defense Methods

The DDoS defense mechanisms discussed above are compared using some important performance metrics like attack detection, attack response, deployment location, robustness and implementation. Here firstly we will discuss these metrics and later use them to compare the DDoS defense system.

Attack Detection: DDoS detection is usually the first step in the mitigation of a DDoS attack. Any DDoS detection technique always attempts to detect an attack by observing anomalous changes in IP attributes or traffic volume because there do not exist clear DDoS attack signatures. The detection is main feature of a defense mechanism, because if we can detect an attack at its initial stage then it can be corrected by deploying a prevention or reaction countermeasure. This provides fast protection to the legitimate users against attacks. In addition, detection helps us to identify the attacker, which can later to be blocked at the source. The defense system which detects attacks quickly, with in time, with accuracy and minimal deployment costs are said to be efficient.

Attack Response: After a DDoS attack has been detected, response techniques attempt to control incoming traffic by packet filtering or rate limit techniques. Based on the studies done, packet filtering techniques can cause more damage to legitimate traffic than rate limit techniques, because it is difficult to distinguish DDoS traffic from normal traffic. Packet filtering task is usually done at the routers based on clearly defined attack signatures. However, DDoS attack traffic cannot be filtered out if it uses packets that request legitimate services [24]. Another common drawback of packet filtering is that it usually needs to be deployed widely in order to protect the victim. Rate limiting is used to control the traffic flow on a network interface. The traffic which is less than or equal to the specified rate is allow to send, whereas traffic which exceeds the rate is either dropped or delayed [25]. The effectiveness of rate limiting to defend DDoS attacks is defined in [26]. Rate limiting can be used as a fast, automatic reaction mechanism to mitigate an attack without any undue penalties for legitimate traffic [26]. In contrast, collateral damage for legitimate traffic is unavoidable in packet filtering because DDoS traffic cannot be easily distinguished from legitimate traffic [27].The goal of attack response is to improve the situation for legitimate users and mitigate the DDoS effect.

Deployment: Deployment of DDoS defense method is an important issue that must be considered. It tells us the place in the network where we can put our DDoS defense system. A practical DDoS defense solution should be easy to deploy in the sense that it minimally interferes with existing Internet protocols and settings. Also, the scale of deployment should be reasonable. Defenses which require local deployment are usually preferred over those which require global deployment. However, if a DDoS defense requires global deployment, then this deployment should be incrementally feasible.

Robustness: Robustness tells us the degree up to which the defense system can resist the attacks. When the defense system is deployed and known to the hackers, then there is possibility that the hacker can compromise the defense system and uses it to further attack on the protected network. Some of the defense systems are less vulnerable to attacks than others. Distributed defense is less vulnerable to DDoS attacks than isolated but still there is a possibility that distributed defense system fails as it can be targeted by the hacker. In case of distributed defense, the information exchanged between defense components are also vulnerable to hackers. So, it depends upon the defense system that how securely they exchange the information.

Implementation: DDoS defense systems are deployed at various locations in different schemes bears an implementation overhead. The defense systems follows different deployment strategies like their defense components are deployed at source/victim side or on different parts in the intermediate networks. DDoS defense sometimes require major changes (such as altering behavior of core routers, deploying new software on all machines in the Internet or changing fundamental Internet protocols) will never be implemented without far more convincing evidence that they would work if their price was paid. In particular, it will be more depressing if the defense system fails to respond to the majority of DDoS attacks even after doing major changes to the internet. This issue points out one serious advantage that target end systems have. They typically cost less to deploy, so if they do not work, less has been lost. Overall, we need to concentrate more on nature and behavior of these attacks and the characteristics of proposed defenses before we should accept anyone's.

The table 2 (divided into two parts) gives deployment based comparison between different DDoS defense mechanisms.

Table 2. Deployment Based Comparisons Between Different DDoS Defense Methods

Deployment Scheme	Scheme Name	Attack Detection	Attack Response
Source Based Defense	D-Ward	Abnormality in traffic	Rate Limiting
	Egress/Ingress filtering	IP Address validity	Rule based packet filtering
Victim Based Defense	Preferential Filtering	Attack traffic graph	Filtering packets with infected edges
	NetBouncer	Legitimacy test for clients	Packet filtering based of legitimacy test
Core Router Based Defense	Perimeter based defense	Traffic Aggregate	Rate Limit Filters
	Controller Agent Model	Signature Matching	Packet Filtering through agents
	Collaborative Change Detection	Change Aggregation tree (CAT)	Packet Filtering
Distributed Defense	ACC & Pushback	Congestion based	Rate limiting
	ASSYST	Packet Classifier Intrusion Detection	Packet Filtering Through Programmable Routers
	DefCom	Traffic Tree Discovery	Distributed rate limiting

Deployment Scheme	Scheme Name	Deployment Location	Robustness	Implementation
Source Based Defense	D-Ward	Source Network	Weak	Difficult
	Egress/Ingress filtering	Source Network	Weak	Difficult
Victim Based Defense	Preferential Filtering	Victim Network	Weak	Easy
	NetBouncer	Victim Network	Weak	Easy
Core Router Based Defense	Perimeter based defense	ISP Core network	Moderate	Moderate
	Controller Agent Model	ISP Core network	Moderate	moderate
	Collaborative Change Detection	ISP Core Network	Moderate	difficult
Distributed Defense	ACC & Pushback	Throughout the network	Strong	Difficult
	ASSYST	Throughout the network	Strong	Difficult
	DefCom	Throughout the network	Strong	Difficult

5 Performance Evaluations

There's a simple argument that which kind of DDoS defense solution is necessary to efficiently protect the network. Source based defense systems detect and filter the attack traffic at the early stages when the attack happens. This is effective, if deployment will cover maximum source networks. But practically it seems to be very difficult to cover all available source networks. Victim network is the best place to detect attack traffic due to its huge volume and easy deployment. But it suffers from high flood rate and itself vulnerable to DDoS attacks. Core routers in the intermediate network are the best places for attack detection and filtration but it require entire coverage, because no single location can capture all attacks. The individual packet scanning also creates additional overhead for the routers. All these schemes perform well in their respective area but we cannot rate anyone best suitable for an individual. All these schemes also have some drawbacks. So we need a defense system which can put their defense components at the following locations.

- Near the target, this is a good position to recognize attacks.
- Near the attackers, this is best place to differentiate between good and bad packets.
- In the center of the network, which achieve high defensive coverage with relatively few deployment points.

Distributed defense systems overcome the shortcomings of intermediate and source/victim end based defense systems. The distributed DDoS defense system spans its defensive components at the above mentioned three locations. Components of distributed defense system are deployed at various locations and cooperate with each other to defend the attacks. Distributed DDoS defense system is only solution to effectively control the flood of attack traffic. Hence, we can say that distributed solution provides effective protection against DDoS attacks than other kind of solutions.

6 Conclusion

This paper classified the various deployment based categories of DDoS defense systems. The categories identified are source, intermediate, victim and distributed networks. A comparison between these categories is identified. The existing defense methods falling under these categories are reviewed and their performance is evaluated on the basis of some metrics. After their performance evaluation the fact is discovered that not any individual location is best for the complete protection against DDoS attacks. We also suggested that we need a distributed defense in which defense components are placed on all the locations to effectively control attack flood. So in the end we make a conclusion that this classification is helpful for an individual in selecting an appropriate defense mechanism

References

1. Karig, D., Lee, R.: Remote Denial of Service Attacks and Countermeasures. Technical Report CEL2001-002, Department of Electrical Engineering, Princeton University (2001)
2. Yahoo on Trail of Site Hackers, `http://www.wired.com/techbiz/media/news/2000/02/34221`
3. Spam block lists bombed to oblivion, `http://www.msnbc.msn.com/id/3088113`
4. Sachdeva, M., Singh, G., Kumar, K.: A Comprehensive Survey of Distributed Defense Techniques Against DDoS Attacks. International Journal of Computer Science and Network Security 9(12) (2009)
5. Canonico, R., Cotroneo, D., Peluso, L., Romano, S., Ventre, G.: Programming Routers to Improve Network Security. In: Proc. of the OPENSIG 2001 Workshop, Next Generation Network Programming (2001)
6. Mirkovic, J., Reiher, P.: A Taxonomy of DDoS attack and DDoS Defense Mechanisms. ACM SIGCOMM Computer Communications Review 34(2), 39–53 (2004)
7. Defenses against distributed denial of service attacks, `http://www.garykessler.net/library/ddos.html`
8. Lin, S., Chieuh, T.: A Survey on Solutions to Distributed Denial of Service Attacks. RPE Technical Report (September 2006)
9. Chen, L., Longstaff, T., Carley, K.: Characterization of Defense Mechanisms Against Distributed Denial of Service Attacks. Computers & Security, 665–678 (2004)
10. Abliz, M.: Internet denial of service attacks and defense mechanisms. University of Pittsburgh Technical Report, No. TR-11-178 (2011)
11. Sachdeva, M., Singh, G., Kumar, K.: Deployment of Distributed Defense Against DDoS attacks in ISP domain. International Journal of Computer Applications 15(2) (2011)
12. Fadlallah, A., Serhrouchni, A.: Denial of service attacks and defense schemes analysis and taxonomy. In: 3rd International Conference: Sciences of Electronics Technologies of Information and Telecomm., TUNISIA (March 2005)
13. Mirkovic, J., Dietrich, S., Dittrich, D.: Internet Denial of Service: Attack and Defense Mechanisms. Prentice Hall (December 2004)
14. Mirkovic, J., Prier, G., Reiher, P.: Source-end DDoS Defense. In: Proceedings of Network Computing and Applications Symposium NCA (2003)
15. Cert advisory ca-2000-01 Denial-of-Service Developments, `http://www.cert.org/advisories/CA-2000-01.html`
16. Sung, M., Xu, J.: IP Traceback-Based Intelligent Packet Filtering: A Novel Technique for Detecting Against Internet DDoS Attacks. In: Proc. of 10th IEEE International Conference on Network Protocols (2002)
17. Chen, S., Song, Q.: Perimeter-Based Defense against High Bandwidth DDoS Attacks. IEEE Transactions on Parallel and Distributed Systems 16(6) (2005)
18. Chen, Y., Hwang, K., Ku, W.: Collaborative Detection of DDoS Attacks over Multiple Network Domains. IEEE Transactions on Parallel and Distributed Systems 18(12) (2007)
19. Tupakula, U., Varadharajan, V.: A Controller Agent Model to Counteract DoS Attacks in Multiple Domains. In: Proc. of Integrated Network Management, IFIP/IEEE 8th International Symposium (2003)
20. Mahajan, R., Bellovin, S., Floyd, S., Ioannidis, J., Paxson, V., Shenker, S.: Controlling High Bandwidth Aggregates in the Network. ACM Computer Communications Review 32(3) (2002)
21. Ioannidis, J., Bellovin, S.: Pushback: Router-Based Defense against DDoS Attacks. In: Proc. of NDSS (February 2002)

22. Mirkovic, J., Robinson, M., Reiher, P., Oikonomou, G.: A Framework for Collaborative DDoS Defense. In: Proc. of the 22nd Annual Computer Security Applications Conference, Miami, Florida, USA, pp. 33–42 (December 2006)
23. Thomas, R., Mark, B., Johnson, T.: Net bouncer: Client-Legitimacy-Based High Performance DDoS Filtering. In: Proc. of the DARPA Information Survivability Conference and Exposition. IEEE (2003)
24. Xiang, Y., Zhou, W., Chowdhury, M.: A Survey of Active and Passive Defence Mechanisms against DDoS Attacks. Technical Report, TR C04/02, School of Information Technology, Deakin University, Australia (March 2004)
25. Evans, J., Filsfils, C.: Deploying IP and MPLS QoS for multiservice networks. Theory and Practice. Morgan Kaufmann (2007)
26. Molsa, J.: Effectiveness of Rate-Limiting in Mitigating Flooding DoS Attacks. In: Proc. of the Third IASTED International Conference on Communications, Internet, and Information Technology, pp. 155–160 (2004)
27. Sterne, D., Djahandari, K., Wilson, B., Babson, B., Schnackenberg, D., Holliday, H., Reid, T.: Autonomic Response to Distributed Denial of Service Attacks. In: Lee, W., Mé, L., Wespi, A. (eds.) RAID 2001. LNCS, vol. 2212, p. 134. Springer, Heidelberg (2001)

WG-8: A Lightweight Stream Cipher for Resource-Constrained Smart Devices

Xinxin Fan, Kalikinkar Mandal, and Guang Gong

Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
{x5fan,kmandal,ggong}@uwaterloo.ca

Abstract. Lightweight cryptographic primitives are essential for securing pervasive embedded devices like RFID tags, smart cards, and wireless sensor nodes. In this paper, we present a lightweight stream cipher WG-8, which is tailored from the well-known Welch-Gong (WG) stream cipher family, for resource-constrained devices. WG-8 inherits the good randomness and cryptographic properties of the WG stream cipher family and is resistant to the most common attacks against stream ciphers. The software implementations of the WG-8 stream cipher on two popular low-power microcontrollers as well as the extensive comparison with other lightweight cryptography implementations highlight that in the context of securing lightweight embedded applications WG-8 has favorable performance and low energy consumption.

Keywords: Lightweight stream cipher, resource-constrained device, cryptanalysis, efficient implementation.

1 Introduction

The Internet of Things (IoT) is an emerging computing and communication paradigm in which smart devices (e.g., RFID tags, smart cards, wireless sensor nodes, etc.) are linked through both wired and wireless networks to the Internet. Those smart devices interact and cooperate with each other to conduct complicated tasks such as sensing the environment, interpreting the data, and responding to events. While the IoT provides new and exciting experience for end users, it also opens up new avenues to hackers and organized crime. Recent attacks to a wide range of smart devices [13, 39] have emphasized that without adequate security the IoT will only become pervasive nightmare.

The challenges for deploying security solutions for smart devices are threefold: 1) The overhead (i.e., the gate count in hardware or the memory footprint in software) of security solutions should be minimal due to the low-cost nature of smart devices; 2) The power consumption of security solutions should be minimal due to the low-power characteristic of smart devices; and 3) The performance of security solutions should be reasonable to support applications and end-user requirements. To address the aforementioned challenges for securing smart devices, a new research direction called *lightweight cryptography* has

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 617–632, 2013.

been established which focuses on designing novel cryptographic algorithms and protocols tailored for implementation in resource-constrained environments.

A host of lightweight symmetric ciphers that particularly target for resource-constrained smart devices have been proposed in the past few years. Early work focuses on optimizing hardware implementations of standardized block ciphers such as AES [17], IDEA [25] and XTEA [22]. Later on, researchers have shown how to modify a classical block cipher like DES [24] for lightweight applications. Recent proposals deal with new low-cost designs, including lightweight block ciphers PRESENT [5], KATAN/KTANTAN [6], PRINTcipher [23], LED [20], and Piccolo [36], lightweight stream ciphers Grain [21], Trivium [7], and MICKEY [3], as well as a lightweight hybrid cipher Hummingbird/Hummingbird-2 [15, 16]. A good research survey about recently published lightweight cryptographic implementations can be found in [14].

In this paper we present the stream cipher WG-8, which is a lightweight variant of the well-known WG stream cipher family [29] as submitted to the eSTREAM project. WG-8 inherits good randomness properties of the WG stream cipher family such as period, balance, ideal two-level autocorrelation, ideal tuple distribution, and exact linear complexity. Moreover, WG-8 is able to resist the most common attacks against stream ciphers including algebraic attack, correlation attack, differential attack, cube attack, distinguish attack, discrete fourier transform attack, and time-memory-data tradeoff attack, thereby providing adequate security for lightweight embedded applications.

We also propose several techniques for efficient implementation of the stream cipher WG-8 on two low-power microcontrollers, including an 8-bit microcontroller ATmega128L from Atmel and a 16-bit microcontroller MSP430 from Texas Instruments. Our experimental results show that WG-8 can achieve high throughput of 185.5 Kbits/s and 95.9 Kbits/s on the above two microcontrollers with energy efficiency of 458 nJ/bit and 125 nJ/bit, respectively. When compared to other lightweight cryptography implementations in the literature, the throughput of the WG-8 is about $2 \sim 15$ times higher and the energy consumption is around $2 \sim 220$ times smaller than those of most previous ciphers.

The remainder of this paper is organized as follows. Section 2 gives a description of the lightweight stream cipher WG-8. Subsequently, in Section 3 we analyze the security of the WG-8 against the most common attacks to stream ciphers. Section 4 describes efficient techniques for implementing the WG-8 stream cipher on low-power microcontrollers and reports our experimental results and comparisons with previous work. Finally, Section 5 concludes this contribution.

2 The Lightweight Stream Cipher WG-8

2.1 Preliminaries

We define the terms and notations that will be used to describe the lightweight stream cipher WG-8 and its architecture as well as to characterize its randomness and cryptographic properties.

- $\mathbb{F}_2 = \{0, 1\}$, the Galois field with two elements 0 and 1.
- $p(x) = x^8 + x^4 + x^3 + x^2 + 1$, a primitive polynomial of degree 8 over $\mathbb{F}_2$.
- $\mathbb{F}_{2^8}$, the extension field of $\mathbb{F}_2$ defined by the primitive polynomial $p(x)$ with 2^8 elements. Each element in $\mathbb{F}_{2^8}$ is represented as an 8-bit binary vector. Let ω be a primitive element of $\mathbb{F}_{2^8}$ such that $p(\omega) = 0$.
- $\mathrm{Tr}(x) = x + x^2 + x^{2^2} + \cdots + x^{2^7}$, the trace function from $\mathbb{F}_{2^8} \mapsto \mathbb{F}_2$.
- $l(x) = x^{20} + x^9 + x^8 + x^7 + x^4 + x^3 + x^2 + x + \omega$, the feedback polynomial of LFSR (which is also a primitive polynomial over $\mathbb{F}_{2^8}$).
- $q(x) = x + x^{2^3+1} + x^{2^6+2^3+1} + x^{2^6-2^3+1} + x^{2^6+2^3-1}$, a permutation polynomial over $\mathbb{F}_{2^8}$.
- WGP-8$(x^d) = q(x^d + 1) + 1$, the WG-8 permutation with decimation d from $\mathbb{F}_{2^8} \mapsto \mathbb{F}_{2^8}$, where d is coprime to $2^8 - 1$.
- WGT-8$(x^d) = \mathrm{Tr}(\text{WGP-8}(x^d)) = \mathrm{Tr}(x^9 + x^{37} + x^{53} + x^{63} + x^{127})$, the WG-8 transformation with decimation d from $\mathbb{F}_{2^8} \to \mathbb{F}_2$, where d is coprime to $2^8 - 1$.
- Polynomial basis (PB) of $\mathbb{F}_{2^8}$: A polynomial basis of $\mathbb{F}_{2^8}$ over $\mathbb{F}_2$ is a basis of the form $\{1, \omega, \omega^2, \cdots, \omega^7\}$.
- Normal basis (NB) of $\mathbb{F}_{2^8}$: A normal basis of $\mathbb{F}_{2^8}$ over $\mathbb{F}_2$ is a basis of the form $\{\theta, \theta^2, \cdots, \theta^{2^7}\}$, where $\theta = \omega^5$ (i.e., a normal element) is used in this work.
- Autocorrelation: The autocorrelation of a binary sequence with period T is defined as the difference between the agreements and disagreements when the symbol 0 maps to 1 and 1 maps to -1. If all the out-of-phase autocorrelation is equal to -1, then the sequence is said to have *ideal two-level autocorrelation.*
- Linear span (LS): The linear span or linear complexity of a binary sequence is defined as the length of the smallest linear feedback shift register (LFSR) which generates the entire binary sequence.
- Nonlinearity: The nonlinearity of a function f is defined as the minimum distance from f to any affine function with the same number of variables.
- Algebraic immunity (AI): The algebraic immunity of a function f is defined as the minimum degree of an annihilator Boolean function g such that g is equivalent to either f or the complement of f (i.e., $fg = 0$ or $(f + 1)g = 0$). In the ideal case, the algebraic immunity of a function f is equal to the degree of f, thus making it immune to algebraic attacks.
- $\oplus$, the bitwise addition operator (i.e., XOR).
- $\otimes$, the multiplication operator over $\mathbb{F}_{2^8}$.

2.2 The Description of the Stream Cipher WG-8

WG-8 is a lightweight variant of the well-known Welch-Gong (WG) stream cipher family with 80-bit secret key and 80-bit initial vector (IV), which can be regarded as a nonlinear filter generator over finite field $\mathbb{F}_{2^8}$. The stream cipher WG-8 consists of a 20-stage LFSR with the feedback polynomial $l(x)$ followed by a WG-8 transformation module with decimation $d = 19$, and operates in two phases, namely an initialization phase and a running phase.

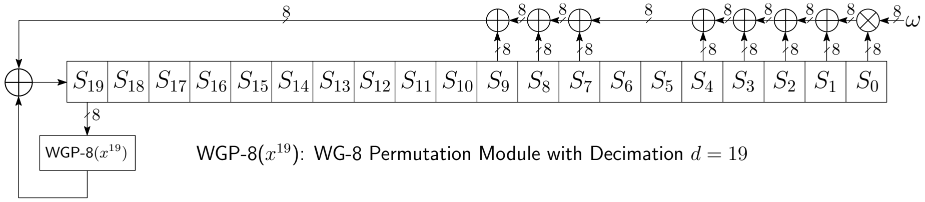


Fig. 1. The Initialization Phase of the Stream Cipher WG-8

Initialization Phase. The key/IV initialization phase of the stream cipher WG-8 is shown in Fig. 1.

Let the 80-bit secret key be $K = (K_{79}, \ldots, K_0)_2$, the 80-bit IV be $IV = (IV_{79}, \ldots, IV_0)_2$, and the internal state of the LFSR be $S_0, \ldots, S_{19} \in \mathbb{F}_{2^8}$, where $S_i = (S_{i,7}, \ldots, S_{i,0})_2$ for $i = 0, \ldots, 19$. The key and IV initialization process is conducted as follows: $S_{2i} = (K_{8i+3}, \ldots, K_{8i}, IV_{8i+3}, \ldots, IV_{8i})_2$ and $S_{2i+1} = (K_{8i+7}, \ldots, K_{8i+4}, IV_{8i+7}, \ldots, IV_{8i+4})_2$ for $i = 0, \ldots, 9$.

Once the LFSR is loaded with the key and IV, the apparatus runs for 40 clock cycles. During each clock cycle, the 8-bit internal state S_{19} passes through the nonlinear WG-8 permutation with decimation $d = 19$ (i.e., the WGP-8(x^{19}) module) and the output is used as the feedback to update the internal state of the LFSR. The LFSR update follows the recursive relation:

$$\begin{aligned} S_{k+20} = (\omega \otimes S_k) \oplus S_{k+1} \oplus S_{k+2} \oplus S_{k+3} \oplus S_{k+4} \oplus \\ S_{k+7} \oplus S_{k+8} \oplus S_{k+9} \oplus \mathsf{WGP\text{-}8}(S_{k+19}^{19}), \qquad 0 \le k < 40. \end{aligned}$$

After the key/IV initialization phase, the stream cipher WG-8 goes into the running phase and 1-bit keystream is generated after each clock cycle.

Running Phase. The running phase of the stream cipher WG-8 is illustrated in Fig. 2. During the running phase, the 8-bit internal state S_{19} passes through the nonlinear WG-8 transformation with decimation $d = 19$ (i.e., the WGT-8(x^{19}) module) and the output is the keystream. Note that the only feedback in the running phase is within the LFSR and the recursive relation for updating the LFSR is given below:

$$S_{k+20} = (\omega \otimes S_k) \oplus S_{k+1} \oplus S_{k+2} \oplus S_{k+3} \oplus S_{k+4} \oplus S_{k+7} \oplus S_{k+8} \oplus S_{k+9}, k \ge 40.$$

The WG-8 transformation module WGT-8(x^{19}) comprises of two sub-modules: a WG-8 permutation module WGP-8(x^{19}) followed by a trace computation module Tr(·). While the WGP-8(x^{19}) module permutes elements over $\mathbb{F}_{2^8}$, the Tr(·) module compresses an 8-bit input to 1-bit keystream.

2.3 Randomness Properties of the WG-8 Keystream

The keystream generated by the stream cipher WG-8 has the following desired randomness properties [8]:

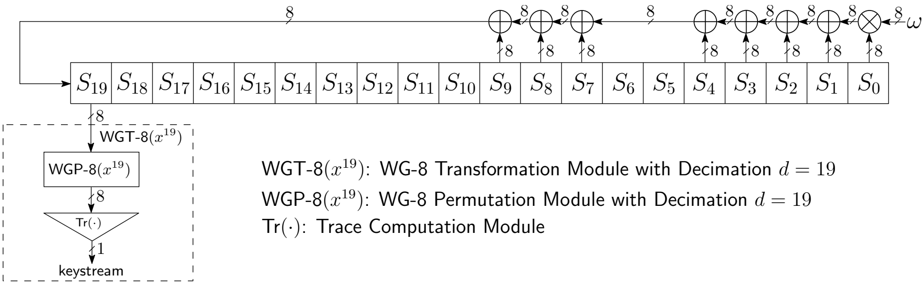

8
8
8
8
8
8
8
8
ω
8
8
8
8
8
8
8
8
S_{19} S_{18} S_{17} S_{16} S_{15} S_{14} S_{13} S_{12} S_{11} S_{10} S_9 S_8 S_7 S_6 S_5 S_4 S_3 S_2 S_1 S_0
8
WGT-8(x^{19})
WGP-8(x^{19})
8
Tr(·)
1
keystream
WGT-8(x^{19}): WG-8 Transformation Module with Decimation $d = 19$
WGP-8(x^{19}): WG-8 Permutation Module with Decimation $d = 19$
Tr(·): Trace Computation Module

Fig. 2. The Running Phase of the Stream Cipher WG-8

1. The keystream has a period of $2^{160} - 1$.
2. The keystream is balanced, i.e., the number of 0's is only one less than the number of 1's in one period of the keystream.
3. The keystream is an ideal two-level autocorrelation sequence.
4. The keystream has an ideal t-tuple ($1 \leq t \leq 20$) distribution, i.e., every possible output t-tuple is equally likely to occur in one period of the keystream.
5. The linear span of the keystream can be determined exactly, which is $2^{33.32}$.

3 Cryptanalysis of the Stream Cipher WG-8

In this section, we analyze the security of the stream cipher WG-8 under the context of lightweight embedded applications.

3.1 Algebraic Attack

The algebraic attack is a powerful attack against LFSR based filtering sequence generators [11]. The goal of the algebraic attack is to form a lower degree multivariate equation by multiplying the filtering function by a low-degree multivariate polynomial. This gives an overdefined system of nonlinear equations for sufficiently many keystreams, which can be solved to recover the internal state of the LFSR. The algebraic immunity of the WGT-8(x^{19}) is equal to 4. According to the algebraic attack, the time complexity and the data complexity for recovering the internal state of the LFSR are about $\frac{7}{64} \cdot \binom{160}{4}^{\log_2 7} = 2^{66.0037}$ and $\binom{160}{4} = 2^{24.65}$, respectively. For applying the fast algebraic attacks [10] to the stream cipher WG-8, one needs to respectively find two multivariate polynomials g and h of degree e and d ($e < d$) such that $f \cdot g = h$. For the WGT-8(x^{19}) and $e = 1$, there does not exist a multivariate polynomial h in 8 variables with degree less than 7. Hence, in order to launch the fast algebraic attack one needs to obtain more keystream bits with a higher complexity. For lightweight embedded applications, it is hard for an attacker to obtain about $2^{24.65}$ keystream bits. Even if the attacker can get those many bits for a fixed key and IV, he needs to perform the operations with the time complexity $2^{66.0037}$, which completely defeats this attack.

3.2 Correlation Attack

In the correlation attack, the objective of an attacker is either to find the correlation between a keystream and an output sequence of an LFSR or to find the correlation among the keystreams [9,27,37]. The stream cipher WG-8 is secure against the correlation among the keystreams as it produces keystreams with 2-level autocorrelation. We now consider the fast correlation attack in which the keystream of the stream cipher is considered as a distorted version of the LFSR output. In the fast correlation attack, the linear approximation of WGT-8(x^{19}) can be used to derive a generator matrix of a linear code that can be decoded by a maximum likelihood decoding (MLD) algorithm. Letting $f(x)$ be a linear function in 8 variables, we have $\Pr(\text{WGT-8}(x^{19})(x) = f(x)) = \frac{(2^8-108)}{2^8} = 0.578125$. Applying the results of [9] for $t = 3$, the amount of keystream (denoted by N) required for the attack to be successful is given by $N \approx (k \cdot 12 \cdot \ln 2)^{\frac{1}{3}} \cdot \epsilon^{-2} \cdot 2^{\frac{160-k}{3}}$ and the decoding complexity is given by $C_{dec} = 2^k \cdot k \cdot \frac{2 \ln 2}{(2\epsilon)^6}$, where $\epsilon = (\Pr(\text{WGT-8}(x^{19}) = f(x)) - 0.5) = 0.078125$ and k is the number of LFSR internal state bits recovered. If we choose a small value of k (e.g., $k = 7$), the number of bits required to launch the attack is about $2^{60.31}$, which is not possible in practice. Similarly, if we choose a large value of k (e.g., $k = 80$), the number of bits required to mount the attack is about $2^{37.15}$. However, the decoding complexity of the attack is approximately $2^{102.68}$, which is worse than the exhaustive search. Hence, the stream cipher WG-8 is secure against the fast correlation attack.

3.3 Differential Attack

The initialization phase in the first design of the WG stream cipher was vulnerable to the chosen IV attack [40], where an attacker can distinguish several output bits by building a distinguisher based on the differential cryptanalysis. This weakness has been fixed in the later design by placing the WG permutation module at the last position of the LFSR [29]. For the proposed stream cipher WG-8, the differential distribution of the WGP-8(x^{19}) is 8-uniform, which provides a maximum 2^{-5} possibility for differential characteristic. During the initialization phase the WGP-8(x^{19}) is applied for 40 times. Thus, after the initialization phase, it would be quite hard for an attacker to distinguish the output keystream since the differentials become complex and contain most key/IV bits.

3.4 Cube Attack

Cube attack [12] is a generic key-recovery attack that can be applied to any cryptosystem, provided that the attacker can obtain a bit of information that can be represented by a low-degree decomposition multivariate polynomial in Algebraic Normal Form (ANF) of the secret and public variables of the target cryptosystem. Note that the nonlinearity of WGP-8(x^{19}) is 92 and the algebraic degrees of the component functions of WGP-8(x^{19}) are 7. Moreover, the ANF representations of 8 component functions contain 133, 113, 146, 124, 137, 109, 122, and 120 terms, respectively, and only the ANF of the second component contains 7 linear terms and other terms are of degree greater than or equal to 2. In the WG-8 stream cipher, after 40 rounds of the initialization phase, the degree of the output polynomial can be very high. As a result, it would be hard for an attacker to collect low-degree relations among the secret key bits.

3.5 Distinguishing Attack

Recently, a distinguishing attack has been proposed against the stream cipher WG-7 [30]. Due to the small number of tap positions in the LFSR of the WG-7, the characteristic polynomial of the LFSR allows an attacker to build a distinguisher for distinguishing a keystream generated by WG-7 from a truly random keystream. For the WG-8 cipher, the characteristic polynomial of the LFSR consists of 8 tap positions and a similar distinguisher as in [30] can be built as

$$F(S_i, ..., S_{i+4}, S_{i+7}, ..., S_{i+9}) = \mathsf{WGT\text{-}8}(\omega \otimes S_i \oplus S_{i+1} \oplus S_{i+2} \oplus S_{i+3} \oplus S_{i+4} \oplus S_{i+7} \\ \oplus S_{i+8} \oplus S_{i+9}) \oplus \mathsf{WGT\text{-}8}(S_i) \oplus \mathsf{WGT\text{-}8}(S_{i+1}) \oplus \mathsf{WGT\text{-}8}(S_{i+2}) \oplus \mathsf{WGT\text{-}8}(S_{i+3}) \oplus \\ \mathsf{WGT\text{-}8}(S_{i+4}) \oplus \mathsf{WGT\text{-}8}(S_{i+7}) \oplus \mathsf{WGT\text{-}8}(S_{i+8}) \oplus \mathsf{WGT\text{-}8}(S_{i+9}),$$

which is a Boolean function in 64 variables. For the distinguisher F, the probability $\Pr(F(x) = 0) = \frac{1}{2} \pm \epsilon$, where $x = (a_0, ..., a_7), a_i \in \mathbb{F}_{2^8}$. Note that the value of ϵ will be quite small due to a huge number of variables in the distinguisher, which requires an attacker to obtain more keystream bits for distinguishing the keystream. However, the computation of the exact value of ϵ is infeasible in this case because the number of possible values of x is 2^{64}. Hence the WG-8 stream cipher is resistant to the distinguishing attack. Note that this type of distinguishing attacks can also be extended to the case in which a distinguisher can be built using a linear relation of a remote term of the LFSR, say S_τ for not large τ, and the sequences addressed in a subset of tap positions of the LFSR, denoted by $I = \{i_1, \cdots, i_t\} \subset \{0, 1, \cdots, 19\}$. In other words, a distinguisher could be built using the linear relation $S_\tau = S_{i_1} + \cdots + S_{i_t}$. Since this property is controlled by the characteristic polynomial of the LFSR, it can be easily teared done by a proper selection of the characteristic polynomial of the LFSR. For our selection of the characteristic polynomial $l(x)$, there is no remote term S_τ for $20 \leq \tau \leq 2^{34}$ for which the size of set I is less than 5. Thus, the WG-8 stream cipher is also resistant to this general distinguishing attack.

3.6 Discrete Fourier Transform Attack

The Discrete Fourier Transform (DFT) attack is a new type of attack to recover the internal state of a filtering generator, which was first proposed by Rønjom and Helleseth in [34] and extended to attacking filtering generators over $\mathbb{F}_{2^n}$ by Gong *et al.* in [19]. For mounting the DFT attack against the WG-8 stream cipher, an attacker needs to obtain $2^{33.32}$ (i.e., the linear complexity) consecutive keystream bits. Hence, the online complexity of this attack for recovering the internal state is $2^{33.32}$, after an offline computation with complexity $2^{48.49}$. For typical lightweight embedded applications like RFID systems, a reader and a tag only exchange 32-bit random numbers in each communication session. Hence, an attacker can never obtain $2^{33.32}$ consecutive keystream bits.

3.7 Time-Memory-Data Tradeoff Attack

The Time-Memory-Data (TMD) tradeoff attack [4] is a generic cryptanalytic attack that is applicable to any stream cipher, especially those with low sampling resistance. The complexity of the TMD tradeoff attack is $O(2^{\frac{n}{2}})$, where n is the size of the internal state. For the WG-8 stream cipher, the size of the internal state is 160-bit and thus the complexity of launching a TMD attack is at least 2^{80}. Moreover, the sampling resistance of the WG-8 stream cipher is high due to the usage of the WGT-8(x^{19}) as the filtering function. The ANF representation of the WGT-8(x^{19}) contains 109 terms, among which only four terms are linear and other terms have degree greater than 2 and less than 8. Hence, only by fixing 7 out of 8 variables can one obtain a linear equation.

4 Efficient Implementation of the Stream Cipher WG-8

In this section, we address efficient implementation of the WG-8 cipher on low-power microcontrollers. For each platform we provide three implementation variants that deal with trade-offs among speed, code size, and energy consumption.

4.1 Implementation of the WG-8 Permutation Module WGP-8(x^{19})

The most complicated WGP-8(x^{19}) module can be implemented using three different methods: a) a 256-byte direct look-up table; b) a 34-byte coset leader based look-up table; or c) tower field (TF) arithmetic.

Directly Look-up Table (DLT) Approach. Depending on the bases used, one can precompute the WG-8 permutation with decimation $d = 19$ by

$$\mathsf{WGP\text{-}8}(x^{19}) = q(x^{19} + 1) + 1$$

for all elements $x \in \mathbb{F}_{2^8}$. Hence, a 256-byte look-up table $T_{\mathsf{WGP\text{-}8}}$ can be generated to compute WGP-8(x^{19}).

Coset Leader Based Look-up Table (CLT) Approach. This approach assumes that a normal basis is used to represent elements in $\mathbb{F}_{2^8}$ and uses the essential property of the WG-8 permutation with decimation d below:

$$\begin{aligned}\mathsf{WGP\text{-}8}\left((x^{2^i})^d\right) &= q\left((x^{2^i})^d + 1\right) + 1 = q\left((x^d)^{2^i} + 1\right) + 1 \\ &= \big(q(x^d + 1)\big)^{2^i} + 1 = \big(q(x^d + 1) + 1\big)^{2^i} = \big(\mathsf{WGP\text{-}8}(x^d)\big)^{2^i} \end{aligned} \qquad (1)$$

for $x \in \mathbb{F}_{2^8}$ and $i = 0, 1, \ldots, 7$. According to the Equation (1), if we know the WG-8 permutation WGP-8(x^d) for an element $x \in \mathbb{F}_{2^8}$, we can easily obtain the WG-8 permutation WGP-8$((x^{2^i})^d)$ for the entire conset $\{x^2, x^{2^2}, \ldots, x^{2^7}\}$ of x by cyclically shifting WGP-8(x^d) to the right by i positions, provided that a normal basis is employed to represent finite field elements. The complete cosets and coset leaders of $\mathbb{F}_{2^8}$ (in hexadecimal notation) are shown in Table 1. We note that under the normal basis representation the elements in $\mathbb{F}_{2^8}$ have been grouped into 34 different cosets except for 0 and 1. Since WGP-8(0) = 0x00 and WGP-8(1) = 0xFF, we only need to generate a 34-byte look-up table $T_{\text{Co-WGP-8}}$ for storing the WG-8 permutation results for each coset leader. Here we present the following Algorithm 1 that uses the table $T_{\text{Co-WGP-8}}$ to compute WGP-8(x^d) for any $x \in \mathbb{F}_{2^8}$.

Table 1. The Cosets and Coset Leaders of $\mathbb{F}_{2^8}$

Coset Leader	Coset							Coset Leader	Coset						
0x00	–	–	–	–	–	–	–	0x27	0x4E	0x9C	0x39	0x72	0xE4	0xC9	0x93
0x01	0x02	0x04	0x08	0x10	0x20	0x40	0x40	0x2B	0x56	0xAC	0x59	0xB2	0x65	0xCA	0x95
0x03	0x06	0x0C	0x18	0x30	0x60	0xC0	0x81	0x2D	0x5A	0xB4	0x69	0xD2	0xA5	0x4B	0x96
0x05	0x0A	0x14	0x28	0x50	0xA0	0x41	0x82	0x2F	0x5E	0xBC	0x79	0xF2	0xE5	0xCB	0x97
0x07	0x0E	0x1C	0x38	0x70	0xE0	0xC1	0x83	0x33	0x66	0xCC	0x99	–	–	–	–
0x09	0x12	0x24	0x48	0x90	0x21	0x42	0x84	0x35	0x6A	0xD4	0xA9	0x53	0xA6	0x4D	0x9A
0x0B	0x16	0x2C	0x58	0xB0	0x61	0xC2	0x85	0x37	0x6E	0xDC	0xB9	0x73	0xE6	0xCD	0x9B
0x0D	0x1A	0x34	0x68	0xD0	0xA1	0x43	0x86	0x3B	0x76	0xEC	0xD9	0xB3	0x67	0xCE	0x9D
0x0F	0x1E	0x3C	0x78	0xF0	0xE1	0xC3	0x87	0x3D	0x74	0xF4	0xE9	0xD3	0xA7	0x4F	0x9E
0x11	0x22	0x44	0x88	–	–	–	–	0x3F	0x7E	0xFC	0xF9	0xF3	0xE7	0xCF	0x9F
0x13	0x26	0x4C	0x98	0x31	0x62	0xC4	0x89	0x55	0xAA	–	–	–	–	–	–
0x15	0x2A	0x54	0xA8	0x51	0xA2	0x45	0x8A	0x57	0xAE	0x5D	0xBA	0x75	0xEA	0xD5	0xAB
0x17	0x2E	0x5C	0xB8	0x71	0xE2	0xC5	0x8B	0x5B	0xB6	0x6D	0xDA	0xB5	0x6B	0xD6	0xAD
0x19	0x23	0x64	0xC8	0x91	0x23	0x46	0x8C	0x5F	0xBE	0x7D	0xFA	0xF5	0xEB	0xD7	0xAF
0x1B	0x36	0x6C	0xD8	0xB1	0x63	0xC6	0x8D	0x6F	0xDE	0xBD	0x7B	0xF6	0xED	0xDB	0xB7
0x1D	0x3A	0x74	0xE8	0xD1	0xA3	0x47	0x8E	0x77	0xEE	0xDD	0xBB	–	–	–	–
0x1F	0x3E	0x7C	0xF8	0xF1	0xE3	0xC7	0x8F	0x7F	0xFE	0xFD	0xFB	0xF7	0xEF	0xDF	0xBF
0x25	0x4A	0x94	0x29	0x52	0xA4	0x49	0x92	0xFF	–	–	–	–	–	–	–

Algorithm 1. Coset Leader Based Look-up Table Approach

Input: $x \in F_{2^8}$, a decimation d, a look-up table $T_{\text{Co-WGP-8}}$
Output: WGP-8(x^d)

```
if $x$ = 0x00 or $x$ = 0xFF then
    return $x$
end if
Find the coset leader $x_c$ of $x$ by cyclically shifting $x$ to the right by $i$ positions,
  where $0 \le i \le 7$ (i.e., $x_c$ is the smallest odd integer in the coset containing $x$.)
Find the position $j$ of $x_c$ being in the table $T_{\text{Co-WGP-8}}$
$a \leftarrow T_{\text{Co-WGP-8}}[j]$
return $a \lll i$
```

Tower Field Arithmetic (TFA) Based Approach. The software implementation of the WGP-8(x^{19}) module involves the arithmetic (i.e., addition, multiplication, and exponentiation) over finite field $\mathbb{F}_{2^8}$. Although we can directly implement all the operations over $\mathbb{F}_{2^8}$, it is well known that using the isomorphic tower constructions of $\mathbb{F}_{2^8}$ might save the memory consumption. Therefore, we investigate the tower construction $\mathbb{F}_{(2^4)^2}$ in this work.

Tower Construction $\mathbb{F}_{(2^4)^2}$ *and Its Arithmetic.* To obtain the tower construction $\mathbb{F}_{(2^4)^2}$, we first construct $\mathbb{F}_{2^4}$ by using an irreducible polynomial $e(X)$ of degree 4 over $\mathbb{F}_2$, and then construct $\mathbb{F}_{(2^4)^2}$ by using a certain irreducible polynomial $f(X)$ of degree 2 over $\mathbb{F}_{2^4}$. In our tower construction, we use $e(X) = X^4+X^3+1$ with its polynomial basis $\{1, \alpha, \alpha^2, \alpha^3\}$ for $\mathbb{F}_{2^4}$ and $f(X) = X^2 + X + \alpha$ with its normal basis $\{\beta, \beta^{16}\}$ for $\mathbb{F}_{(2^4)^2}$, where $\alpha = \omega^{119} \in \mathbb{F}_{2^4}$ and $\beta = \omega^7 \in \mathbb{F}_{(2^4)^2}$ are zeros of the polynomials $e(X)$ and $f(X)$, respectively.

Arithmetic operations in $\mathbb{F}_{2^4}$. The arithmetic in $\mathbb{F}_{2^4}$ is conducted with the aid of a 4 × 4 exponentiation table T_{exp} and a 4 × 4 logarithm table T_{log}. While the table T_{exp} stores exponentiation $\alpha^i, i = 0, 1, \ldots, 14$, the table T_{log} keeps the exponent i for each $\alpha^i, i = 0, 1, \ldots, 14$. Let $A = a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3$ and $B = b_0 + b_1\alpha + b_2\alpha^2 + b_3\alpha^3$ be two non-zero elements in $\mathbb{F}_{2^4}$, where $a_i, b_i \in \mathbb{F}_2, i = 0, 1, 2, 3$. We can perform the arithmetic in $\mathbb{F}_{2^4}$ as follows:

$$\begin{aligned} AB &= T_{exp}[(T_{log}[(a_0, a_1, a_2, a_3)] + T_{log}[(b_0, b_1, b_2, b_3)]) \bmod 15], \\ A^2 &= T_{exp}[(T_{log}[(a_0, a_1, a_2, a_3)] \ll 1) \bmod 15], \\ \alpha A &= T_{exp}[(T_{log}[(a_0, a_1, a_2, a_3)] + 1) \bmod 15]. \end{aligned}$$

Arithmetic operations in $\mathbb{F}_{(2^4)^2}$. Let $A = a_0\beta + a_1\beta^{16}$ and $B = b_0\beta + b_1\beta^{16}$, where $a_0, a_1, b_0, b_1 \in \mathbb{F}_{2^4}$. A multiplication AB in $\mathbb{F}_{(2^4)^2}$ is computed as follows:

$$AB = (a_0\beta + a_1\beta^{16})(b_0\beta + b_1\beta^{16}) = (c\alpha \oplus a_0b_0)\beta + (c\alpha \oplus a_1b_1)\beta^{16},$$

where $c = (a_0 \oplus a_1)(b_0 \oplus b_1)$. For a non-zero element $A \in \mathbb{F}_{(2^4)^2}$, the squaring of A is calculated as follows:

$$A^2 = (a_0\beta + a_1\beta^{16})^2 = [(a_0 \oplus a_1)^2\alpha \oplus a_0^2]\beta + [(a_0 \oplus a_1)^2\alpha \oplus a_1^2]\beta^{16}.$$

The Frobenius mapping of A with respect to $\mathbb{F}_{2^4}$, which is the 16th power operation, is computed as follows:

$$A^{2^4} = (a_0\beta + a_1\beta^{16})^{16} = a_0\beta^{16} + a_1\beta^{256} = a_1\beta + a_0\beta^{16}.$$

Implementation of WGP-8(x^{19}) *Module.* For an element $x \in \mathbb{F}_{2^8}$, the WGP-8(x^{19}) can be computed as follows:

$$\mathsf{WGP\text{-}8}(x^{19}) = q(x^{19}+1)+1 = y + y^{2^3+1} + y^{2^6}(y^{2^3+1} + y^{2^3-1}) + y^{2^3(2^3-1)+1} + 1,$$

where $y = x^{19} + 1 = x^{2^4} \cdot x^2 \cdot x + 1$. Note that for the tower construction $\mathbb{F}_{(2^4)^2}$, 1 can be denoted by the vector $(1, 0, 0, 0, 1, 0, 0, 0)$. Therefore, the addition with 1 under the TF representation is equivalent to XORing with a constant 0x88.

Table 2. Trace Computation of an Element $x \in F_{2^8}$ Using Different Bases

Basis	Element Representation	$\mathrm{Tr}(x)$
Polynomial Basis (PB)	$x_0 + x_1\omega + \cdots + x_7\omega^7$	x_5
Normal Basis (NB)	$x_0\theta + x_1\theta^2 + \cdots + x_7\theta^{2^7}$	$\bigoplus_{i=0}^{7} x_i$
Tower Field (TF)	$(x_0 + x_1\alpha + x_2\alpha^2 + x_3\alpha^3)\beta +$ $(x_4 + x_5\alpha + x_6\alpha^2 + x_7\alpha^3)\beta^{16}$	$x_1 \oplus x_2 \oplus x_3 \oplus x_5 \oplus x_6 \oplus x_7$

4.2 Implementation of the Trace Computation Module Tr(·)

Depending on the bases chosen, the trace of an element $x \in F_{2^8}$ can be computed as shown in Table 4.2.

4.3 Implementation of the Multiplication by ω Module

The multiplication by ω module can be implemented using either finite field arithmetic or an 8 × 8 look-up table.

Multiplication by ω Using Finite Field Arithmetic. We consider the following three cases when the PB, NB, and TF are used to represent finite field elements, respectively. With the PB representation, the multiplication of an element $x \in \mathbb{F}_{2^8}$ by ω can be computed as follows:

$$\begin{aligned} x \cdot \omega &= x_0\omega + x_1\omega^2 + \cdots + x_6\omega^7 + x_7\omega^8 \\ &= x_7 + x_0\omega + (x_1 \oplus x_7)\omega^2 + (x_2 \oplus x_7)\omega^3 + \\ &\quad (x_3 \oplus x_7)\omega^4 + x_4\omega^5 + x_5\omega^6 + x_6\omega^7. \end{aligned} \tag{2}$$

Therefore, the result of $x \cdot \omega$ is represented as an 8-bit vector $(x_7, x_0, x_1 \oplus x_7, x_2 \oplus x_7, x_3 \oplus x_7, x_4, x_5, x_6)$ with respect the PB.

With the NB representation, the multiplication of an element $x \in \mathbb{F}_{2^8}$ by ω can be calculated as follows:

$$x \cdot \omega = (x_0\theta + x_1\theta^2 + \cdots + x_6\theta^{2^6} + x_7\theta^{2^7}) \cdot \omega = \mathbf{M} \cdot (x_0, x_1, \cdots, x_6, x_7)^T, \tag{3}$$

where the matrix $\mathbf{M}$ is given below.

With the TF representation, the multiplication of an element $x \in \mathbb{F}_{2^8}$ by ω can be calculated as follows:

$$\begin{aligned} x \cdot \omega &= [(x_0 + x_1\alpha + x_2\alpha^2 + x_3\alpha^3)\beta + (x_4 + x_5\alpha + x_6\alpha^2 + x_7\alpha^3)\beta^{16}] \cdot \omega \\ &= \mathbf{M}' \cdot (x_0, x_1, \cdots, x_6, x_7)^T, \end{aligned} \tag{4}$$

where the matrix $\mathbf{M}'$ is given below.

$$\mathbf{M} = \begin{pmatrix} 1&1&1&0&1&0&1&1 \\ 0&0&0&0&1&1&1&0 \\ 1&0&1&0&1&0&0&1 \\ 1&0&1&1&1&0&0&0 \\ 0&0&1&0&1&1&1&0 \\ 0&1&1&0&0&1&1&1 \\ 1&0&1&1&1&1&0&0 \\ 0&1&1&0&1&0&1&1 \end{pmatrix} \quad \text{and} \quad \mathbf{M}' = \begin{pmatrix} 1&0&1&1&1&0&0&1 \\ 0&1&0&1&1&1&0&0 \\ 1&0&1&0&0&1&1&0 \\ 0&1&1&0&0&0&1&0 \\ 1&0&0&1&0&1&1&1 \\ 1&1&0&0&0&0&1&1 \\ 0&1&1&0&0&0&0&1 \\ 0&0&1&0&1&1&1&1 \end{pmatrix}.$$

Multiplication by ω Using Look-Up Tables. Based on the Equations (2)–(4), one can generate 256-byte look-up tables with respect to the chosen bases.

4.4 Implementation Platforms and Development Tools

In this section, we briefly describe two low-power microcontrollers for implementing the WG-8 stream cipher as well as the corresponding development tools.

8-Bit Microcontroller ATmega128L and Development Tool. The low-power 8-bit microcontroller ATmega128L [1] from Atmel is based on the AVR enhanced RISC architecture with 128 Kbytes of In-System Self-Programmable Flash, 4 Kbytes EEPROM and 8 Kbytes Internal SRAM. It is equipped with 133 highly-optimized instructions and most of them can be executed within one clock cycle. Moreover, the clock frequency of the ATmega128L can run from 0 to 8 MHz and the power supplies can go from 2.7 to 5.5 V. We use the latest integrated development environment Atmel Studio 6.0 [2] from Atmel for implementing and testing the performance of the WG-8 on the target platform.

16-Bit Microcontroller MSP430F1611 and Development Tool. The 16-bit microcontroller MSP430F1611 [38] from Texas Instruments has a traditional von-Neumann architecture with 48 Kbytes Flash memory and 10 Kbytes RAM. All special function registers, peripherals, RAM and Flash/ROM share the same address space. The clock frequency of the MSP430F1611 ranges from 0 to 8 MHz and the power supplies can go from 1.8 to 3.6 V. The MSP430F1611 features 27 instructions and 7 different addressing modes that provide great flexibility in data manipulation. To implement and simulate the WG-8 on the target platform, we use the CrossWorks for MSP430 Version 2.1 from Rowley Associates [35].

4.5 Experimental Results and Comparisons

In this section, we report our experimental results for implementing the stream cipher WG-8 on the low-power microcontrollers ATmega128L and MSP430F1611 and compare our results with other lightweight-cryptography implementations on the same or similar platforms. We focus on three major performance criteria for implementing cryptographic primitives on resource-constrained environments, namely throughput, code size, and energy consumption (i.e., energy/bit). Table 3 compares our implementation results with previous work in terms of the aforementioned three performance criteria. Note that we estimate the per bit energy consumptions by the formula: energy/bit $= \frac{\text{Supply Voltage}\times\text{Current}\times\text{Cycles}}{\text{Clock Frequency}\times\text{Number of Bits}}$, which is based on the typical current consumption of a low-power microcontroller for the given clock frequency and supply voltage.

From Table 3, we note that on 8-bit ATmega microcontrollers the throughput of WG-8 is about $2 \sim 15$ times higher than that of stream ciphers Grain, Trivium, Salsa20, and WG-7, block ciphers PRESENT-80 and XTEA as well as the hybrid cipher Hummingbird, whereas the energy consumption of WG-8 is around

Table 3. Performance Comparison of Lightweight-Cryptography Implementations on Low-Power Microcontrollers

Low-Power Microcontroller	Cryptographic Primitive	Clock Freq. [MHz]	Opt. Goal/ Method	Memory Usage [byte]		Setup [cycle]	Throughput [Kbits/sec]	Energy/Bit [nJ]
				Flash	SRAM			
ATmega	AES [31]	8 MHz	RAM	1,912	176	789	475.6	179
			Speed	1,912	256	747	513.8	165
	PRESENT-80 [33]		Size	1,474	32	–	0.99	85,819
			Speed	2,398	528	–	66.7	1,274
	Hummingbird [15]		Size	1,308	–	14,735	34.9	2,433
			Speed	10,918	–	8,182	91.5	929
	Hummingbird-2 [16]		RAM	3,600	114	2,970	171.8	495
			Speed	3,200	1,500	1,800	258.6	329
	XTEA [32]		Speed	820	–	–	51.7	1,645
	Grain [32]		Speed	778	20	107,336	12.9	6,556
	Trivium [32]		Speed	424	36	775,726	12.0	7,066
	Salsa20 [28]		Speed	3,842	258	318	83.7	101,564
	WG-7 [26]		Size	938	–	20,917	34.0	2,497
	WG-8		TFA	2,450	20	99,702	3.58	23,739
			CLT	2,238	148	10,683	31.7	2,683
			DLT	**1,984**	**20**	**1,379**	**185.5**	**458**
MSP430	PRINTcipher-48 [18]	8 MHz	Speed	6,424	48	–	4.5	153
	AES [18]		Speed	10,898	218	–	78.0	154
	PRESENT-80 [18]		Speed	6,424	288	–	19.4	619
	KLEIN-64 [18]		Speed	6,424	288	–	65.0	185
	Hummingbird [15]		Size	1,064	–	9,667	53.0	226
			Speed	1,360	–	4,824	104.9	114
	Hummingbird-2 [16]		Size	770	50	5,984	84.2	143
			Speed	3,648	114	1,361	356.5	34
	WG-7 [26]		Size	1,050	–	18,379	21.0	572
	WG-8		TFA	2,110	20	127,944	2.44	4,926
			CLT	2,628	148	15,265	10.8	1,107
			DLT	**1,558**	**20**	**3,604**	**95.9**	**125**

$2 \sim 220$ times smaller than that of those ciphers. Moreover, WG-8 has the comparable throughput and energy efficiency with the hybrid cipher Hummingbird-2 (optimized with assembly language). On the 8-bit platform, WG-8 is less efficient than AES in terms of throughput and energy consumption. The main reason is that WG-8 is a bit-oriented stream cipher whereas AES is a block cipher with block size 128-bit. Furthermore, the code size of WG-8 is medium and the SRAM usage of WG-8 is small among all the lightweight implementations.

On 16-bit MSP430 microcontrollers, the throughput of WG-8 is about $1 \sim 20$ times higher than that of the stream cipher WG-7 as well as block ciphers PRINTcipher-48, AES, PRESENT-80, and KLEIN-64, whereas the energy efficiency is comparable with that of those ciphers. While WG-8 has similar throughput and energy efficiency as the hybrid cipher Hummingbird, it is less efficient when compared to the Hummingbird-2 cipher. The main reason comes from the optimization with the assembly language in the speed-optimized Hummingbird-2 implementation. Furthermore, the code size of WG-8 is about $2 \sim 7$ times smaller than block ciphers PRINTcipher-48, AES, PRESENT-80, and KLEIN-64 as well as the hybrid cipher Hummingbird-2, and is comparable with the Hummingbird cipher. Regarding to the SRAM usage, the stream cipher WG-8 is superior to other block cipher and stream ciphers.

In addition, for the three implementation variants, we note that on both 8-bit and 16-bit platforms the DLT method is consistently better than both CLT and TFA methods with respect to throughput and energy consumption. The reason lies in the efficient memory access for look-up tables on both microcontrollers.

5 Conclusion

In this paper, we present a lightweight stream cipher WG-8 targeted for resource-constrained devices like RFID tags, smart cards, and wireless sensor nodes, which inherits all the good randomness and cryptographic properties of the well-known WG stream cipher family. A detailed cryptanalysis shows that WG-8 is resistant to the most common attacks against stream ciphers. Moreover, the software implementations on low-power microcontrollers demonstrate the high performance and low energy consumption of the WG-8 stream cipher, when compared to most of previous block ciphers and stream ciphers. Therefore, the stream cipher WG-8 is a competitive candidate for securing pervasive embedded applications.

References

1. Atmel Corporation, ATmega128(L): 8-bit Atmel Microcontroller with 128 KBytes In-System Programmable Flash (2011), http://www.atmel.com/Images/doc2467.pdf
2. Atmel Corporation, Atmel Studio 6 – The Integrated Development Environment (2012), http://www.atmel.com/microsite/atmel_studio6/
3. Babbage, S., Dodd, M.: The Stream Cipher MICKEY 2.0, ECRYPT Stream Cipher (2006), http://www.ecrypt.eu.org/stream/p3ciphers/mickey/mickey_p3.pdf
4. Biryukov, A., Shamir, A.: Cryptanalytic Time/Memory/Data Tradeoffs for Stream Ciphers. In: Okamoto, T. (ed.) ASIACRYPT 2000. LNCS, vol. 1976, pp. 1–13. Springer, Heidelberg (2000)
5. Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A., Robshaw, M., Seurin, Y., Vikkelsoe, C.: PRESENT: An ultra-lightweight block cipher. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 450–466. Springer, Heidelberg (2007)
6. De Cannière, C., Dunkelman, O., Knežević, M.: KATAN and KTANTAN — A Family of Small and Efficient Hardware-Oriented Block Ciphers. In: Clavier, C., Gaj, K. (eds.) CHES 2009. LNCS, vol. 5747, pp. 272–288. Springer, Heidelberg (2009)
7. De Cannière, C., Preneel, B.: Trivium – A Stream Cipher Construction Inspired by Block Cipher Design Principles. ECRYPT Stream Cipher (2005), http://www.ecrypt.eu.org/stream/papersdir/2006/021.pdf
8. Chen, L., Gong, G.: Communication System Security. Chapman & Hall/CRC, Boca Raton (2012)
9. Chepyzhov, V.V., Johansson, T., Smeets, B.: A Simple Algorithm for Fast Correlation Attacks on Stream Ciphers. In: Schneier, B. (ed.) FSE 2000. LNCS, vol. 1978, pp. 181–195. Springer, Heidelberg (2001)
10. Courtois, N.T.: Fast Algebraic Attacks on Stream Ciphers with Linear Feedback. In: Boneh, D. (ed.) CRYPTO 2003. LNCS, vol. 2729, pp. 176–194. Springer, Heidelberg (2003)

11. Courtois, N., Meier, W.: Algebraic Attacks on Stream Ciphers with Linear Feedback. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 345–359. Springer, Heidelberg (2003)
12. Dinur, I., Shamir, A.: Cube Attacks on Tweakable Black Box Polynomials. In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 278–299. Springer, Heidelberg (2009)
13. Driessen, B., Hund, R., Willems, C., Paar, C., Holz, T.: Don't Trust Satellite Phones: A Security Analysis of Two Satphone Standards. In: The 33th IEEE Symposium on Security and Privacy - S&P 2012, pp. 128–142 (2012)
14. Eisenbarth, T., Kumar, S., Paar, C., Poschmann, A., Uhsadel, L.: A Survey of Lightweight-Cryptography Implementations. IEEE Design & Test of Computers 24(6), 522–533 (2007)
15. Engels, D., Fan, X., Gong, G., Hu, H., Smith, E.M.: Hummingbird: Ultra-Lightweight Cryptography for Resource- Constrained Devices. In: Sion, R., Curtmola, R., Dietrich, S., Kiayias, A., Miret, J.M., Sako, K., Sebé, F. (eds.) RLCPS, WECSR, and WLC 2010. LNCS, vol. 6054, pp. 3–18. Springer, Heidelberg (2010)
16. Engels, D., Saarinen, M.-J.O., Schweitzer, P., Smith, E.M.: The Hummingbird-2 Lightweight Authenticated Encryption Algorithm. In: Juels, A., Paar, C. (eds.) RFIDSec 2011. LNCS, vol. 7055, pp. 19–31. Springer, Heidelberg (2012)
17. Feldhofer, M., Wolkerstorfer, J., Rijmen, V.: AES Implementation on a Grain of Sand. IEE Proceedings Information Security 15(1), 13–20 (2005)
18. Gong, Z., Nikova, S., Law, Y.: KLEIN: A New Family of Lightweight Block Ciphers. In: Juels, A., Paar, C. (eds.) RFIDSec 2011. LNCS, vol. 7055, pp. 1–18. Springer, Heidelberg (2012)
19. Gong, G., Rønjom, S., Helleseth, T., Hu, H.: Fast Discrete Fourier Spectra Attacks on Stream Ciphers. IEEE Transactions on Information Theory 57(8), 5555–5565 (2011)
20. Guo, J., Peyrin, T., Poschmann, A., Robshaw, M.: The LED Block Cipher. In: Preneel, B., Takagi, T. (eds.) CHES 2011. LNCS, vol. 6917, pp. 326–341. Springer, Heidelberg (2011)
21. Hell, M., Johansson, T., Meier, W.: Grain: A Stream Cipher for Constrained Environments. International Journal of Wireless and Mobile Computing 2(1), 86–93 (2007)
22. Kaps, J.-P.: Chai-tea, Cryptographic Hardware Implementations of xTEA. In: Chowdhury, D.R., Rijmen, V., Das, A. (eds.) INDOCRYPT 2008. LNCS, vol. 5365, pp. 363–375. Springer, Heidelberg (2008)
23. Knudsen, L., Leander, G., Poschmann, A., Robshaw, M.J.B.: PRINTcipher: A Block Cipher for IC-Printing. In: Mangard, S., Standaert, F.-X. (eds.) CHES 2010. LNCS, vol. 6225, pp. 16–32. Springer, Heidelberg (2010)
24. Leander, G., Paar, C., Poschmann, A., Schramm, K.: New Lightweight DES Variants. In: Biryukov, A. (ed.) FSE 2007. LNCS, vol. 4593, pp. 196–210. Springer, Heidelberg (2007)
25. Liu, D., Yang, Y., Wang, J., Min, H.: A Mutual Authentication Protocol for RFID Using IDEA, Auto-ID Labs White Paper, WP-HARDWARE-048 (March 2009), http://www.autoidlabs.org/uploads/media/AUTOIDLABS-WP-HARDWARE-048.pdf
26. Luo, Y., Chai, Q., Gong, G., Lai, X.: WG-7: A Lightweight Stream Cipher with Good Cryptographic Properties. In: IEEE Global Communications Conference – GLOBECOM 2010, pp. 1–6 (2010)
27. Meier, W., Staffelbach, O.: Fast Correlation Attacks on Certain Stream Ciphers. Journal of Cryptology 1(3), 159–176 (1989)

28. Meiser, G., Eisenbarth, T., Lemke-Rust, K., Paar, C.: Efficient Implementation of eSTREAM Ciphers on 8-bit AVR Microcontrollers. In: International Symposium on Industrial Embedded Systems – SIES 2008, pp. 58–66 (2008)
29. Nawaz, Y., Gong, G.: WG: A Family of Stream Ciphers with Designed Randomness Properties. Information Science 178(7), 1903–1916 (2008)
30. Orumiehchiha, M.A., Pieprzyk, J., Steinfeld, R.: Cryptanalysis of WG-7: A Lightweight Stream Cipher. Cryptography and Communications 4(3-4), 277–285 (2012)
31. Osvik, D.A., Bos, J.W., Stefan, D., Canright, D.: Fast Software AES Encryption. In: Hong, S., Iwata, T. (eds.) FSE 2010. LNCS, vol. 6147, pp. 75–93. Springer, Heidelberg (2010)
32. Otte, D.: AVR-Crypto-Lib (2012), `http://www.das-labor.org/wiki/AVR-Crypto-Lib/en`
33. Poschmann, A.: Lightweight Cryptography – Cryptographic Engineering for a Pervasive World, Ph.D. Thesis, Department of Electrical Engineering and Information Science, Ruhr-Universitäet Bochum, Bochum, Germany (2009)
34. Rønjom, S., Helleseth, T.: A New Attack on the Filtering Generator. IEEE Transactions on Information Theory 53(5), 1752–1758 (2007)
35. Rowley Associates, CrossWorks for MSP430 (2012), `http://www.rowley.co.uk/msp430/`
36. Shibutani, K., Isobe, T., Hiwatari, H., Mitsuda, A., Akishita, T., Shirai, T.: *Piccolo*: An Ultra-Lightweight Blockcipher. In: Preneel, B., Takagi, T. (eds.) CHES 2011. LNCS, vol. 6917, pp. 342–357. Springer, Heidelberg (2011)
37. Siegenthaler, T.: Decrypting a Class of Stream Ciphers Using Ciphertext Only. IEEE Transactions on Computers 34(1), 81–85 (1985)
38. Texas Instuments Inc., MSP430F15x, MSP430F16x, MSP430F161x Mixed Signal Microcontroller (2011), `http://www.ti.com/lit/ds/symlink/msp430f1611.pdf`
39. Verdult, R., Garcia, F.D., Balasch, J.: Gone in 360 Seconds: Hijacking with Hitag2. In: The 21st USENIX Security Symposium - USENIX Security 2012, pp. 237–252. USENIX Association (2012)
40. Wu, H., Preneel, B.: Chosen IV Attack on Stream Cipher WG, ECRYPT Stream Cipher Project Report, 2005/045, `http://cr.yp.to/streamciphers/wg/045.pdf`

Protecting PLM Data Throughout Their Lifecycle

Rohit Ranchal and Bharat Bhargava

Purdue University, Computer Sciences and CERIAS
West Lafayette, IN 47906, USA
{rranchal,bbshail}@purdue.edu

Abstract. Enterprises operate in a global economy with their operations dispersed across internal processes and external partners. Product Lifecycle Management (PLM) systems play a significant role in modern product development and management. There are multiple stages in product lifecycle that streamline by sharing data among PLM entities. Shared data may contain highly sensitive information such as trade secrets, intellectual property, private organizational or personal information. In large enterprise systems, it is difficult to understand and track data dissemination. Data sharing across global partners complicates and magnifies the problem further. The effect of shared data being leaked is one of the key risks. Existing approaches ensure security within the domain of an organization and don't address protection in a decentralized environment. We propose an approach for secure data dissemination using the Active Bundle scheme. This approach enables organizations to securely share information in their PLM steps and protects it throughout the product lifecycle.

Keywords: PLM, active bundle, data dissemination, security, privacy.

1 Introduction

Modern product development is highly complex and increased competition has driven organizations to focus on core competences. A single organization can no longer efficiently handle product development and management in its entirety. Organizations focus on their expertise and outsource other activities to partner organizations specialized in these activities. This is done to achieve faster product development cycles, lower development costs and improve quality of service. Organizations rely on a complex web of distributed collaboration among internal business processes and external service providers to develop and deliver their products and services. This complex web of interactions is realized through Product Lifecycle Management (PLM) systems.

PLM is an information management solution that supports product development and management by streamlining the flow of product related information along all the stages of its lifecycle [1]. It provides a shared platform to connect various participating entities over the entire lifecycle of the product from concept to retirement. Product development is accompanied by many changes such as changes in customer demands, errors in design and planning, resource availability etc. Thus PLM

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 633–642, 2013.

deals with an enormous quantity of data flow. Effective collaboration and product management in a PLM system requires product data to be dynamically shared, readily accessible and inherently secure. This includes all the product lifecycle stages that plan, build, manage, maintain, service and protect the product. Each stage involves interactions among entities that use available information, generate new information and share it further. Each entity defines specific information usage requirements before sharing it further. Being highly sensitive in nature, this information is not only necessary to develop and deliver products but is also responsible for improving efficiency, driving business decisions and maintaining competitiveness in the market.

1.1 Information Flow in PLM

Modern PLM systems engage global partners and employ cross-domain information exchange, where the information is dispersed across multiple entities and not under the control of a single owner. PLM aims to provide control over this information and tools to manage the overall lifecycle. Organizations like to keep track of their information flow to know how their information is used, with whom it is shared, and what actions are applied to it. It can be safely assumed that complete control over information flow, information usage and information tracking is possible in a trusted domain. Each entity in PLM has its own trusted domain but when data leaves this domain, there is very little or no visibility and control over this data. This data can be further shared with other entities in the PLM stages making it impossible for the entities in the earlier stages to track or know about its current state. The magnitude of data, complexity of processes and global distribution of entities in the PLM systems further complicate data sharing and dissemination. It is very difficult to support data sharing across partners and ensure its security at the same time. The threat of shared data being compromised is one of the key risks in PLM systems. Unauthorized information disclosures or data leakage can lead to huge financial and business losses and can be threatening to an organization's reputation. Thus, it is imperative that the information in PLM systems be securely shared and protected according to its owner's policies leading to trustworthy PLM systems.

1.2 Motivation

There have been numerous cases of product management systems being compromised resulting in substantial damages to the organizations. One such incident happened with Foxconn, which assembles about 40 percent of the consumer electronics products in the world. The hackers penetrated the Foxconn network and stole sensitive data including contact details of Foxconn’s global sales managers, usernames, IP addresses, client e-mails and purchases. This data could have been used to place fraudulent orders from the Foxconn’s clients. Foxconn had to take its services offline to prevent further damage [3]. Even the most reputed companies such as Apple, HP, Sony etc have shipped pre-owned laptops, hard drives, and other devices with viruses, worms, and trojans on them. A recent report published by Verizon indicates that there has been an increase in the number of hacking attacks and data breaches across the globe [4]. According to the report there were 855 data breaches in 2011 that involved more than 174 million compromised records. The report also reflects the product development challenges for the organizations conducting global business. Organization insiders (the employees and contractors of an organization) account for 80% of computer and Internet crimes [5]. Security weaknesses in information dissemination are major threats that could be exploited by the malicious insiders [6]. PLM data resides in the globally connected networks and such incidents can have highly damaging consequences.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 describes the proposed approach. Section 4 presents the prototype implementation. Section 5 discusses the resilience of the proposed approach. Section 6 concludes the paper.

2 Related Work

Information in the PLM systems is scattered across various entities that participate in product development and management. One of the main challenges for PLM systems is getting complete control over the information. Tracking the information flow and ensuring its protection throughout its lifecycle is a significant issue with dynamic cross-domain information exchange. Each entity has specific security requirements. It is very difficult to know or compare the information security controls of actual and potential partners against the information owner's policies such as the capability of a partner to protect the information, level of protection actually being applied, outsourcing of information by the partner to other entities, compliance to regulatory and legal policies etc. Researchers have investigated issues with PLM in VIDOP project [7, 8]. In [9], the authors discuss security limitations of distributed PLM solutions. In [10], authors propose a solution based on the use of a Trusted Third Party (TTP) known as Management Entity. TTP is responsible for enforcing owner policies on the shared information. Substantial studies have been done to address information sharing and access control in the area of collaborative design in [11, 12] and supply chain in [2, 13, 14].

2.1 Issues with Current Solutions

Existing research lacks in addressing security issues with PLM. Available solutions typically operate in isolation and focus on data protection within the organization and do not extend to its partners. Security policies are defined to protect shared information, restrict access and regulate its usage. Each entity defines policies for its information. Examples of policies include: prevent data leaks to outsiders, enforce authorization to shared information, and control dissemination content. Information is protected according to the information holder's policies but there are often multiple information receivers for which the owner specifies multiple policies. These receivers can further disseminate the information. This requires policy communication, negotiation and enforcement in different (including unknown or untrusted) domains but the available solutions are unable to ensure the correct policy enforcement and control information dissemination in unknown or untrusted domains. Information security standards are constantly evolving and there are no global security standards for PLM systems. Multiple, overlapping, disparate standards and the differences in the implementation of security controls add more complexity to the protection mechanisms. This can result in duplication of security controls and inefficiencies but more importantly may leave security gaps in the information flow [15].

Organizations do vendor comparisons, rely on service level agreements or contracts, perform audits, or hire specialized companies to identify vulnerabilities, detect violations and ensure conformance of business processes to specific security and privacy requirements. Efficient manual verification is possible on processes in the same domain and could only be performed on processes that are composed of small number of activities and generate less data. Such approaches don't work in unknown or untrusted domains and are insufficient to ensure data security throughout its lifecycle. We propose a data-centric approach that addresses the above-mentioned issues. It provides a security blanket around the information that traverses the PLM entities and protects it throughout its lifecycle. It enables PLM entities to receive the respective information without revealing extra information.

3 Proposed Data Sharing and Dissemination Mechanism

The main challenges in information sharing mechanisms are their limitations to provide control over data after it leaves the trusted domain. Common protection mechanisms consider data as passive entities that are unable to protect themselves. They require another active and trusted entity– a trusted processor, a trusted memory module, a trusted application or a TTP to provide protection and ensure correct policy enforcement in foreign domains. But trusting this entity requires taking risks. If the trusted entity is compromised, all the sensitive information is also compromised. Moreover, trusted entities (except TTP) should be installed in all the PLM entities, which may not be feasible for external partners. We propose a data-centric approach

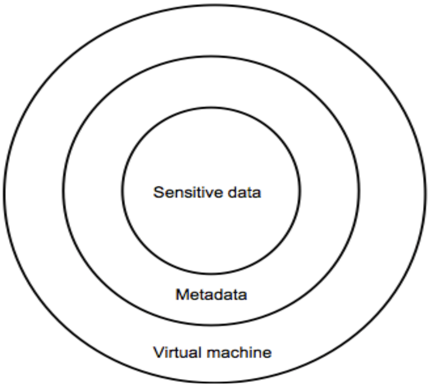


Fig. 1. Basic structure of an Active Bundle (AB)

that transforms passive data into an active entity that is able to protect itself. It enables dynamic data dissemination decisions. The granularity of the data being shared with an entity is determined by the respective dissemination policy of the data owner. The proposed approach is based on the Active Bundle (AB) scheme [l6, 17].

3.1 Overview of the Active Bundle Scheme

The active bundle is a robust and an extensible scheme that can be used to securely disseminate data across multiple domains. An Active Bundle (more detailed description can be found in [16, 17]), the structure of which is shown in Fig. 1, is a data protection mechanism that encapsulates sensitive data with metadata and a virtual machine.

Sensitive Data. It is the digital content that needs to be protected from privacy violations, data leaks and unauthorized disclosures. The digital content can include documents, pieces of code, images, audio, video files etc. The content in the sensitive data can have several sub-elements, each with different dissemination requirements. For instance, certain part of the data could be shared with marketing department such as product specifications, pricing etc and a different part of data could be shared with production department such as design documents.

Metadata. It describes the active bundle and its privacy policies. The metadata includes (but is not limited to) the following components (details available in [17]): (a) provenance metadata; (b) integrity check metadata; (c) access control metadata; (d) dissemination control metadata; (e) life duration value; (f) security metadata (including: security server id; encryption algorithm used by the Virtual Machine; encrypted pseudo-random number generator; trust server id used to validate the trust level and the role of a host; and trust level threshold required to access data in an active bundle); and (g) other application-dependent and context-dependent metadata. For instance, the access control metadata is used to ascertain the content for a specific receiver according to receiver's authorization.

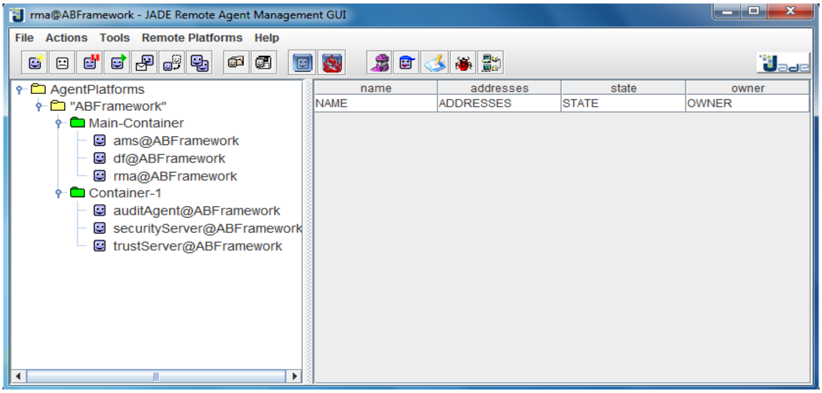
Virtual Machine (VM). It is the protection and policy enforcement mechanism that manages and controls the program code enclosed in a bundle. Its main functions include: (a) enforcing bundle access control policies through apoptosis (self-destruction) or data filtering (e.g., disclosing to a receiver only the portion of sensitive data that it is entitled to access); (b) enforcing bundle dissemination policies; and (c) validating bundle integrity.

3.2 Working of the Active Bundle Scheme

Unlike other solutions, the AB scheme does not require a client application on the receiver to execute its code. It can work like an applet, a jar file or a mobile agent. An active bundle is created with sensitive data, security policies and sent from one PLM entity to another. When arriving at the receiver entity, an active bundle ascertains the entity's trust level through a TTP. Using its disclosure policy, it decides whether the entity is eligible to access all or part of bundle’s data, and which portion of sensitive data can be revealed to it. An active bundle may realize that its security is about to be compromised. E.g., it may discover that its self-integrity check fails, or the trust level of the receiver entity is too low. In response, the bundle may choose to apoptosize, that is, perform atomically a clean self-destruction (one that is complete and leaves no traces usable for an attacker) [17]. An active bundle after completing its task on this entity travels to the next entity in the PLM chain. The information about next entity to be traversed can either be specified in advance in the bundle (e.g. during its creation) or dynamically decided.

4 Active Bundle Prototype Implementation

The AB prototype has been implemented using the mobile agent framework Jade (http://jade.tilab.com/) [17, 18]. A mobile agent is a software object that contains code and carried data and is able to perform computations on visited hosts, transport itself from one host to another, and interact with and use capabilities of visited hosts [17]. The system contains: AB Coordinator; AB Creator; AB Destination; Directory Facilitator (DF); AB Services: Security Services Agent (SSA), Trust Evaluation Agent (TEA), and Audit Services Agents (ASA); and an AB. The components are distributed among Jade containers. Fig. 2 shows the GUI for agent management. It shows the containers and registered agents (DF, SSA, TEA, ASA) in the system. They can be setup on a single host or distributed on different hosts.



- AgentPlatforms
  - "ABFramework"
    - Main-Container
      - ams@ABFramework
      - df@ABFramework
      - rma@ABFramework
    - Container-1
      - auditAgent@ABFramework
      - securityServer@ABFramework
      - trustServer@ABFramework

name	addresses	state	owner
NAME	ADDRESSES	STATE	OWNER

Fig. 2. AB Framework GUI using JADE

AB Coordinator hosts the Jade DF, providing a yellow pages service. It is used by agents to register/deregister their services, and to search for services and destinations.

AB Creator is an application that accepts from a user, input including sensitive data, metadata, and the destination and transforms it according to the attributes of AB's structure and includes the code for VM. Next, it registers the AB with DF.

AB Destination hosts a container and receives ABs sent by the creator.

AB Services are TTPs implemented as three agents: (a) SSA that contains security related information about ABs. It is used for encrypting and decrypting ABs. Each AB is described in SSA using the following information: name, decryption key, and the threshold trust level that a receiver must satisfy to access data from AB. (b) TEA answers requests from SSA about the trust level of a specified host, which could be obtained using a trust management system [17, 20]. (c) ASA records and monitors activities of ABs. It receives audit information from ABs, and records this information for analysis by authorized entities (e.g., AB owners, or auditors) and to support dynamic metadata updates.

AB is a mobile agent constructed by AB Creator that has a set of attributes and operations.

4.1 System Functioning

With the assumption that there is a secure communication channel between the AB and the TTPs, below we provide a description of AB prototype functioning.

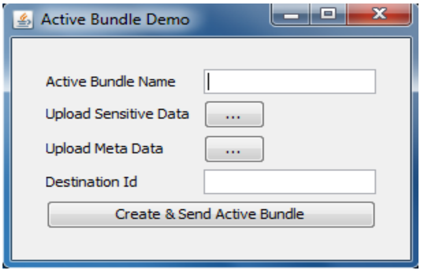
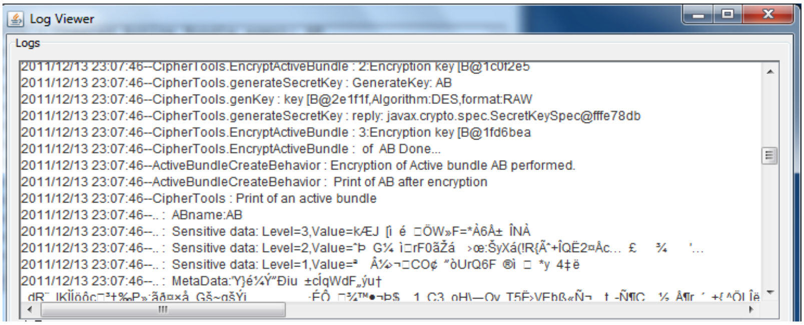


Fig. 3. AB creation GUI

Initialization of an AB: An owner of sensitive data provides AB Creator with sensitive data and metadata as shown in Fig. 3. Sensitive data contains multiple versions of the content to support selective and controlled dissemination based on access policies and receiver authorization. Simple custom DTDs are defined to specify sensitive data and metadata policies in XML format. AB Creator constructs an AB by putting together data, metadata, and a VM. After this stage, the AB becomes an active entity (since it has its own VM) that can perform the remaining steps.

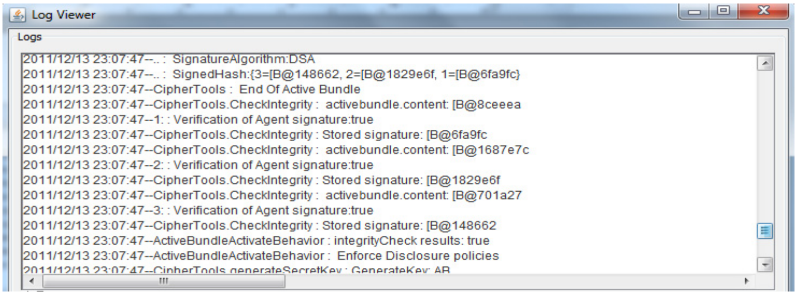


```
2011/12/13 23:07:46--CipherTools.EncryptActiveBundle : 2:Encryption key [B@1c0f2e5
2011/12/13 23:07:46--CipherTools.generateSecretKey : GenerateKey: AB
2011/12/13 23:07:46--CipherTools.genKey : key [B@2e1f1f,Algorithm:DES,format:RAW
2011/12/13 23:07:46--CipherTools.generateSecretKey : reply: javax.crypto.spec.SecretKeySpec@fffe78db
2011/12/13 23:07:46--CipherTools.EncryptActiveBundle : 3:Encryption key [B@1fd6bea
2011/12/13 23:07:46--CipherTools.EncryptActiveBundle :  of  AB Done...
2011/12/13 23:07:46--ActiveBundleCreateBehavior : Encryption of Active bundle AB performed.
2011/12/13 23:07:46--ActiveBundleCreateBehavior :  Print of AB after encryption
2011/12/13 23:07:46--CipherTools : Print of an active bundle
2011/12/13 23:07:46--.. :  ABname:AB
2011/12/13 23:07:46--.. :  Sensitive data: Level=3,Value=kÆJ  [ì  é  □ÖW»F=*À6Å±  ÎNÀ
2011/12/13 23:07:46--.. :  Sensitive data: Level=2,Value=ˆÞ  G¼  ì□rF0ãŽá   ›œ:ŠyXá(!R{Ã˜+ÎQË2¤Åc...  £     ¾      '...
2011/12/13 23:07:46--.. :  Sensitive data: Level=1,Value=ª   Â¼›¬□CO¢  ˝òUrQ6F  ®ì  □  *y  4‡ë
2011/12/13 23:07:46--.. :  MetaData:'Y}é¼Ý˝Ðiu  ±cÍqWdF„ýu†
 dR¯ IKÏîõôc□³†‰P»:ãð¤×å  Gš~gšÝi              ·ÉÔ  □¾™•¬Þ$   1  C3  oH\—Ov  T5Ê›VEhß«Ñ¬   †  -Ñ¶C    ½  Å¶r  ´  +{ ^ÖI Îë
```

Fig. 4. AB processing logs GUI

Building an AB: The steps taken in the process of building an AB are as follows (Fig. 4 shows the logs for this step):

- The AB gets two pairs of public/private keys from Security Service Agent (SSA) for each version of the sensitive data, where the first pair of keys is used for encrypting the data included in it and the second pair of keys is used for signing/verifying the data signature. The reason for having two key pairs is to prevent attackers from modifying AB's sensitive data and signing it again with the private decryption key of the data owner.
- The AB sends a request to SSA asking it to record the AB's security information. The AB's identity data includes its name, decryption keys, and the trust level that a receiver must satisfy to use the AB. The goal is to keep the decryption keys and other auxiliary data for ABs in a trusted location. The decryption keys are given only to receivers that are eligible to access the AB.
- The AB computes a hash value for each version of the sensitive data and signs them using the signature keys. The signature certifies that sensitive data is from its owner.
- The AB encrypts sensitive data using the encryption key.



```
2011/12/13 23:07:47--.. :  SignatureAlgorithm:DSA
2011/12/13 23:07:47--.. :  SignedHash:{3=[B@148662, 2=[B@1829e6f, 1=[B@6fa9fc}
2011/12/13 23:07:47--CipherTools :  End Of Active Bundle
2011/12/13 23:07:47--CipherTools.CheckIntegrity :  activebundle.content: [B@8ceeea
2011/12/13 23:07:47--1: : Verification of Agent signature:true
2011/12/13 23:07:47--CipherTools.CheckIntegrity : Stored signature: [B@6fa9fc
2011/12/13 23:07:47--CipherTools.CheckIntegrity :  activebundle.content: [B@1687e7c
2011/12/13 23:07:47--2: : Verification of Agent signature:true
2011/12/13 23:07:47--CipherTools.CheckIntegrity : Stored signature: [B@1829e6f
2011/12/13 23:07:47--CipherTools.CheckIntegrity :  activebundle.content: [B@701a27
2011/12/13 23:07:47--3: : Verification of Agent signature:true
2011/12/13 23:07:47--CipherTools.CheckIntegrity : Stored signature: [B@148662
2011/12/13 23:07:47--ActiveBundleActivateBehavior : integrityCheck results: true
2011/12/13 23:07:47--ActiveBundleActivateBehavior :  Enforce Disclosure policies
2011/12/13 23:07:47--CipherTools.generateSecretKey : GenerateKey: AB
```

Fig. 5. AB processing logs GUI

Enabling an AB: After arriving at the destination, AB enables itself (Fig. 5 shows the logs for this step). The steps of the enabling algorithm are as follows:

- AB sends a request to SSA asking for the security information on AB and the receiver's trust level.
- AB checks if the receiver's trust level is lower than the minimal trust level required for AB access. If so, the AB apoptosizes; otherwise, it executes the next step.
- AB checks integrity of its sensitive data. It computes the hash value for sensitive data and it verifies the AB's signed hash value by comparing it to the computed hash value. If verification fails, AB apoptosizes; otherwise, the AB decrypts the appropriate version of data according to access policies and receiver authorization.
- AB enforces its privacy policies and provides the data to the receiver.
- AB sends audit information to ASA. This information includes AB's name, the receiver's identity, and the name of the event being audited.

5 Resilience of the Proposed Approach

The current AB approach relies on TTP. This brings in all TTP related issues such as loss of control, lack of trust etc [18]. We are investigating approaches to decrease reliance on TTP making ABs more self-protected entities. One such approach based on predicates over encrypted data and multiparty computing for Identity Management is discussed in [18]. Another approach (used in Vanish [19]) is to use Shamir's threshold secret sharing technique [21] to split decryption keys into N shares, and use a threshold t (< N) that defines the minimum number of shares required to reconstruct the key. The reconstructed key can then be used to decrypt the AB data. Key shares can be stored in a distributed hash table (DHT) system (such as Vuze), where each share is stored at a separate node. The main advantages of using DHT are decentralized and asynchronous communication and its large-scale geographic distribution making the practical attacks on key shares impossible. The malicious hosts can attack and alter the AB VM to gain unauthorized access to bundle's data. They can also deny the VM execution. In this case, data are not disclosed to unauthorized entities but legitimate entities are denied access to these data. The main challenge in implementing the AB's VM is assuring that a visited host executes the AB's VM code faithfully and correctly. We are investigating different approaches to address this. The idea is to intertwine the VM code and data together to make it incomprehensible and use obfuscation to hide data and program code within a scrambled code so that it still works, but provably cannot be reverse engineered [22].

6 Conclusion

PLM requires cross-company collaboration but sharing information externally raises numerous security concerns. Available solutions focus on data protection within the organization but do not address external information sharing. We need better technologies that provide security along with flexibility and adaptability to the PLM systems. The security mechanisms should not interfere with the existing information sharing and collaboration mechanisms. They should be able to protect shared information throughout its lifecycle. We propose the use of AB scheme that can be adopted in existing systems. AB security capabilities include protecting data throughout its lifecycle, self-protection on unknown/untrusted receivers, controlled data dissemination, selective data dissemination, activity monitoring, and dynamic policy adjustment. It enables organizations and their partners to share product information in a secure PLM environment with confidence that each participant's information is protected in accordance with its policies. The future work involves extending the prototype for AB scheme that doesn't rely on TTP, deploying it in a PLM system and performance evaluation of the system.

Acknowledgement. This research was supported by a grant from PLM Center of Excellence and CERIAS at Purdue University. Special thanks to Guneshi Wickramaarachchi and Lotfi Ben Othmane for their contributions to this research.

References

1. Ameri, F., Dutta, D.: Product lifecycle management: closing the knowledge loops. Computer-Aided Design & Applications 2(5), 577–590 (2005)
2. Atallah, M.J., Elmongui, H.G., Deshpande, V., Schwarz, L.B.: Secure supply-chain protocols. In: IEEE International Conference on E- Commerce, pp. 293–302 (2003)
3. Hackers attack Foxconn for the laughs (2012), `http://www.macworld.com/article/1165298/foxconn_reportedly_hacked_by_group_critical_of_working_conditions.html`
4. Verizon: 2012 Data Breach Investigations Report (2012), `http://www.verizonbusiness.com/resources/reports/rp_data-breach-investigations-report-2012_en_xg.pdf?CMP=DMC-SMB_Z_ZZ_ZZ_Z_TV_N_Z037`
5. Carr, J.: Strategies and Issues: Thwarting Insider Attacks. Network Magazine (2002)
6. Cappelli, D., Moore, A., Trzeciak, R., Shimeall, T.J.: Common Sense Guide to Prevention and Detection of Insider Threats, V. 3.1. Carnegie Mellon University (2009)
7. Woerner, J., Woern, H.: Distributed and secure co-operative engineering in virtual plant production. In: Advanced Production Management Systems: Conference on Collaborative Systems for Production Management, pp. 175–187 (2002)
8. Pels, H.: Federated Product Data Management in Multi-company Projects. Advances in Design, 281–291 (2006)
9. Leong, K.K., Yu, K.M., Lee, W.B.: A security model for distributed product data management system. Computers in Industry 50(2), 179–193 (2003)
10. Rouibah, K., Ould-Ali, S.: Dynamic data sharing and security in a collaborative product definition management system. Robotics and Computer-Integrated Manufacturing 23(2), 217–233 (2007)
11. Cera, C.D., Kim, T., Han, J., Regli, W.C.: Role-based viewing envelopes for information protection in collaborative modeling. Computer-Aided Design 36(9), 873–886 (2004)
12. Brustoloni, J.C., Nnaji, B.O.: Intellectual property protection in collaborative design through lean information modeling and sharing. Journal of Computing and Information Science in Engineering 6, 149 (2006)
13. Iyer, A.V., Ye, J.: Assessing the value of information sharing in a promotional retail environment. Manufacturing & Service Operations Management 2(2), 128–143 (2000)
14. Cachon, G.P., Fisher, M.: Supply chain inventory management and the value of shared information. Management Science 46(8), 1032–1048 (2000)
15. Browne, N., Crespigny, M., Reavis, J., Roemer, K., Samani, R.: Business Assurance for the 21st Century: Navigating the Information Assurance landscape. Information Security Forum (2011)
16. Ben Othmane, L., Lilien, L.: Protecting Privacy of Sensitive Data Dissemination Using Active Bundles. In: World Congress on Privacy, Security, Trust and the Management of e-Business, pp. 202–213 (2009)
17. Ben Othmane, L.: Active Bundles for Protecting Confidentiality of Sensitive Data Throughout Their Lifecycle. Ph.D. Thesis, Western Michigan University (2010)
18. Ranchal, R., Bhargava, B., Ben Othmane, L., Lilien, L., Kim, A., Kang, M., Linderman, M.: Protection of identity information in cloud computing without trusted third party. In: 29th IEEE Symposium on Reliable Distributed Systems (2010)
19. Geambasu, R., Kohno, T., Levy, A., Levy, H.M.: Vanish: Increasing data privacy with self-destructing data. In: 18th USENIX Security Symposium, p. 56 (2009)
20. Bhargava, B., Angin, P., Ranchal, R., Sivakumar, R., Linderman, M., Sinclair, A.: A trust based approach for secure data dissemination in a mobile peer-to-peer network of AVs. International Journal of Next-generation Computing 3(1) (2012)
21. Shamir, A.: How to Share a Secret. Communications of the ACM 22(11), 612–613 (1979)
22. Heffner, K., Collberg, C.: The Obfuscation Executive. Information Security 3225, 428–440 (2004)

Filtering Nonlinear Feedback Shift Registers Using Welch-Gong Transformations for Securing RFID Applications

Kalikinkar Mandal and Guang Gong

Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Ontario, N2L 3G1, Canada
{kmandal,ggong}@uwaterloo.ca

Abstract. Pseudorandom number generators play an important role to provide security and privacy on radio frequency identification (RFID) tags. In particular, the EPC Class 1 Generation 2 (EPC C1 Gen2) standard uses a pseudorandom number generator in the tag identification protocol. In this paper, we first present a pseudorandom number generator, named the filtering nonlinear feedback shift register using Welch-Gong (WG) transformations (filtering WG-NLFSR) and the filtering WG7-NLFSR for EPC C1 Gen2 RFID tags. We then investigate the periodicity of a sequence generated by the filtering WG-NLFSR by considering the model, named nonlinear feedback shift registers using Welch-Gong (WG) transformations (WG-NLFSR). The periodicity of WG-NLFSR sequences is investigated in two ways. Firstly, we perform the cycle decomposition of WG-NLFSR recurrence relations over different finite fields by computer simulations where the nonlinear recurrence relation is composed of a characteristic polynomial and a WG transformation module. Secondly, we conduct an empirical study on the period distribution of the sequences generated by the WG-NLFSR. The empirical study states that a sequence with period bounded below by the square root of the maximum period can be generated by the WG-NLFSR with high probability for any initial state.

Keywords: Nonlinear feedback shift registers, pseudorandom sequence generators, stream ciphers, WG-7 stream cipher.

1 Introduction

A pseudorandom sequence generator is a heart of a stream cipher, which is used for generating random-looking binary keystreams that are used to encrypt binary message streams by XORing the plaintexts with the keystreams in a bit by bit fashion to produce the ciphertexts. In practice, linear and nonlinear feedback shift registers (LFSRs/NLFSRs) have been widely used as basic building blocks for constructing stream ciphers. For instance, well-known stream ciphers, namely Grain, Trivium and Mickey in the eSTREAM project use NLFSRs as their building blocks [4].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 643–657, 2013.

The randomness properties of a sequence generated by an LFSR have been well studied and understood [5,6], however, the randomness properties of a sequence generated by an arbitrary NLFSR are not known and hard to determine. As an example, the cycle decomposition of an arbitrary NLFSR is not well understood and it is hard to determine the number of cycles and the lengths of the cycles in a cycle decomposition of an NLFSR. In the theory of NLFSRs, the cycle decomposition of NLFSRs is an important property to investigate first, since each cycle can be considered as a sequence and the cycles' lengths determine the periods of the sequences.

Several pseudorandom number generators have been proposed in the literature for EPC C1 Gen2 RFID tags [1,12,13,17]. Che *et al.*'s proposal [1] consists of an oscillator-based true random number generator (TRNG) and an LFSR of 16-stage where the TRNG is implemented using an analog circuit. In their design, one true random bit is added to each component of an LFSR generated 16-bit pseudorandom number. Due to the linear structure of the PRNG, the PRNG has been attacked by Melia-Segui *et al.* [13] with high success probability $\frac{(n+1)}{8n}$, wher n is the length of the LFSR. To avoid such an attack, Melia-Segui *et al.* [13] proposed a design by employing eight primitive polynomials to an LFSR where in each clock cycle one primitive polynomial is chosen based on a true random number generator. In [17], Peris-Lopez *et al.* proposed a PRNG named LAMED for RFID tags, which can generate 32-bit random numbers as well as 16-bit random numbers. The internal state of LAMED is 64-bit including a 32-bit key and a 32-bit IV. LAMED always outputs a 32-bit random number, a 16-bit number is obtained by dividing 32-bit number into two equal halves and XORing these two halves together. Recently, Mandal *et al.* [12] designed a PRNG named Warbler based on nonlinear feedback shift registers for RFID tags. In their design, three NLFSRs are used, two of them work over the binary field and the other one is defined over a finite field. The internal state of Warbler consists of 65 bits and 16-bit random numbers are produced by taking disjoint sequences of 16 bits.

In this paper, we present a family of pseudorandom sequence generators, named the filtering nonlinear feedback shift registers using Welch-Gong (WG) transformations (henceforth called filtering WG-NLFSR) for EPC Class 1 Generation 2 RFID tags. In particular, the filtering WG7-NLFSR is composed of a nonlinear feedback shift register of length 23 and a WG transformation module over the field $\mathbb{F}_{2^7}$. Due to the nonlinear state update of the filtering WG-NLFSR, the period of a sequence generated by the filtering WG-NLFSR is not known in general. We investigate the periodicity of a sequence generated by the filtering WG-NLFSR by considering the model, named nonlinear feedback shift registers using Welch-Gong (WG) transformations (WG-NLFSR). The design of the WG-NLFSR was inspired by the key initialization phase of the WG cipher, which was submitted to the eSTREAM project [4,15]. In the WG-NLFSR, the nonlinear recurrence relation is composed of a primitive polynomial and a nonlinear WG permutation. Due to the nonlinear property of the recurrence relation, the WG-NLFSR will be resistant to the powerful cryptanalytic attacks such as algebraic attacks, cube attacks, correlation attacks, and discrete fourier transformation attacks. Another objective of this paper is to study the periodicity of an output sequence produced by the WG-NLFSR. The periodicity of WG-NLFSR sequences is investigated in two steps. Firstly, we perform the complete cycle decomposition for different nonlinear recurrence relations by computer simulations. It is observed that, for a proper selection of a characteristic polynomial, a sequence with period greater than the square root of the maximum period can be generated by the WG-NLFSR. Secondly, we conduct an empirical study for investigating the period distribution of WG-NLFSR sequences. In the empirical study, we consider different WG-NLFSR recurrence relations over different finite fields and compute the probability distribution for different cases. Our empirical study shows that, with high probability, the WG-NLFSR generates sequences with periods bounded below by the square root of the maximum period.

The remainder of the paper is organized as follows. In Section 2, we define some terms and notations that will be used in the paper. In Section 3, we describe a general model of the filtering WG-NLFSR and a pseudorandom number generator, the filtering WG7-NLFSR. In Section 4, we study the periodicity of the WG-NLFSR sequences by performing the cycle decomposition of WG-NLFSR recurrence relations and by conducting an empirical study on the period distribution of WG-NLFSR sequences. Finally, in Section 5, we conclude the paper.

2 Preliminaries

In this section, we define the terms and notations that will be used in this paper to describe the filtering WG-NLFSR.

Notations:

- $\mathbb{F}_2 = \{0, 1\}$: the Galois field with 2 elements.
- $\mathbb{F}_{2^t}$: a finite field with 2^t elements, which is defined by α with $g(\alpha) = 0$, where $g(x)$ be a primitive polynomial of degree t over the field $\mathbb{F}_2$.
- $p(x) = c_0 + c_1x + \cdots + c_{n-1}x^{n-1} + x^n$: a characteristic polynomial over $\mathbb{F}_{2^t}$.
- $N = 2^{nt} - 1$: the maximum period of a nonzero sequence generated by an n-stage NLFSR over $\mathbb{F}_{2^t}$.
- $\mathbb{S} = \{(x_0, x_1, ..., x_{n-1}) \mid x_i \in \mathbb{F}_{2^t}\}$: the set of all states of the WG-NLFSR with $|\mathbb{S}| = N + 1$.

The Welch-Gong (WG) Transformation

Let $\mathrm{Tr}(x) = x + x^2 + x^{2^2} + \cdots + x^{2^{t-1}}$ be the trace function mapping from $\mathbb{F}_{2^t}$ to $\mathbb{F}_2$. Let t be a positive integer with $t \pmod 3 \neq 0$ and $3k \equiv 1 \bmod t$ for some integer k. We define the function h from $\mathbb{F}_{2^t}$ to $\mathbb{F}_{2^t}$ by $h(x) = x + x^{q_1} + x^{q_2} + x^{q_3} + x^{q_4}$ and the exponents are given by $q_1 = 2^k + 1, q_2 = 2^{2k} + 2^k + 1, q_3 = 2^{2k} - 2^k + 1, q_4 = 2^{2k} + 2^k - 1$. Then the function, from $\mathbb{F}_{2^t}$ to $\mathbb{F}_{2^t}$, defined as

$$\mathrm{WGP}(x) = h(x + 1) + 1$$

is known as the *WG permutation* and the function, from $\mathbb{F}_{2^t}$ to $\mathbb{F}_2$, defined by

$$WG(x) = \mathrm{Tr}(\mathrm{WGP}(x)), x \in \mathbb{F}_{2^t}$$

is known as the *WG transformation* [8]. The WG transformation has good cryptographic properties such as high nonlinearity, algebraic degree, and at least 1-order resiliency for a proper choice of basis. Moreover, a WG sequence has high linear complexity.

3 The Filtering WG-NLFSR

In this section we first give a general description of the filtering WG-NLFSR, which has two components including a characteristic polynomial and a WG transformation module. Then we present a pseudorandom number generator named the filtering WG7-NLFSR for EPC C1 Gen 2 RFID tags.

3.1 General Description of the Filtering WG-NLFSR

The filtering WG-NLFSR is a family of word-oriented pseudorandom sequence generators, where an internal state consists of n cells, each of which contains t bits. The total number of bits in an internal state of the filtering WG-NLFSR is $n \cdot t$. Moreover, the internal state is updated by a nonlinear recurrence relation, which is composed of a characteristic polynomial and a nonlinear WG permutation over $\mathbb{F}_{2^t}$. An overview of the architecture is shown in Fig. 1.

Let $\mathbf{a} = \{a_i\}_{i\geq 0}, a_i \in \mathbb{F}_{2^t}$ be a sequence generated by the n-stage nonlinear recurrence relation, which is defined as

$$a_{n+k} = c_0 a_k + c_1 a_{k+1} + \cdots + c_{n-1} a_{n-1+k} + \mathrm{WGP}(a_{n-1+k}),\ a_i \in \mathbb{F}_{2^t},\ k \geq 0,\ (1)$$

where $\mathrm{WGP}(x)$ is the WG permutation and $(a_0, a_1, ..., a_{n-1})$ is the *initial state.* The filtering WG-NLFSR sequence $\{b_i\}$ is defined by $b_i = WG(a_i)$, where $WG(x)$ is the WG transformation.

It is not hard to show that the period of $\{b_i\}$ produced by the filtering WG-NLFSR is the same as the period of $\mathbf{a}$. We note that the output sequence $\mathbf{a}$ cannot directly be used without applying the filter function because after n clock cycles one can have access to the internal state of the NLFSR, which allows an attacker to generate the whole sequence for a key.

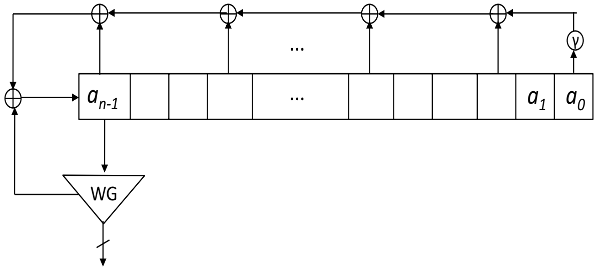


Fig. 1. Architecture of the Filtering WG-NLFSR

3.2 The Filtering WG7-NLFSR

We now give the mathematical details of the filtering WG7-NLFSR which is similar to the WG-7 stream cipher [11]. The main difference between the WG-7 stream cipher and the filtering WG7-NLFSR is that the WG-7 stream cipher uses the nonlinear feedback only at the initialization phase, but the filtering WG7-NLFSR always uses the nonlinear feedback function. The filtering WG7-NLFSR is composed of a nonlinear feedback shift register of length 23 and the WG transformation over the finite field $\mathbb{F}_{2^7}$. The finite field $\mathbb{F}_{2^7}$ is defined by the primitive polynomial $t(x) = x^7 + x + 1$ over $\mathbb{F}_2$.

Let $h(x) = x + x^{33} + x^{39} + x^{41} + x^{104}$. Then, the nonlinear WG permutation with decimation 3, from $\mathbb{F}_{2^7}$ to $\mathbb{F}_{2^7}$, is defined by $\mathsf{WGP7}(x^3) = h(x^3 + 1) + 1$, and the WG transformation over $\mathbb{F}_{2^7}$ is defined as

$$\mathsf{WG7}(x) = \mathrm{Tr}(\mathsf{WGP7}(x^3)) = \mathrm{Tr}(x^3 + x^9 + x^{21} + x^{57} + x^{87}), x \in \mathbb{F}_{2^7}$$

where $\mathrm{Tr}(x) = x + x^2 + x^4 + x^8 + x^{16} + x^{32} + x^{64}$ is the mapping from $\mathbb{F}_{2^7}$ to $\mathbb{F}_2$. We denote by $\{a_i\}$ the sequence generated by the NLFSR, which is defined as

$$a_{i+23} = \gamma a_i + a_{i+11} + \mathsf{WGP7}(a_{i+22}), a_i \in \mathbb{F}_{2^7} \tag{2}$$

where $p(x) = x^{23} + x^{11} + \gamma$ is a primitive polynomial over $\mathbb{F}_{2^7}$ and $t(\gamma) = 0$. A binary filtering WG7-NLFSR sequence $\{s_i\}$ is produced by filtering through the WG transformation $\mathsf{WG7}$, i.e., $s_i = \mathsf{WG7}(a_i), i \geq 0$.

The key length and the IV length of the filtering WG7-NLFSR are 80 bits and 81 bits, respectively. We represent an 80-bit key as $K_{0,1,\ldots,79}$ and an 81-bit initial vector as $IV_{0,1,\ldots,80}$. The key and an IV are loaded into the NLFSR as follows. For $0 \leq j \leq 10$, $a_{2j} = (K_{7j,7j+1,7j+2,7j+3}, IV_{7j,7j+1,7j+2})$ and $a_{2j+1} = (K_{7j+4,7j+5,7j+6}, IV_{7j+3,7j+4,7j+5,7j+6})$ and $a_{22} = (K_{77,78,79}, IV_{77,78,79,80})$. After loading the key and the IV, the filtering WG7-NLFSR is run for 46 clock cycles without any output. At 47-th clock cycle, the filtering WG7-NLFSR outputs the first bit.

Due to the nonlinear WG permutation $\mathsf{WGP7}(x)$ in recurrence relation (2), the period of the sequence $\{a_i\}$ is not known in general and is hard to know the exact cycle decomposition because of the large internal state. In Section 4, we will see a general investigation of the periodicity of a sequence produced by a nonlinear recurrence relation of the above type. As the keystream bits are generated by a purely nonlinear feedback function, it will be resistant to powerful cryptanalytic attacks such as algebraic attacks, correlation attacks, cube attacks and discrete fourier transformation attacks [2,7,14,16].

The mathematical functions used in the filtering WG7-NLFSR are the same as the functions used in the WG-7 stream cipher and the nonlinear WG permutation feedback does not increase any extra cost (as it is implemented for the key initialization), the implementation will be the same as the WG-7 stream cipher. For details of the WG-7 stream cipher implementation, we refer the reader to [11]. For easy reference, we reproduce the comparison data given in [11] in Table 8 as Appendix A, which indicates a microcontroller implementation comparison of the WG-7 stream cipher with other ciphers. The implementation includes the 4-bit MARC4 ATAM893-D microcontroller (*a* in Table 8) and the 8-bit AVR microcontroller ATmega8 (*b* in Table 8) from Atmel.

3.3 Application of the Filtering WG7-NLFSR

The EPCglobal Class 1 Generation 2 (EPC C1 Gen2) is an RFID standard. The tag identification protocol in the EPC C1 Gen2 standard uses a couple of 16-bit random numbers for identifying low cost passive RFID tags. Passive RFID tags get power from the reader at the beginning of the communication. Most of the existing random number generators are based on an LFSR and a true random number generator. Moreover, a true random number generator consumes more power, occupies more area and the throughput is low. For such resource-constrained environments, the filtering WG7-NLFSR can be used as a pseudorandom number generator for generating 16-bit random numbers. The 16-bit random numbers are generated by taking disjoint 16-bit sequences from the filtering WG7-NLFSR sequence $\{s_i\}$. Based on the implementation given in [11,10], it is confirmed that the filtering WG7-NLFSR is a suitable candidate for RFID tags.

4 Period Analysis of the WG-NLFSR

In order to study the periodicity of a filtering WG-NLFSR sequence, we need to investigate the period property of a sequence produced by recurrence relation (1). We redefine the nonlinear recurrence relation for the WG-NLFSR over the field $\mathbb{F}_{2^t}$ as follows. Let $\mathbf{a} = \{a_i\}_{i\geq 0}, a_i \in \mathbb{F}_{2^t}$ be a sequence generated by an n-stage nonlinear recurrence relation, which is defined as

$$a_{n+k} = c_0a_k + c_1a_{k+1} + \cdots + c_{n-1}a_{n-1+k} + \mathrm{WGP}(a_{n-1+k}),\ a_i \in \mathbb{F}_{2^t},\ k \geq 0, \quad (3)$$

where $\mathrm{WGP}(x)$ is the WG permutation, $t \pmod 3 \neq 0$, and $(a_0, a_1, ..., a_{n-1})$ is the *initial state*. We call the nonlinear recurrence relation (3) a *WG-NLFSR recurrence relation*. A block diagram of the WG-NLFSR sequence generator is shown in Fig. 2. Note that a WG-NLFSR recurrence relation is uniquely determined by the characteristic polynomial $p(x)$ and WG permutation. For a fixed WG permutation, the recurrence relation is different if the characteristic polynomial is different.

Due to the nonlinear term $\mathrm{WGP}(\cdot)$ in the recurrence relation (3), the period of the sequence $\mathbf{a}$ is not equal to the period of the polynomial $p(x)$. In particular, the period of $\mathbf{a}$ depends on three factors: the characteristic polynomial $p(x)$, the WG permutation $\mathrm{WGP}(x)$, and the initial state. To investigate the period of sequence $\mathbf{a}$, we need to study the cycle decomposition of the recurrence relation.

Remark 1. In recurrence relation (3), any permutation over a finite field $\mathbb{F}_{2^t}$ can be used. We here used WG permutation as a WG transformation has excellent cryptographic properties and which can be used for both updating the internal state and filtering the output sequences.

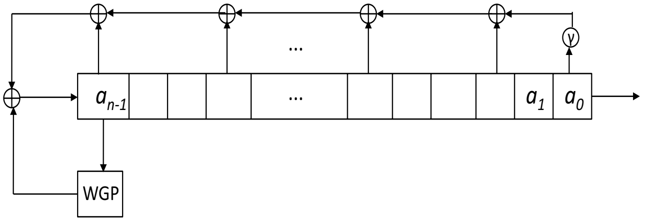


Fig. 2. Architecture of the WG-NLFSR

4.1 Cycle Decomposition of the WG-NLFSR

It is not hard to show that the recurrence relation (3) generates sequences with no branch. Thus, the recurrence relation partitions the whole state space $\mathbb{S}$ into a finite number of disjoint cycles, which is known as the cycle decomposition of the recurrence relation [5]. We denote by Ω the cycle decomposition of the recurrence relation (3), where $\Omega = \{C_1, C_2, \cdots, C_r\}$ with $\mathbb{S} = C_1 \cup C_2 \cup \cdots \cup C_r$ and $C_i \cap C_j = \phi$, $1 \leq i \neq j \leq r$. For an arbitrary recurrence relation, the value of r is not determined. Let $L_i = |C_i|$ be the number of states in $C_i, i = 1, 2, ..., r$. Using any state of C_i, all other states in C_i can be generated by recurrence relation (3). Thus, C_i can be considered as a sequence with period L_i. (For details of cycle decompositions, see [5].)

We here perform computer simulations for investigating the cycle structure of recurrence relation (3). Considering recurrence relation (3) over fields $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$, we present the cycle decompositions for different characteristic polynomials in Tables 1 - 4, where Y represents YES and N represents NO. In tables, the primitive elements α, β and the WG transformations over fields $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$ are defined in Section 4.2. The computer simulations show that for a fixed WGP(x) and a proper selection of a characteristic polynomial, a sequence with period lower bounded by $\sqrt{N}$ can be generated by the recurrence relation (3), where a proper selection of a characteristic polynomial is meant by a characteristic polynomial in the recurrence relation (3) for which the lengths of all cycles are greater than or equal to $\sqrt{N}$. It is noticed that the long period of a sequence generated by recurrence relation (3) does not depend on the irreducibility of the characteristic polynomial. In the recurrence relation, there exists a hidden relation between the coefficients of a characteristic polynomial and the exponents of the WG permutation and that hidden relation can determine a construction of a nonlinear feedback function, which will generate a sequence with a bounded period. Unfortunately, we are not yet able to explore the hidden relation.

4.2 Period Distribution of the WG-NLFSR

In this section, we conduct an empirical study on the period distribution of the sequences generated by recurrence relation (3) by considering the recurrence relation for different characteristic polynomials. In the cycle decomposition, we

Table 1. Complete cycle decompositions of the WG-NLFSR for $n = 3$ over $\mathbb{F}_{2^5}$

Index	Characteristic Polynomial	Irreducible	Cycle decomposition, Ω
1	$1 + \alpha^{14}x + \alpha^{21}x^2 + x^3$	N	23779, 6710, 2276, 1
2	$\alpha^4 + \alpha^{17}x + \alpha^{19}x^2 + x^3$	N	32762, 1, 4
3	$\alpha^{20} + \alpha^2x + \alpha^{25}x^2 + x^3$	Y	15236, 14762, 2769
4	$1 + \alpha^7x + \alpha^{26}x^2 + x^3$	N	23779, 6710, 2276, 1
5	$\alpha^3 + \alpha^{26}x + \alpha^9x^2 + x^3$	N	32750, 4, 2, 3, 1
6	$\alpha^5 + \alpha^{20}x + \alpha^{15}x^2 + x^3$	Y	32754, 4, 3, 5, 1
7	$\alpha^7 + \alpha^{16}x + \alpha^{18}x^2 + x^3$	Y	32762, 4, 1

Table 2. Complete cycle decompositions of the WG-NLFSR for $n = 4$ over $\mathbb{F}_{2^5}$

Index	Characteristic Polynomial	Irreducible	Cycle decomposition, Ω
1	$\alpha + \alpha^4x + \alpha^{26}x^2 + \alpha^{25}x^3 + x^4$	N	39070, 363841, 546171, 99492, 1
2	$\alpha + \alpha^7x + \alpha x^2 + \alpha x^3 + x^4$	N	590707, 379331, 46734, 22986, 8815, 1
3	$\alpha + \alpha^8x + \alpha^2x^2 + \alpha^4x^3 + x^4$	N	615325, 114129, 91408, 227712, 1
4	$\alpha + \alpha^8x + \alpha^{30}x^2 + \alpha^{10}x^3 + x^4$	N	298643, 549045, 141353, 59533, 1
5	$\alpha + \alpha^{18}x + \alpha^4x^2 + \alpha^{22}x^3 + x^4$	N	664966, 54862, 268380, 34846, 25518, 2, 1
6	$\alpha + \alpha^{25}x + \alpha^7x^2 + \alpha^{21}x^3 + x^4$	N	430236, 609318, 3194, 5826, 1
7	$\alpha + \alpha^{28}x + \alpha^5x^2 + \alpha^{25}x^3 + x^4$	N	914718, 91230, 42623, 1, 3
8	$\alpha^3 + \alpha^8x^2 + \alpha^{25}x^3 + x^4$	Y	463471, 585093, 5, 1
9	$\alpha^3 + \alpha^8x + \alpha^{10}x^2 + \alpha^{25}x^3 + x^4$	N	490152, 522883, 30947, 4592, 1
10	$\alpha^3 + \alpha^{20}x + \alpha^{10}x^2 + \alpha^{26}x^3 + x^4$	Y	178444, 870118, 1, 4, 3
11	$\alpha^3 + \alpha^{20}x + \alpha^{13}x^2 + \alpha^{18}x^3 + x^4$	N	636187, 81945, 312587, 17855, 1
12	$\alpha^3 + \alpha^{25}x + \alpha^{18}x^2 + \alpha^{17}x^3 + x^4$	N	62628, 105531, 880413, 2, 1
13	$\alpha^3 + \alpha^{25}x + \alpha^{20}x^2 + \alpha^{22}x^3 + x^4$	N	1048562, 7, 6
*14	$\alpha^5 + \alpha^{14}x + \alpha^{12}x^3 + x^4$	N	1030097, 9736, 8742
15	$\alpha^5 + \alpha^{16}x + \alpha^{14}x^2 + \alpha^3x^3 + x^4$	N	981057, 53724, 13788, 2, 4
16	$\alpha^7 + \alpha^{10}x + \alpha^2x^2 + \alpha^{21}x^3 + x^4$	N	1048570, 4, 1
17	$\alpha^7 + \alpha^{17}x + \alpha^{14}x^2 + \alpha^{15}x^3 + x^4$	N	953457, 80759, 14347, 7, 2, 1
18	$\alpha^7 + \alpha^{19}x + \alpha^{15}x^2 + \alpha^{24}x^3 + x^4$	N	1048572, 2, 1
19	$\alpha^{11} + \alpha x + \alpha^8x^2 + \alpha^8x^3 + x^4$	N	940556, 108007, 9, 1, 2
20	$\alpha^{11} + \alpha^7x + \alpha^4x^2 + \alpha^{28}x^3 + x^4$	N	125158, 635317, 249323, 38772, 1, 3
21	$\alpha^{11} + \alpha^{10}x + \alpha^{26}x^2 + \alpha^{11}x^3 + x^4$	N	554609, 493933, 16, 1
22	$\alpha^{11} + \alpha^{11}x + \alpha^2x^2 + \alpha^{14}x^3 + x^4$	N	696972, 337871, 13730, 1
23	$\alpha^{11} + \alpha^{14}x + \alpha^{18}x^2 + \alpha^8x^3 + x^4$	N	240673, 726854, 81046, 1
24	$\alpha^{11} + \alpha^{15}x + \alpha^3x^2 + \alpha^{12}x^3 + x^4$	Y	1005347, 43222, 3, 1, 2
25	$\alpha^{11} + \alpha^{15}x + \alpha^{20}x^2 + \alpha^9x^3 + x^4$	N	835608, 212956, 9, 1
26	$\alpha^{11} + \alpha^{18}x + \alpha^7x^2 + \alpha^{23}x^3 + x^4$	N	895975, 152596, 2, 1
27	$\alpha^{11} + \alpha^{20}x + \alpha^{21}x^2 + \alpha^{28}x^3 + x^4$	Y	289429, 510434, 84330, 164381, 1
28	$\alpha^{11} + \alpha^{27}x + \alpha^{23}x^2 + \alpha^8x^3 + x^4$	Y	835558, 213010, 2, 4, 1
29	$\alpha^{15} + x + \alpha^{14}x^2 + x^4$	N	1008690, 39884, 1
30	$\alpha^{15} + \alpha^8x + \alpha^{14}x^2 + \alpha^8x^3 + x^4$	N	881607, 166967, 1
31	$\alpha^{15} + \alpha^{15}x + \alpha^8x^2 + \alpha^{20}x^3 + x^4$	N	675115, 373449, 2, 3, 1
32	$\alpha^{15} + \alpha^{16}x + \alpha^{11}x^2 + \alpha^{13}x^3 + x^4$	N	922952, 57138, 44338, 24136, 6, 4, 1
33	$\alpha^{15} + \alpha^{24}x + \alpha^{15}x^2 + \alpha^{25}x^3 + x^4$	Y	1048571, 3, 1

have observed that there exist many characteristic polynomials for which the recurrence relation can generate sequences with periods bounded below by $\sqrt{N}$, where N is the maximum period. However, we do not know the relation between the WG permutation and such characteristic polynomials in general. We here intend to study the probability distribution of period of at least $\sqrt{N}$. That is,

Table 3. Complete cycle decompositions of the WG-NLFSR for $n = 5$ over $\mathbb{F}_{2^5}$

Index	Characteristic Polynomial	Irreducible	Cycle decomposition, Ω
1	$\alpha + \alpha^{18}x^2 + \alpha^{10}x^3 + \alpha^{14}x^4 + x^5$	N	24934939, 8057211, 501740, 60539, 2
2	$\alpha + \alpha^{21}x^2 + \alpha^{26}x^3 + \alpha^{20}x^4 + x^5$	N	33324081, 215923, 14354, 6, 67
3	$\alpha + \alpha x + \alpha^5x^2 + \alpha^{21}x^3 + \alpha^5x^4 + x^5$ N		23683815, 9430226, 180678, 255311, 4401
4	$\alpha + \alpha^{28}x^2 + \alpha^{19}x^3 + \alpha^{19}x^4 + x^5$	Y	33137436, 416935, 29, 23, 1, 6
5	$\alpha + x + \alpha x^2 + \alpha^{22}x^3 + \alpha^9x^4 + x^5$	N	33509677, 42891, 1740, 118, 2, 1
6	$\alpha + \alpha x + x^2 + \alpha^5x^3 + \alpha^{20}x^4 + x^5$	N	32438885, 802371, 136113, 154148, 22912, 1
7	$\alpha + \alpha^4x^2 + \alpha^8x^3 + \alpha^{18}x^4 + x^5$	N	20018544, 12576215, 661370, 252630, 45560, 111, 1
8	$\alpha + \alpha^5x^2 + \alpha^{24}x^3 + x^4 + x^5$	N	31853496, 1026340, 616630, 10591, 46360, 1001, 13
9	$\alpha + \alpha^6x^2 + \alpha^{28}x^3 + \alpha^4x^4 + x^5$	N	27060025, 539828, 5044304, 853141, 57062, 70, 1
10	$\alpha + \alpha^{13}x^2 + \alpha^{14}x^3 + \alpha^4x^4 + x^5$	N	1614083, 26744592, 5172342, 23352, 59, 2, 1
11	$\alpha + \alpha^{16}x^2 + \alpha^2x^3 + \alpha^{24}x^4 + x^5$	N	26604921, 60903, 5881770, 980844, 25982, 4, 7
12	$\alpha + \alpha^{18}x^2 + \alpha^{20}x^3 + \alpha^{24}x^4 + x^5$	N	13669238, 17126821, 2416848, 289074, 52395, 54, 1
13	$\alpha + x + \alpha^{11}x^3 + \alpha^{15}x^4 + x^5$	Y	29770970, 2699894, 1000613, 62602, 20324, 23, 5
14	$\alpha + x + x^2 + \alpha^5x^3 + \alpha^{18}x^4 + x^5$	N	9244135, 9425167, 10061666, 4589985, 233472, 1, 5
15	$\alpha + x + x^2 + \alpha^{22}x^3 + \alpha^{13}x^4 + x^5$	Y	32786392, 758058, 9835, 132, 11, 2, 1
16	$\alpha + x + \alpha x^2 + \alpha^{16}x^3 + \alpha^{20}x^4 + x^5$	N	33188710, 351685, 13861, 166, 6, 2, 1
17	$\alpha + x + \alpha^4x^2 + \alpha^{28}x^3 + \alpha^{18}x^4 + x^5$	Y	33554268, 45, 17, 2, 1, 29, 3
*18	$\alpha + x + \alpha^{11}x^2 + \alpha^{25}x^3 + \alpha^{19}x^4 + x^5$	N	1711633, 17174871, 11626420, 2069636, 659633, 275686, 36552
19	$\alpha + x + \alpha^{12}x^2 + \alpha^{30}x^3 + \alpha^{20}x^4 + x^5$	N	26385451, 704023, 262540, 3728330, 2474077, 8, 2
20	$\alpha + x + \alpha^{13}x^2 + \alpha x^3 + \alpha^{17}x^4 + x^5$	N	31083249, 2470874, 281, 11, 6, 9, 1
21	$\alpha + x + \alpha^{16}x^2 + \alpha^{20}x^3 + \alpha^{30}x^4 + x^5$	N	32645326, 634069, 54804, 88483, 74357, 57391, 1
22	$\alpha + x + \alpha^{19}x^2 + \alpha^{27}x^3 + \alpha^{12}x^4 + x^5$	N	30290671, 609570, 384964, 554062, 1570249, 144914, 1
*23	$\alpha + x + \alpha^{25}x^2 + \alpha x^3 + \alpha^{27}x^4 + x^5$	N	6758906, 19951473, 853356, 5840681, 5929, 75633, 68453
24	$\alpha + \alpha^8x^2 + \alpha^{21}x^3 + \alpha^{13}x^4 + x^5$	N	31959770, 1594335, 112, 173, 7, 17, 9, 1
25	$\alpha + x + \alpha^2x^2 + \alpha^{12}x^3 + \alpha^{20}x^4 + x^5$	N	14631594, 17557700, 1270630, 23428, 50395, 20669, 11, 2
26	$\alpha + x + \alpha^3x^2 + \alpha^{10}x^3 + \alpha^{10}x^4 + x^5$	N	8613690, 17190010, 7681297, 17715, 34521, 17155, 41, 1
27	$\alpha + x + \alpha^{17}x^2 + \alpha^{10}x^3 + \alpha^9x^4 + x^5$	N	31934521, 1487357, 11327, 64353, 56840, 28, 3, 1
28	$\alpha + x + \alpha^{21}x^2 + \alpha^{16}x^3 + \alpha^{29}x^4 + x^5$	Y	11545515, 21015426, 720059, 240858, 32564, 3, 2, 1
29	$\alpha + \alpha x + \alpha^3x^2 + x^3 + \alpha^{12}x^4 + x^5$	N	20341385, 6807881, 4023518, 1776187, 598917, 6539, 2, 1

we want to compute what the success probability is that for any initial state of the recurrence relation, the WG-NLFSR can generate a sequence with period lower bounded by $\sqrt{N}$. The main goal of performing this empirical study is that it can convey a general behavior of this type of recurrence relations.

Procedure for Computing the Success Probability for the Period $\geq \sqrt{N}$. We calculate the probability distribution of period as follows. For a WG-NLFSR recurrence relation, we perform the complete cycle decomposition by computer simulations. We first compute the complete cycle decompositions for different characteristic polynomials with the same WG permutation, where different characteristic polynomials are chosen randomly. Then, using the cycle decomposition we calculate the expected success probability and the standard

Table 4. Complete cycle decompositions of the WG-NLFSR for $n = 3$ over $\mathbb{F}_{2^7}$

Index	Characteristic Polynomial	Irreducible	Cycle decomposition, Ω
1	$\beta + \beta x + \beta^{116}x^2 + x^3$	Y	1972915, 124227, 9
2	$\beta + \beta^4 x + \beta^2 x^2 + x^3$	N	281885, 213421, 858081, 306286, 24446, 327239, 11564, 58233, 15994, 1
3	$\beta + \beta^4 x + \beta^{111}x^2 + x^3$	N	1862053, 21922, 161976, 38595, 12601, 2
4	$\beta + \beta^7 x + \beta^{43}x^2 + x^3$	Y	1548601, 335992, 200230, 12315, 3, 1
*5	$\beta + \beta^{21}x + \beta^{121}x^2 + x^3$	Y	1482387, 331576, 283188
6	$\beta + \beta^{55}x + \beta^{45}x^2 + x^3$	N	2079604, 17535, 5, 3, 4
7	$\beta + \beta^{80}x + \beta^{84}x^2 + x^3$	Y	2097095, 52, 2, 1
8	$\beta + \beta^{81}x + \beta^8 x^2 + x^3$	Y	245680,143280,675851,1003363, 20428,8546,1
9	$\beta + \beta^{91}x + \beta^7 x^2 + x^3$	N	1980490, 75492, 41167, 1
10	$\beta^3 + \beta^2 x + x^3$	Y	1923727, 173414, 7, 2, 1
*11	$\beta^3 + \beta^4 x + \beta^{83}x^2 + x^3$	N	2043475, 38142, 15534
12	$\beta^3 + \beta^{54}x + \beta^{84}x^2 + x^3$	Y	1892847, 184935, 19367, 1
13	$\beta^3 + \beta^{87}x + \beta^{38}x^2 + x^3$	N	2082246, 14900, 3, 1
14	$\beta^9 + \beta^{69}x + \beta^{69}x^2 + x^3$	N	1956446, 140682, 16, 6, 1
15	$\beta^9 + \beta^{70}x + \beta^{21}x^2 + x^3$	Y	1918311, 174964, 3872, 3, 1
16	$\beta^9 + \beta^{101}x + \beta^{84}x^2 + x^3$	Y	1955962, 141168, 14, 4, 3
17	$\beta^9 + \beta^{115}x + \beta^{29}x^2 + x^3$	Y	1610286, 486846, 16, 2, 1
18	$\beta^5 + \beta^{78}x + \beta^{118}x^2 + x^3$	N	1780061, 274339, 42749, 1
19	$\beta^9 + \beta^{20}x + \beta^{121}x^2 + x^3$	N	678904, 1418237, 4, 3
20	$\beta^{11} + \beta^{30}x + \beta^4 x^2 + x^3$	Y	624809, 1446046, 26294, 1
21	$\beta^{21} + \beta^{99}x + \beta^{59}x^2 + x^3$	N	2038686, 58448, 9, 4
22	$\beta + \beta^{25}x + \beta^{81}x^2 + x^3$	N	191464, 1328016, 460109, 117558, 4
23	$\beta^3 + \beta^{17}x + \beta x^2 + x^3$	Y	1576062, 356525, 140941, 23621, 2
*24	$\beta^3 + \beta^{112}x + \beta^{44}x^2 + x^3$	Y	93674, 1203620, 395834, 392354, 11669
25	$\beta^7 + \beta^{46}x + \beta^{84}x^2 + x^3$	N	1023858, 706836, 334068, 32387, 2
*26	$\beta^{11} + \beta^{53}x + \beta^{13}x^2 + x^3$	N	162697, 1628279, 72007, 114484, 119684
*27	$\beta^{27} + \beta^{28}x + \beta^{90}x^2 + x^3$	N	1393588, 534559, 116786, 34123, 18095
28	$\beta^{27} + \beta^{48}x + \beta^{91}x^2 + x^3$	Y	658722, 1230400, 176058, 31965, 6
29	$\beta + \beta^4 x + \beta^{111}x^2 + x^3$	N	1862053, 21922, 161976, 38595, 12601, 2
*30	$\beta^{23} + \beta^{62}x + \beta^{46}x^2 + x^3$	N	450219, 149546, 530547, 287938, 648238, 13859, 16804
*31	$\beta^{21} + \beta^5 x + \beta^{75}x^2 + x^3$	Y	668870, 643111, 86400, 73493, 343991, 277419, 3867
*32	$\beta^{19} + \beta^{118}x + \beta^{15}x^2 + x^3$	N	283412, 1296087, 431294, 23925, 25440, 24900, 12093

deviation (SD) of the period greater than or equal to $\sqrt{N}$. We note that the success probability is equal to one when the lengths of all the cycles are greater than or equal to $\sqrt{N}$. The details of the success probability calculation is described in the following procedure.

Let D be a random variable which represents the number of distinct characteristic polynomials of the same degree. For each characteristic polynomial, the success probability of the period greater than or equal to $\sqrt{N}$ is computed as follows:

Procedure 1.

1. Compute $\{C_1, C_2, ..., C_r\}$, which is the cycle decomposition of the characteristic polynomial with $L_i = |C_i|,\ i = 1, 2, ..., r$.
2. Add all L_j's which are less than $\sqrt{N}$ and let the sum be L_{sum}.
3. The success probability of the period bounded below by $\sqrt{N}$ for any initial state is $1 - \frac{L_{sum}}{N}$.

We then compute the expectation and standard deviation (SD) for the period of D success probabilities. Let D_{mean} and D_{SD} be the expectation and standard deviation, respectively. Then, we use the histogram with (D_{mean}, D_{SD}) to represent the probability distribution of the period. In the following subsection, we present the experimental results by the above procedure.

Period Distribution of the WG-NLFSR over the Field $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$. In this subsection, we compute the expected success probability of period by the above Procedure 1 for the recurrence relation of length $n = 3, 4$ and 5 over the field $\mathbb{F}_{2^5}$ and for the recurrence relation of length $n = 3, 4$ over the field $\mathbb{F}_{2^7}$. In Table 5, the WG permutations over fields $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$ are defined.

Table 5. Parameter descriptions

t	Primitive element of $\mathbb{F}_{2^t}$	Primitive Polynomial	WG permutations
5	α	$\alpha^5 + \alpha^3 + 1 = 0$	$\text{WGP5}(x) = x + (x+1)^5 + (x+1)^{13} + (x+1)^{19} + (x+1)^{21}$
7	β	$\beta^7 + \beta + 1 = 0$	$\text{WGP7}(x) = x + (x+1)^{33} + (x+1)^{39} + (x+1)^{41} + (x+1)^{104}$

We consider the n-stage recurrence relation (3) with WGP5(x) and WGP7(x) as the WG permutation over the field $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$, respectively. Our simulation results for $n = 3, 4$ and 5 over the field $\mathbb{F}_{2^5}$ are given in Table 6. Similarly, the simulation results for $n = 3$ and 4 over $\mathbb{F}_{2^7}$ are given in Table 7. In Tables 6 and 7, we provide the number of characteristic polynomials (D), the expected success probability (D_{mean}), the standard deviation (D_{SD}), and the maximum value of the sum of all smaller length cycles which are less than $\sqrt{N}$ (L_{sum}). In addition, the average number of cycles in the cycle decomposition of the WG-NLFSR recurrence relation is presented. Our experimental results show that the numerical value for the average number of cycles is very close to the average number of cycles generated by the random sampling (let $\overline{r}_s$ denote the expected number of cycles generated by the random sampling, then $\overline{r}_s \approx \ln N$, see [5]).

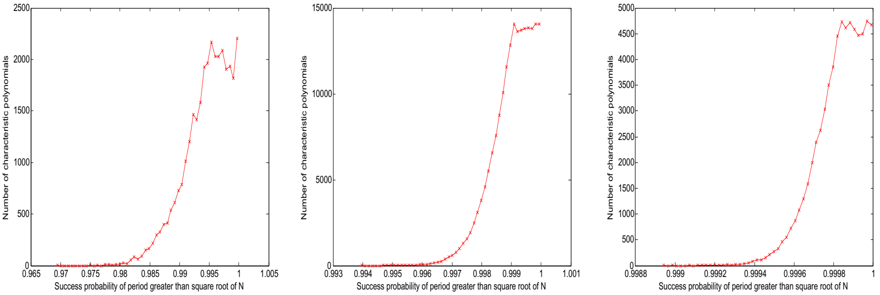
Table 6. The summary of simulation results over $\mathbb{F}_{2^5}$

Length, n	Max period, N	D	D_{mean}	D_{SD}	L_{sum}	Avg.#of cycles	$\overline{r}_s$
3	$2^{15} - 1$	31744	0.9945	0.0039	1011	10.51	10.38
4	$2^{20} - 1$	197296	0.9990	0.00069	6394	13.95	13.86
5	$2^{25} - 1$	66888	0.9998	0.00012	35828	17.44	17.32

For $n = 3$, the success probability of period lower bounded by $\sqrt{N}$ is depicted in Fig. 3a in the form of a histogram. In figures, the x-axis represents the success probability values and the y-axis represents the number of characteristic polynomials that have been taken. In the histogram, it can be observed that for most characteristic polynomials the recurrence relation produces sequences with period of at least $\sqrt{N}$ when the success probability is greater than 0.985. The empirical result for $n = 3$ in Table 6 says that if an arbitrary characteristic polynomial is chosen in the WG-NLFSR recurrence relation with WGP5(x),

Table 7. The summary of simulation results over $\mathbb{F}_{2^7}$

Length, n	Max period, N	D	D_{mean}	SD	L_{sum}	Avg.#of cycle	$\overline{r}_s$
3	$2^{21}-1$	294912	0.9993	0.00049	35828	14.49	14.55
4	$2^{28}-1$	7337	0.9999	0.00004	82216	19.38	19.41



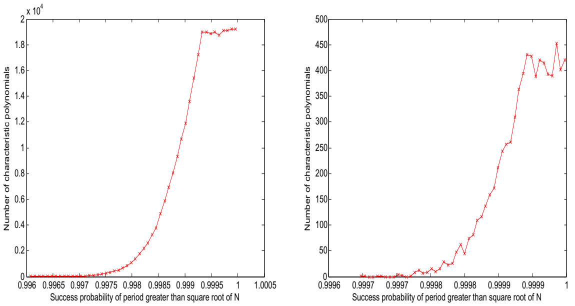
(a) $n = 3$ (b) $n = 4$ (c) $n = 5$

Fig. 3. Distribution of the period $\geq \sqrt{N}$ for $t = 5$, (a) $n = 3$, (b) $n = 4$ and (c) $n = 5$

then, with expected probability 0.9945, the recurrence relation can generate a sequence with period lower bounded by $\sqrt{N}$.

In a similar fashion, the probability distributions of period for $n = 4$ and 5 over $\mathbb{F}_{2^5}$ in Figs. 3b and 3c, and $n = 3$ and 4 over $\mathbb{F}_{2^7}$ in Figs. 4a and 4b are depicted in the form of a histogram along with the expected success probability. For $n = 4$ and 5, the expected success probabilities of the period are given by 0.990 and 0.9998, respectively, which are greater than the expected success probability for $n = 3$.

The empirical analysis shows that with a high probability the WG-NLFSR can generate a sequence with period at least $\sqrt{N}$ for a large length of the NLFSR. In



(a) $n = 3$ (b) $n = 4$

Fig. 4. Distribution of the period $\geq \sqrt{N}$ for $t = 7$, (a) $n = 3$, (b) $n = 4$

particular, with very high probability, the filtering WG7-NLFSR can generate a sequence with period at least $2^{80.5}$.

5 Conclusions

In this paper, we presented a family of pseudorandom number generators named the filtering WG-NLFSR and the filtering WG7-NLFSR for EPC C1 Gen2 RFID tags. Due to the nonlinear feedback for the state update, the filtering WG-NLFSR and filtering WG7-NLFSR will be resistant to the powerful cryptanalytic attacks. In order to investigate the periodicity of the filtering WG7-NLFSR sequence, we introduced the WG-NLFSR, which generates sequences over the finite field. The periodicity of WG-NLFSR sequences is investigated by performing the complete cycle decomposition of the WG-NLFSR recurrence relations and by conducting an empirical study on the period distribution of WG-NLFSR sequences. In the cycle decomposition, we observed that there are many characteristic polynomials in which the cycle lengths are close to the maximum period or bounded below by $\sqrt{N}$ and we listed some characteristic polynomials over the fields $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$. In the empirical study, the period distribution of the WG-NLFSR sequences over the field $\mathbb{F}_{2^5}$ and $\mathbb{F}_{2^7}$ for different lengths of the shift registers are conducted. Moreover, the empirical study reveals that, with high probability, the filtering WG7-NLFSR can generate sequences with periods bounded below by $2^{80.5}$. To the best of our knowledge, this is the first study in the literature on the cycle decomposition and the distribution of a period of a sequence generated by the nonlinear feedback shift register over an extension field.

References

1. Che, W., Deng, H., Tan, X., Wang, J.: A Random Number Generator for Application in RFID Tags. In: Networked RFID Systems and Lightweight Cryptography, ch. 16, pp. 279–287. Springer (2008)
2. Dinur, I., Shamir, A.: Cube Attacks on Tweakable Black Box Polynomials. In: Joux, A. (ed.) EUROCRYPT 2009. LNCS, vol. 5479, pp. 278–299. Springer, Heidelberg (2009)
3. EPCglobal. EPC Radio-Frequency Identification Protocol Class-1 Generation-2 UHF RFID for Communication at 860-960 MHz (2008), `http://www.epcglobalinc.org/`
4. The eStream Project, `http://www.ecrypt.eu.org/stream/`
5. Golomb, S.W.: Shift Register Sequences. Aegean Park Press, Laguna Hills (1981)
6. Golomb, S.W., Gong, G.: Signal Design for Good Correlation: For Wireless Communication, Cryptography, and Radar. Cambridge University Press, New York (2004)
7. Gong, G., Rønjom, S., Helleseth, T., Hu, H.: Fast Discrete Fourier Spectra Attacks on Stream Ciphers. IEEE Transactions on Information Theory 57(8), 5555–5565 (2011)

8. Gong, G., Youssef, A.: Cryptographic Properties of the Welch-Gong Transformation Sequence Generators. IEEE Transactions on Information Theory 48(11), 2837–2846 (2002)
9. Juels, A.: RFID Security and Privacy: A Research Survey. IEEE Journal on Selected Areas in Communications (J-SAC) 24(2), 381–394 (2006)
10. Lam, C., Aagaard, M., Gong, G.: Hardware Implementations of Multi-output Welch-Gong Ciphers, CACR Technical Report (2011), http://www.cacr.math.uwaterloo.ca/
11. Luo, Y., Chai, Q., Gong, G., Lai, X.: WG-7: A Lightweight Stream Cipher with Good Cryptographic Properties. In: IEEE Global Communications Conference – GLOBECOM 2010, pp. 1–6 (2010)
12. Mandal, K., Fan, X., Gong, G.: Warbler: A Lightweight Pseudorandom Number Generator for EPC Class 1 Gen 2 RFID Tags. In: Radio Frequency Identification System Security: RFIDsec 2011 Asia Workshop Proceedings (Cryptology and Information Security), November 7-8 (2012)
13. Melia-Segui, J., Garcia-Alfaro, J., Herrera-Joancomarti, J.: Analysis and Improvement of a Pseudorandom Number Generator for EPC Gen2 Tags. In: Sion, R. (ed.) FC 2010. LNCS, vol. 6052, pp. 34–46. Springer, Heidelberg (2010)
14. Meier, W., Staffelbach, O.: Fast Correlation Attacks on Certain Stream Ciphers. Journal of Cryptology, 159–176 (1989)
15. Nawaz, Y., Gong, G.: WG: A Family of Stream Ciphers with Designed Randomness Properties. Information Science 178(7), 1903–1916 (2008)
16. Courtois, N., Meier, W.: Algebraic Attacks on Stream Ciphers with Linear Feedback. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 345–359. Springer, Heidelberg (2003)
17. Peris-Lopez, P., Hernandez-Castro, J.C., Estevez-Tapiador, J.M., Ribagorda, A.: LAMED - A PRNG for EPC Class-1 Generation-2 RFID Specification. Computer Standard Interfaces 31, 88–97 (2009)
18. Ranasinghe, D.C., Cole, P.H.: An Evaluation Framework. In: Networked RFID Systems and Lightweight Cryptography, pp. 157–167. Springer (2008)

Appendix A.

We here present the performance comparison table for the WG-7 stream cipher from [11]. Another implementation of WG-7 stream cipher can be seen in [10].

Table 8. Comparison of WG-7 and other lightweight ciphers [11]

a	Cipher	Cost of Resources		Init. [cycles]	Thru.put [bits/sec]
		Code	EXP/RET		
	PRESENT@2MHz PRESENT@0.5MHz	841	25/4	230	2,297 574
	HB@2MHz HB@0.5MHz	1,532	9/7	22,949	5543 1,386
	WG-7@2MHz **WG-7**@0.5MHz	1,097	7/4	10,084	9,852 2,463
b		Flash	SRAM		
	AES@8MHz	6,664	88	7,149	81,432
	Salsa20@8MHz	3,842	258	318	83,688
	XTEA@8MHz	820	0	–	51655
	PRESENT@8MHz	2,398	528	–	53,361
	Size+HB@8MHz	1,308	0	14,735	34,934
	Speed+HB@8MHz	10,918	0	8,182	91,494
	GRAIN@8MHz	778	20	107,366	12,966
	TRIVIUM@8MHz	424	36	775,726	12,030
	WG-7@8MHz	1,100	0	10074	280,087

Performance Analysis of Cryptographic Acceleration in Multicore Environment

Yashpal Dutta and Varun Sethi

Freescale Semiconductor Inc.
Noida, Uttar Pradesh-201301, India
{yashpal.dutta,varun.sethi}@freescale.com

Abstract. With the increased capability to meet processing requirements and convergence of multiple servers, Multicore platforms are getting popular in the embedded space. Seamless performance scaling is assumed by a system designer while migrating to a Multicore system. This may not be true, especially with the ever increasing cryptographic requirement of security servers in embedded space. Cryptographic computational requirements are being pushed beyond the capabilities of general purpose processors. Thus many of the advanced Multicore platforms also provide hardware cryptographic accelerators. On Multicore platforms it's possible to use the crypto accelerator in a SMP or an AMP configuration. In case of SMP configuration, OS controls the cryptographic accelerator sharing across multiple applications. In a virtualized environment the crypto accelerator can be shared across multiple guest operating systems under the supervision of the hypervisor. Hypervisor utilizes the services of an IOMMU to isolate crypto operations and data across various guest OS partitions. A proper analysis of each of the design configurations is required in order to select the best possible option while designing security server over an embedded system. The paper covers cryptographic processing for security servers on SMP Linux and in a virtualized environment (running with a hypervisor [6]).

Keywords: Cryptography, OpenSSL, Cryptodev, Multicore, Hypervisor.

1 Introduction

In an embedded environment, cryptographic processing plays a critical role. The cryptographic applications range from the ones requiring basic cryptographic operation like AES, TDES or MD5 to ones requiring complex protocol level processing like Apache Web Server[5] working with SSL/TLS protocols or Strongswan[6] providing IPSec stack working with cryptographic algorithms for VPN. The cryptographic operations generally consume a lot of CPU cycles. A single core platform becomes bottleneck in security server applications and thus there is a need of multiple servers working in a load sharing environment to meet expected processing requirements. Multicore platform helps in convergence of multiple servers in a single system. CPU may not always be a bottleneck when cryptographic

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 658–667, 2013.

operations are performed in a system, but critical platform resources like cache, memory modules and platform bus may also contribute to performance challenges.

In case of a SMP operating system, hardware resources are shared across multiple user-space processes. In hypervisor controlled virtualized configuration, it is possible to both physically partition and share resources across various guest OS partitions. The resources on a server platform include CPUs, memory, Caches and IO devices including crypto accelerators. The cumulative server performance improves with optimized resource partitioning.

In this paper we discuss cryptographic processing design options available for SMP and AMP configurations on Multicore platforms. We specifically look at ways cryptographic processing can be optimally offloaded to cryptographic accelerator in each of these configurations. The paper concludes with the experimental results, pros and cons of security server running on Multicore platform under various design approaches.

2 Building Blocks

The section covers software and hardware components involved in our experiments under various configurations.

2.1 OpenSSL Library

OpenSSL[2] is an open source toolkit for SSL2.0/3.0, TLS1.0. OpenSSL is one of the standard user-space cryptography library supporting large number of general purpose symmetric ciphers operations (AES, DES, TDES etc), digests operations (MD2, MD4, MD5, SHA1 etc) and asymmetric cipher operations (RSA, DSA, DH etc). It provides interface to offload cryptographic operations to hardware accelerator with its engine interface.

Cryptodev [3] engine is one of the engine interfaces used by OpenSSL for offloading cryptographic operations to hardware accelerators. The module supports multiple hardware accelerators registered with CryptoAPI. Other engine available for offloading cryptographic operations from OpenSSL library to hardware accelerators are proprietary engine interface, AF_ALG and OCF-Linux [4].

The CryptoAPI infrastructure provides cryptographic operation handling within Linux kernel. The infrastructure includes a mechanism to register and offload cipher operation supported by a cryptographic accelerator driver. If multiple drivers support the same crypto operation, the driver with the highest priority is selected for the operation. Support exists for handling symmetric ciphers and digests in the Linux kernel. This support is available in Linux kernel version 2.6 onwards.

Cryptographic accelerator driver initializes the hardware accelerator and registers supported cipher operations to CryptoAPI infrastructure in Linux kernel. Cryptographic operation is only requested by hardware accelerator if operation is registered by driver with CryptoAPI. Figure 1 below shows various layers involved in OpenSSL cryptographic processing in our experiments.

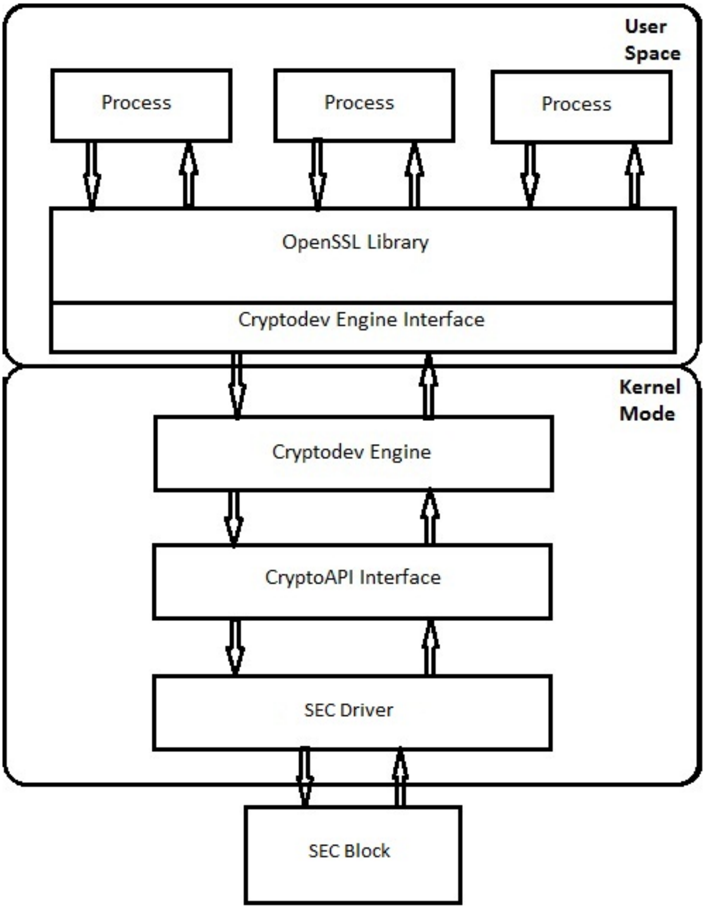


Fig. 1. OpenSSL Library with Cryptodev Engine

2.2 Cryptographic Acceleration

A cryptographic hardware accelerator offloads actual math operation related to cryptographic operation. The accelerator may implement common cryptographic processing including symmetric Cipher operations (e.g. AES, DES, TDES etc), Digest operations (MD5, SHA1, SHA-256 etc), Public key Ciphers (RSA, DSA, DH etc), Protocol Offloading (e.g. IPSec, SSL, TLS, DTLS etc) and random number generations.

For this paper, we used Freescale Multicore platform with integrated security accelerator called SEC. Figure 2 shows high level block diagram of SEC block. Parallel sub-blocks processing multiple cryptographic Job's in parallel help in scaling performance. The Job Queue controller unit checks whether the cryptographic jobs can be processed in parallel. The decision depends on whether there is inter-dependency among two cryptographic jobs. This is true for situations like multi-pass hashing on a sequence of buffers requiring init, update and final-state processing.

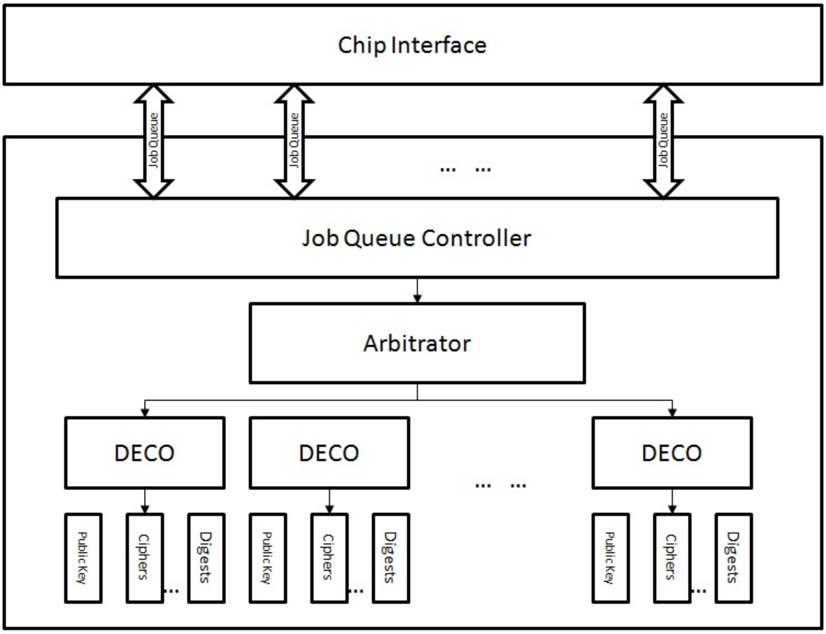


Fig. 2. Freescale SEC Block

2.3 Hypervisor

Freescale Embedded Hypervisor [6] is used to run OS in supervised configuration. This provides easy porting of an OS to hypervisor with minimal performance impact.

3 Cryptographic Configuration Options

3.1 SMP Configuration

In SMP configuration, all the processing threads use cryptographic support provided by cryptographic accelerator. The system resources like system bus and caches are used without any partitioning [8]. Thus, a lower priority security process can starve high priority process by consuming shared system resources. E.g. cached data corresponding to a high priority flow may get evicted by cache line corresponding to a low priority flow. Such issues can be mitigated by proper QoS implementation at hardware level to handle high priority crypto requests before lower priority process. Figure 3 shows cryptographic accelerator access by multiple processing running under SMP configuration.

The performance of a cryptographic process is impacted by presence of other processes under SMP configuration which can impact scheduling of the processes on CPUs. Proper configuration of scheduling parameters of various processing running on CPU helps mitigate performance impact.

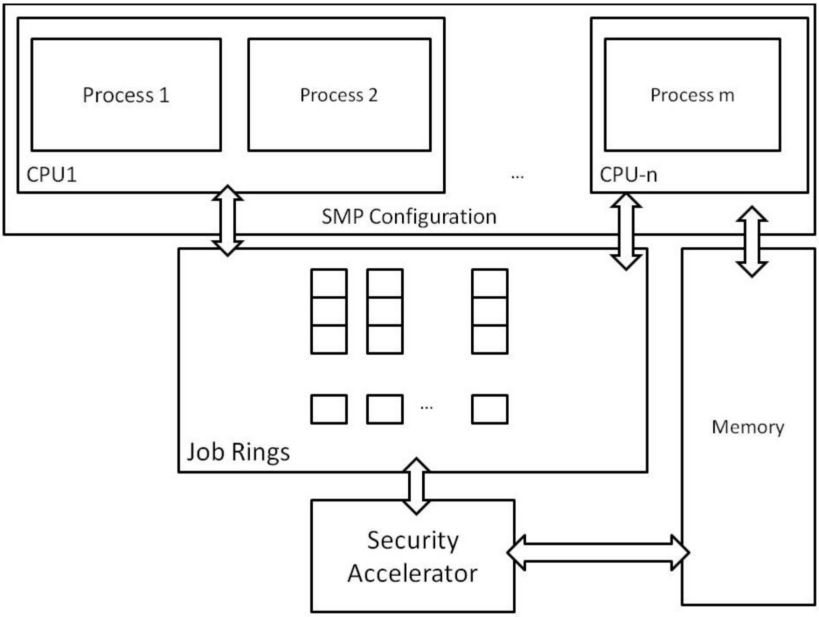


Fig. 3. Security accelerator sharing in SMP Configuration

3.2 Virtualized Configuration

In a virtualized environment the cryptographic accelerator can be shared across multiple guest operating systems under the supervision of the hypervisor. Hypervisor utilizes the services of an IOMMU to isolate cryptographic operations and data across various guest operating system partitions. IOMMU ensures that unless memory is shared, no two partitions can access each other's memory. An attempt to access memory of other guest partition is checked and violation is raised by IOMMU.

The partition virtualized environment offloads its cryptographic acceleration requirements with the notion that it owns the block. If system resources could be partitioned among guest partitions, the issues like cacheline eviction by unrelated partition are avoided and thus reduce interference among partitions. Figure 4 shows a virtualized configuration under hypervisor control.

One issue with such a partitioning is that the global configuration space associated with security block must be owned by one partition which can work as a control partition for the platform. Assumption made by all partitions offloading their security operation to security block is that security block is already initialized by control partition.

Access to hypervisor controlled resources like MMU or interrupt controllers can lead to significant overhead, thus impacting performance of the cryptographic applications.

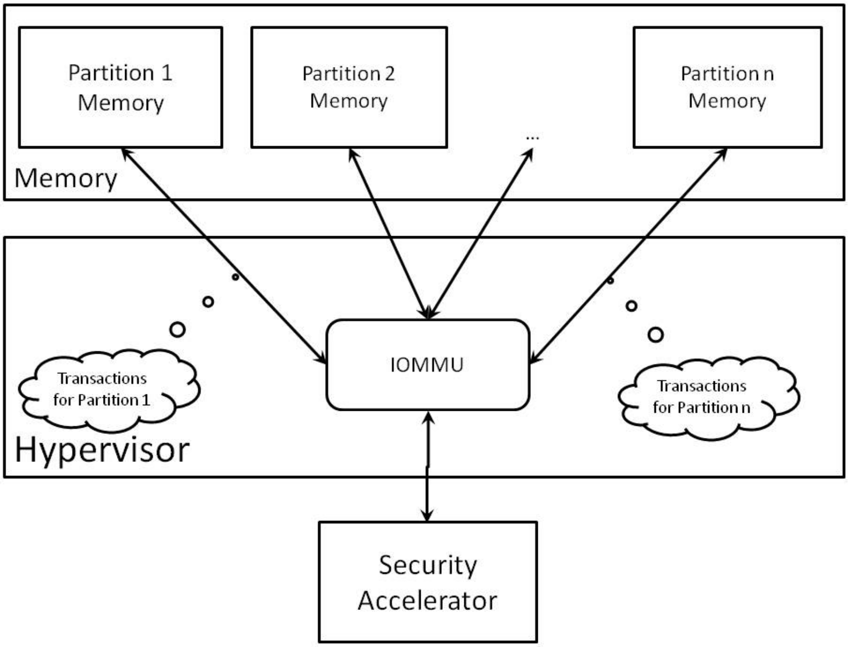


Fig. 4. Virtualization of Security hardware accelerator

4 Experimental Setup

We performed our experiments on Freescale QorIQ P4080 Multicore platform. This platform has eight e500mc cores each running at 1.5GHz. Each core has 32KB L1 instruction and data cache and a 128 KB unified backside L2 cache. The platform has a shared 2MB platform cache. For our experiments, we used a system with 4GB DDR with DDR bank interleaving. The crypto accelerator block used is the Freescale security block IP.

4.1 Setup for the SMP Configuration

The experiments were performed on Linux kernel version 3.0.48. OpenSSL version 1.0.1c and cryptodev version 1.5 were used for the experiments.

4.2 Setup for the AMP configuration

The virtualized setup consisted of lightweight baremetal executive guest partitions running on Freescale embedded hypervisor.

5 Performance Results

5.1 Results on Native SMP Configuration

The results are obtained using speed test built-in OpenSSL with and without cryptodev engine. Results show that performance gain varies for different size of buffer and number of parallel threads. The results in figure 5 shows MD5 and SHA1 performance comparison for offloaded operation against software based cryptographic implementation with single thread of speed test.

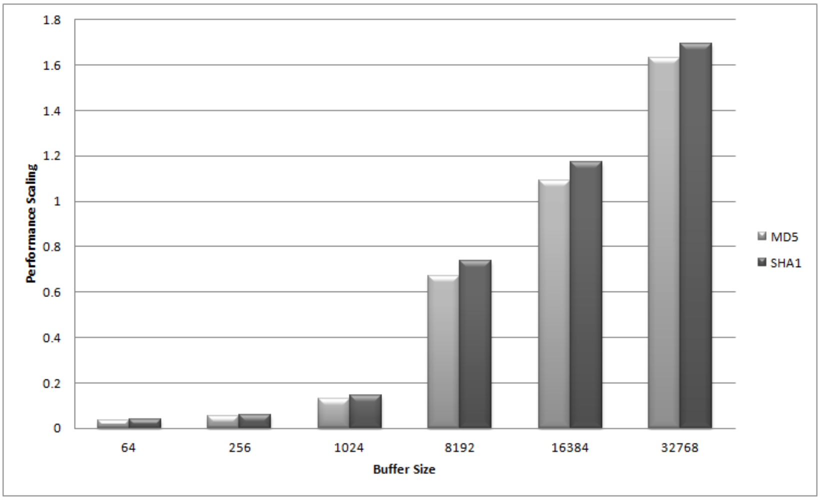


Fig. 5. MD5 and SHA1 Performance Scaling

Figure 6 shows performance scaling results for AES-CBC and DES-CBC cryptographic operations. From experimental results in graphs, we can see that cost of offloading gives less CPU bandwidth saving for small sized packet than large sized packets.

In presence of multiple accelerator sub-blocks capable of performing parallel processing of different and unrelated cryptographic operations, the performance of application threads can be improved.

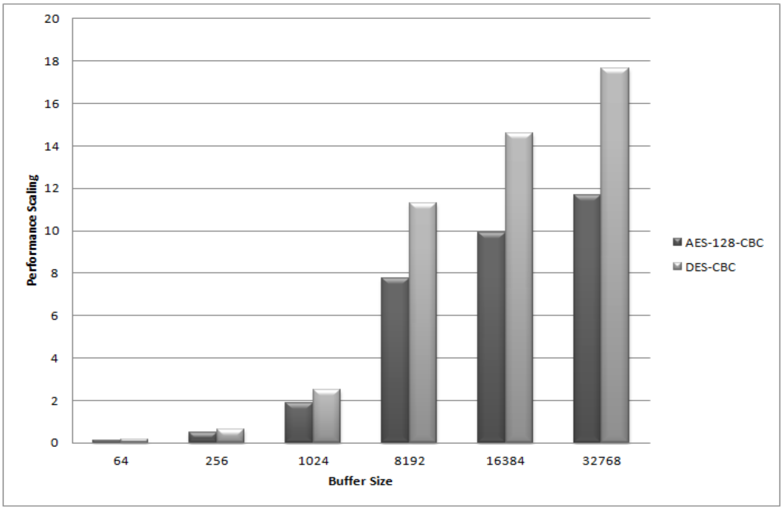


Fig. 6. Performance scaling of single threaded AES-128-CBC and DES-CBC

Figure 7 above shows performance scaling with security application threads running in parallel natively on Multicore platform.

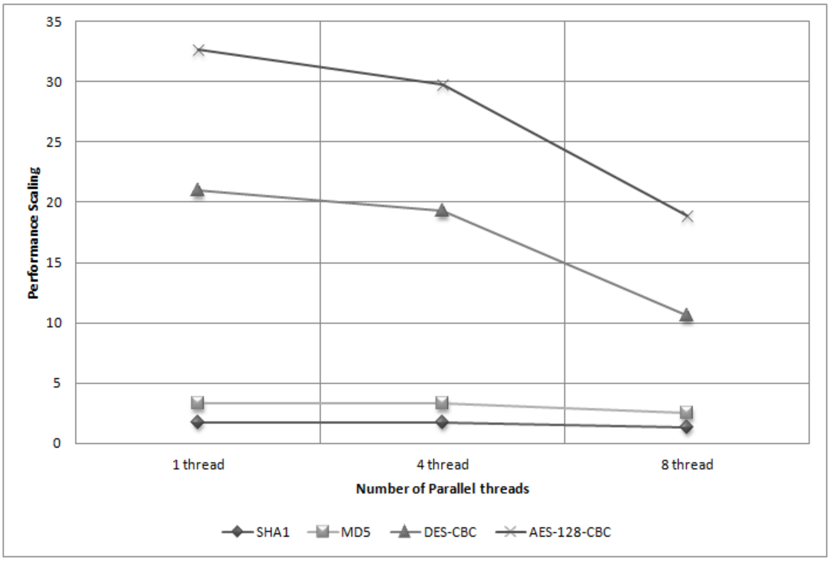


Fig. 7. Performance scaling with multiple threads

5.2 Results on Supervised AMP partition

Figure 8 below shows performance scaling with a proprietary security application running under Freescale Hypervisor. The application exercises security accelerator for IPSec protocol processing. Performance scales almost linearly for small sized frames but for large size frame, the performance scales for 3 partitions and additional cores/partition don't give performance benefit due to bandwidth limitation of hardware accelerator.

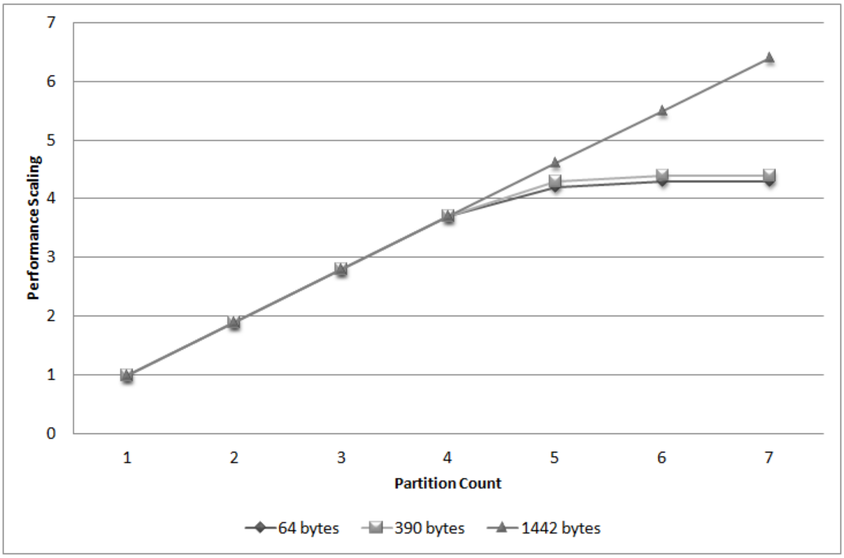


Fig. 8. IPSec Performance scaling of Virtualized Security Block

6 Conclusion

On Multicore platforms, there are multiple options available for Security Server design. Convergence of multiple security servers under SMP and virtualized configuration is possible. The sections above looked at both configurations on Freescale Multicore platforms. The paper also shared performance scaling of OpenSSL library on Multicore platform with and without hardware acceleration. It could be seen that performance scaling stops beyond a number of CPUs with hardware accelerator. This is the point where CPUs were generating more cryptographic operation requests for hardware accelerator than it could handle and there is a need for congestion support to avoid overloading on cryptographic acceleration [8]. Virtualization of cryptographic accelerator helps in isolation and security of application domains running in separate partitions. This separate traffic domain in virtualization for a partition reduces performance impact on traffic running on other partition which is lacking in SMP Linux.

References

1. Freescale P4080 QorIQ processor: http://www.freescale.com/webapp/sps/site/prod_summary.jsp?code=P4080
2. OpenSSL project: http://www.openssl.org/
3. Cryptodev engine: http://home.gna.org/cryptodev-linux/
4. OCF-Linux project: http://ocf-linux.sourceforge.net/
5. Apache Web Server: http://www.apache.org
6. Freescale Embedded Hypervisor: http://cache.freescale.com/files/32bit/doc/white_paper/EMBEDDED_HYPERVISOR.pdf?fsrch=1&sr=2
7. Strongswan project: http://www.strongswan.org/
8. Dutta, Y., Malik, S.: Hemant Agrawal: Multicore Development Challenges in Embedded Space, ARM TechCon-2012 (2012)

A Review on Wireless Network Security

Sandeep Sharma, Rajesh Mishra, and Karan Singh

School of ICT, Gautam Buddha University
Greater Noida, Gautam Budh Nagar, U.P., India
{sandeepsharma,rmishra,karan}@gbu.ac.in

Abstract. Computer network is very essential part of our life by which we can share the information via different technologies such as wired or wireless. Generally the wireless is mostly adopted technology by us due to various advantages like ease of installation, mobility, reconfigure ability, low infrastructural cost etc. but suffers from more attacks as the wireless channel is open. Therefore, many researchers are working in this hot area to secure the wireless communication. In this paper, we discuss the WEP, WPA, WPA2 and the RSA protocols and give the comparative study.

Keywords: Wireless Network, Network Security, Attack, Wireless Authentication, EAP, WEP, WPA, TKIP.

1 Introduction

In recent years the number of the computer users increases drastically and exponentially due to their interest in the internet usability and computing needs. The proliferation of laptop computers and PDA's has caused an increase in the range of the places where the people performing computing like schools, colleges ,business centres and even in the houses. Wireless networks offer mobility to the users due to which everybody wants to join it. As the number of the users are increasing hence the security of the message is the main concern. The devices comprises of the wireless network are available to the potential intruders unintended information. Although a number of cryptographic algorithms are available which provides a high level of security, still there is a need of and also modifiable for such intrusions. If the intruder is within the range, he can listen to the more secure algorithm. When connectivity to the network is needed, wireless networks is preferred over its wired counterpart and here comes the popular IEEE 802.11 standards is used in the picture. The IEEE 802.11 standard defines the protocols for two types of networks: Ad-hoc networks and Infrastructure networks. The Ad-hoc network is a simple network where communication is established between the stations in the given coverage region without using a server or wireless Access Point (AP).This standard provides the way to all the stations to have a fair access to the wireless network. It provides the method to initialize a request to use the media to ensure that all the users in the Base Service Set (BSS) can have maximum throughput. The Infrastructure networks uses the wireless Access Point (AP) which acts as an controller to control allocation of the transmit-time for all the

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 668–681, 2013.

stations and allows the mobile terminals to roam here and there in their own cell and from one cell to another cell. The access point is used to handle traffic from the mobile terminals to the wired or wireless backbone of the infrastructure network. The wireless access point routes all the data between the stations and other stations or to and from the network server. Before communicating data, the wireless client must establish association and only after an association two wireless stations can exchange data between them. In the infrastructure mode, the client associate with an access point which is a 2 step process and involves three stages:

- Unauthenticated and unassociated
- Authenticated and unassociated
- Authenticated and associated

The transitions from one stage to another takes place by the exchange of messages called as management frames. After a fixed time interval all Access Points (APs) transmits a frame known as beacon management frame which is listen by the client in the coverage region. All the network names i.e. the service set identifiers (SSID) which contains the beacon frames are used to identify the network to be associated with. The client-access point authentication is then done by the exchange of several management frames as the part of the authentication process. There are two types of the authentication which are Open System Authentication (OSA) and Shared Key Authentication (SKA).After the authentication gets successful the client moves into the second stage, authenticated and unassocaiated stage. And after the client sends an association request frame and the access point responds with an association response frame the stage enters from the second stage to the third stage. After the completion of the third stage client becomes a peer and can transmit the data frames.

Fig. 1. A Wireless LAN

The paper is arranged in the following way: we begin with the discussion about the attacks in the wireless LAN in Section 2, and the security goals in Section 3. In the section 4, we are providing different security mechanisms in 802.11 standards. We present comparative summary of WEP, WPA and RSA security protocols in the Section 5 and finally concludes the paper in Section 6.

2 Attack in WLAN

Attack is defined as a potential for violation of security, which exists when there is a circumstance, capability, action or event that could breach security and cause harm, where as a threat is a possible danger that might exploit vulnerability. Attack is an assault on the system security that derives from an intelligent threat i.e. an intelligent act that is a deliberate attempt to evade security service and violate the security policy of the system. Attacks in the wireless networks can be classified into two main parts: active and passive.

2.1 Active Attacks:

An active attack occurs when an unauthorized party makes modifications to a message, data stream, or file. In the active attack the attacker first receive the information from the system and then modify it. The different categories of active attack are as follows:

- *Masquerade*: where one entity pretends to be a different entity.
- *Replay*: This involves the passive capture of a data unit and its subsequent retransmission to produce an unauthorized effect.
- *Modification of messages*: It means that some of the portion of the legitimate message is altered or that message is delayed or reordered to produce an unauthorized effect.
- *Denial of service*: It prevents the normal use of the management of the communication facilities. Another form is the disruption of an entire network, either by disabling the network or by overloading it with messages so as to degrade the performance. It is discussed in [10, 38, 44]
- *Alteration*: This involves some change in the original message.

2.2 Passive Attacks:

A passive attack is an attack in which an unauthorized party gains access to an asset but does not modify its content or engage in communication with any node in the network. Passive attacks involve eavesdropping and traffic analysis. Eavesdropping is when the attacker monitors packet transmissions for the message content.

- *Traffic Analysis:* In this type of the attack the attacker try to figure out the similarities between the messages to come up with some sort of pattern that provides some clues regarding the communication that is taking place between the legitimate transmitter and receiver.
- *Release of the message contents:* In this type of the attack, the secret message between two entities is exposed to the unwanted intruder.

A passive attack is normally undetectable, while an active attack can usually be detected. Even though it is possible for one to detect an active attack that does not mean an active attack is preventable. In the client-attacker environment some form of communication is set up between an attacker and one or more nodes in the network. Effectively, active attack involves changing data in the packet.

3 Security Goals

Security is one of the critical attributes of any communication network. The security aspect comes into the scene when it is necessary to protect the information transmission from an opponent who may present a threat to confidentiality, authentication and so on. The major security attributes are Confidentiality, Integrity and Availability which is commonly known as (CIA).Along with the CIA the other attributes includes Authenticity and Accountability. These security attributes can be defined as follows:

- ***Confidentiality*:** This term covers two related concepts

  Data confidentiality: Assures that private or confidential information is not made available or disclosed to unauthorized individuals.
  Privacy: Assures that individuals control or influence what information related to them may be collected and stored and by whom and to whom that information may be disclosed.

- ***Integrity:*** This term covers two related concepts:

  *Data integrity***:** Assures that information and programs are changed only in a specified and authorized manner.
  System integrity: Assures that a system performs its intended function in an unimpaired manner, free from deliberate or inadvertent unauthorized manipulation of the system.

- ***Availability:*** Assures that systems work promptly and service is not denied to the authorize users.
- ***Authenticity:*** The property of being genuine and being able to be verified and trusted, confidence in the validity of a transmission, a message, or message originator. This means verifying that the message is coming from a trusted source or legitimate user.
- ***Accountability:*** The security goal that generates the requirement for actions of an entity to be traced uniquely to that entity. This supports non-repudiation, deterrence, fault isolation, intrusion detection and prevention, and after-action recovery and legal action. Because truly secure systems are not yet an achievable goal, we must be able to trace a security breach to a responsible party. Systems must keep records of their activities to permit later forensic analysis to trace security breaches or to aid in transaction disputes.

4 Security Mechanisms in IEEE 802.11 Standards

IEEE 802.11 provides several mechanisms to provide a secure environment for the wireless network access and this section discusses all of them in short.

4.1 Wired Equivalent Privacy (WEP) Protocol

WEP provides data encryption and integrity protection for the 802.11 standards. It is proved to be unsecure protocol and hence vulnerable to network attacks and can be cracked easily [1, 2, 3]. WEP with the 802.1X is called as the dynamic WEP which in a non standard technology that some of the vendors were using to overcome the weaknesses of the static WEP. Whether it is a static WEP or dynamic WEP, both of them have security issues and hence there is a need of more secure protocols such as WPA/WPA2.WEP is less secure and uses 40 or 104 bit encryption scheme in the IEEE 802.11 standards [4].WEP weaknesses are as follows:

- It does not prevent forgery of the packets.
- It does not prevent the replay attack in which the Attackers can simply record the packet and replay them as desired and they will be accepted by the legitimate user.
- WEP uses RC4 improperly and the key used for the encryptions are very weak and can be brute-forced on standard computers in hours or minutes using the freely available softwares on the internet.
- WEP reuses initialization vectors. A variety of available cryptanalytic methods can decrypt data without knowing the encryption key.
- WEP allows modification in the message without knowing the encryption key by an attacker.
- Key management is a lack and updating is very poor.
- Problem related to the RC-4 algorithm.
- Easy to forge the authentication messages.

4.2 The WPA and WPA2 Protocol

In 2003, the Wi-Fi Alliance [19, 20] introduced a new protocol, Wi-Fi Protected Access (WPA) as a strong standard-based interoperable Wi-Fi Security Mechanism. WPA addressed all the vulnerabilities which were not addressed by the WEP.WPA protocol also provides authentication and replaces WEP with its strong encryption technology called as Temporal Key Integrity Protocol (TKIP) with the Message Integrity Check (MIC). For the mutual authentication of the clients WPA uses either IEEE802.11X/Extensible Authentication Protocol (EAP) authentication or the Pre-Shared Key (PSK), [3].

In 2004, WPA2 was launched by the Wi-Fi Security and like the WPA it supports 802.1X/EAP authentication or PSK technology [6]. It also includes the advanced encryption mechanism using the Counter-Mode/CBC-MAK Protocol (CCMP) called the Advanced Encryption Standard (AES) [9].

4.3 Attacks Handling with WPA and WPA2 Protocol

Both WPA and WPA2 protects the wireless networks from variety of attacks such as man-in-the-middle, authentication forging, replay, key collisions, weak keys, packet forging, and brute-force attacks.WPA/WPA2 addresses all the weaknesses of the original WEP protocol which has weak authentication and imperfect and inefficient encryption key implementation.

It uses TKIP which has enhanced the encryption algorithm and authentication method with the 802.1X/EAP authentications. TKIP uses a 128 bit per packet key per user per session to provide strong encryption.

Table 1. Comparative chart showing WPA and WPA2 modes

	WPA	WPA2
Enterprise Mode	Authentication: IEEE 802.1X/EAP Encryption: TKIP/MIC	Authentication: IEEE 802.1X/EAP Encryption: AES-CCMP
Personal Mode	Authentication: PSK Encryption: TKIP/MIC	Authentication: PSK Encryption: AES-CCMP

4.4 An Overview of the WPA/WPA2 Authentication Process

The authentication process in WPA and WPA2 has the following components

- **The Client Supplicant:** It is a software that is installed on the client to implement the IEEE 802.1X protocol framework and on or more Extensible Authentication Protocol (EAP) methods.
- **Access Point**: These are the service point trough which we can have the network access after successful authentication and authorization process.
- **Authentication Server**: WPA and WPA2 use IEEE 802.1X authentication with the EAP types which provides the mutual authentication on the wireless network. The authentication server stores the list of the names and credentials of the authorized users against which the server verifies the authentic user and denies the unauthentic one. For this purpose a Remote Authentication Dial-in User Service (RADIUS) Server is generally used.

In the WPA2 the mutual authentication is initiated by the user to be associated with the access point. The access point denies the request and blocks the user until the user is authenticated. Then the client provides credentials to the access point which is then communicated to the RADIUS server which uses the 802.1 X/EAP frameworks for authentication. This is the Extensible Authentication Protocol which finally gives the mutual authentication of the wireless client with the server via the access point. After the credentials were checked, the client joins the wireless network the WLAN. Once

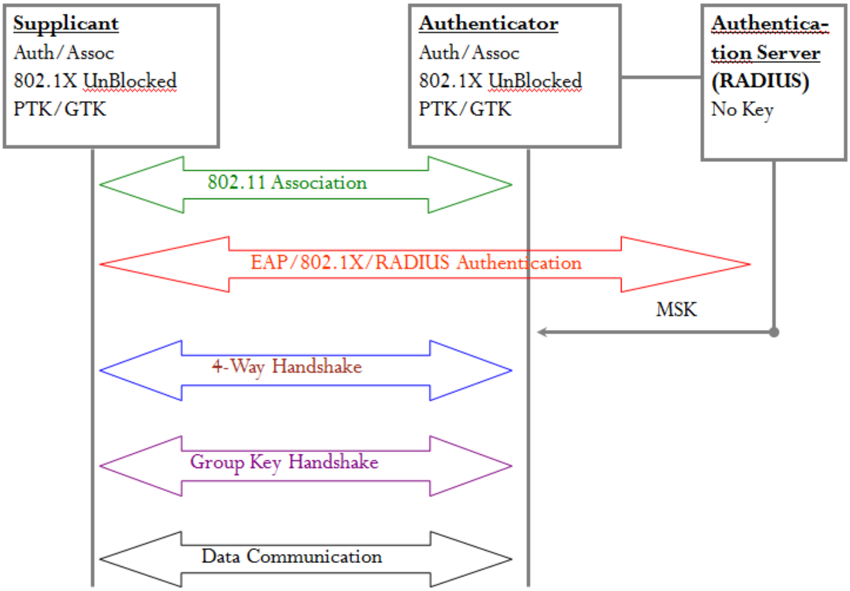


Fig. 2. Authentication process of WPA/WPA2

the wireless client has been authenticated, the authentication server and the client simultaneously generate a Pair-wise Maser Key (PMK). A 4-way handshake is established between the user [15, 22] and the access point and then the encryption keys are generated with the installation of the TKIP in the WPA or with the AES in the WPA2 environment. As the client sends data on the network, encryption protects the data exchanged between the cline and the access point (AP).

4.5 The Functioning of the WPA Encryption with the TKIP

WPA uses the TKIP protocol for the encryption, for which it uses a 128 bit per packet key per user per session instead of the 40/104 bit key in the predecessor WEP. The WPA uses a method which generates dynamic keys and removes the possibility of the key prediction by a potential intruder in the wireless network.WPA protocol also have a provision to check against the capturing, altering and relay/resending of the data packets through the use of the Message Integrity Check (MIC).In the OSI reference model of the network, the WPA protocol works on the Media Access Control (MAC) layer. The MIC provides a strong mathematical function which is computed at the transmitting and the receiving end and if it does not match with the MIC then the data is considered to be tempered by the intruder and hence the packet is dropped.

4.6 The Functioning of the WPA2 Encryption with the AES

The WPA2 protocol uses the AES which is a block cipher, a type of the symmetric key cipher (which uses the same key to encrypt a plain text and to decrypt the cipher text) that uses a group of bits of fixed length called the blocks [5]. AES employ a block size of 128 bits with 3 possible key lengths: 128,192 and 256. For the WPA2 implementation of the AES, a 128 bit key is used which includes 4 stages that makes a round. Each of these rounds are then goes through 10,12 or 14 iterations depending upon the key size, for example ,the WPA2/802.11i implementation of the AES , each round is iterated 10 times. The AES employs CCMP which enables a single key to be used for both the encryption and authentication. CCMP includes the Counter Mode (CTR) that is used for the data encryption and the Cipher Block Chaining Message Authentication Code (CBC-MAC) to provide the data integrity. The AES uses a 48-bit initialization vector (IV) which takes 2^{120} operations to be performed in order to break the AES key, making it a secure cryptographic algorithm for the wireless scenario [23].

4.7 Selecting the EAP

The Extensible Authentication Protocol (EAP) supported by the IEEE 802.1x includes Extensible Authentication Protocol-Transport Layer Security (EAP-TLS), Extensible

Table 2. Summary of the EAP types

Parameters	PEAP	EAP-TLS	EAP-TTLS
User Authentication	OTP,LDAP, NDS,	LDAP, NT Domains,	OTP, LDAP, NDS, NT Domains
Database and Server	NT omains, Active Directory	Active Directory	Active Directory
Native Operating System Support	Windows XP, 2000	Windows XP, 2000	Windows XP, 2000, ME, 98, WinCE, Pocket PC2000, Mobile 2003
User Authentication Method	Password or OTP	Digital Certificate	Password or OTP
Authentication Transaction Overhead	Moderate	Substantial	Moderate
Management Deployment Complexity	Moderate Digital Certificate For Server	Substantial Digital Certificate Per Client and For Server	Moderate Digital Certificate For Server
Single Sign On	Yes	Yes	Yes

Authentication Protocol-Tunnelled Transport Layer Security (EAP-TTLS), Protected-EAP or simply PEAPv.0 or PEAPv.1, Extensible Authentication Protocol-Message Digest 5 (EAP-MD5) etc. [24, 42]. Different supplicants and networks use different EAP types which offer different advantages, disadvantages and their overheads. Some are good where the access is controlled by simple passwords and some proves to be the best when the client-server certificate is required. The EAP type adopted depends upon the type of the network environment and the security level required. Table 2 give us a comparative study of PEAP, EAP-TLS and EAP-TTLS on parameters such as the user authentication, database and the server, operating system support, user authentication methods, authentication overheads and deployment complexity etc.

4.8 EAP Overview

EAP was originally proposed for the point-to-point (PPP) protocol for an optional authentication phase after the PPP link is fored.EAP supports a variety of authentication methods such as token card, one-time password, certificate, public key authentication and smart cards. As shown in the figure 2, there can be various authentication mechanisms in the authentication layer such as the TLS, TTLS, MD5 etc. and can be modified to enter a new member.

4.9 Robust Security Networks (RSNs)

In 2004, the 802.11i was introduced that uses the concept of a Robust Security Network (RSN), where wireless devices need to handle additional capabilities [44]. This

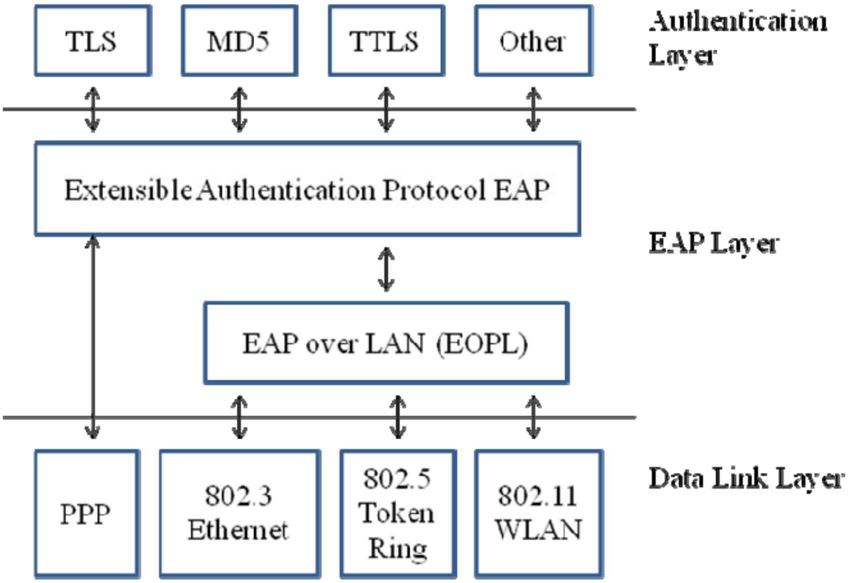


Fig. 3. EAP and its associated layers

new standard and architecture utilizes the IEEE 802.1X standard for access control and Advanced Encryption Standard (AES) for encryption. It uses a pair-wise key exchange (4 way handshake) protocol utilizing 802.1X for mutual authentication and key management process. 802.11i allows for various network implementations and can use TKIP, but by default RSN uses AES (Advanced Encryption Standard) and CCMP (Counter Mode CBC MAC Protocol) and it is this which provides for a stronger and scalable solution to the security problem.

4.10 Working of RSN

RSN uses dynamic negotiation of authentication and encryption algorithms between the access points (APs) and the mobile devices. The authentication schemes are based on 802.1X and Extensible Authentication Protocol (EAP). The encryption algorithm is Advanced Encryption Standard (AES). Dynamic negotiation of authentication and encryption algorithms lets RSN evolve with the state of the art in security of the network. Using dynamic negotiation, 802.1X, EAP and AES, RSN is considerably stronger than WEP and WPA. However, RSN would run very feebly on the legacy devices. Unfortunately only the latest devices have the capability required to accelerate the algorithms in clients and access points, providing the performance expected of today's WLAN products.

4.11 RSN Assessment

WPA had improved security of legacy devices to a modestly acceptable level with one exception (pass phrases not less than 20 characters), but RSN is the future of the wireless security (over-the-air security) for 802.11 WLANs.

5 Comparison of WEP, WPA AND RSN Security Protocols

WEP has been regarded as a collapse in wireless security, as it has been accepted by the IEEE that WEP was not designed to provide full security. The original WEP security standard, using RC4 cipher is widely considered to be vulnerable and broken due to the use of the insecure IV usage.

It uses 40 bits of encryption key RC4 cipher by default (with vendor specific longer key support exceptions), concatenates key with IV values per packet sent over the wireless channel, with no key management mechanism embedded, having no automatic or periodic key change attribute associated with it, causing re-use and easy to capture small sized IVs that leads to key deciphering to the third parties. The data integrity check mechanism of WEP is not cipher protected and uses CRC-32; ICV providing no header integrity control mechanism and be short of the replay attack prevention method [12].

Table 3. Comparison summary of WEP, WPA and RSA

Features of Mechanism	**WEP**	**WPA**	**RSN**
Encryption Cipher Mechanism	RC4 (Vulnerable - IV Usage)	RC4 / TKIP	AES /CCMP CCMP /TKIP
Encryption Key size	40 bits *	128 bits	128 bits
Encryption Key Per Packet	Concatenated	Mixed	No need
Encryption Key Management	None	802.1x	802.1x
Encryption Key Change	None	For Each Packet	No need
IV Size	24 bits	48 bits	48 bits
Authentication	Weak	802.1x - EAP	802.1x - EAP
Data Integrity	CRC 32 - ICV	MIC (Michael)	CCM
Header Integrity	None	MIC (Michael)	CCM
Replay Attack Prevention	None	IV Sequence	IV Sequence
* Some vendors apply 104 and 232 bits key, where the 802.11 Requires 40 bits of encryption key.			

WPA is a provisional solution to the WEP vulnerability uses a subset of 802.11i features and had been generally assumed as a major security improvement in wireless environment. WPA has various enhancements over WEP. Namely, RC4 ñ TKIP encryption cipher mechanism, 128 bits of key size, mixed type of encryption key per packet usage, 802.1x dynamic key management mechanism, 48 bits of IV size, 802.1x ñ EAP usage for authentication, providing data integrity and header integrity, ciphering aspect via MIC that is inserted into TKIP and IV sequence mechanism to prevent replay attacks and support for existing wireless infrastructures. Table-3 gives the comparison of WEP, WPA and RSN Security Protocols. RSN seems to be the strongest contender among all the security protocol for wireless networks as far as all previously declared vulnerabilities and drawbacks associated to WEP and WPA are concerned. After the 802.11i standard is ratified, RSN is accepted as the concluding solution to wireless security, expected to provide the robust security required for wireless environments. RSN provides all the advantages of WPA in addition to stronger encryption through the implementation of AES, roaming support and CCM mechanism for data and header integrity. WPA supports existing wireless infrastructures. WPA deployments over current WEP installations provide cost effective and hassle free shifts where vendors can transit to the WPA standard through a software or firmware upgrade. For RSN this is not the case. It requires extra hardware upgrade in order to implement AES.

6 Conclusions

The objective of this paper is to make aware the readers about the wireless network security and the security protocols used in the wireless network such as WEP, WPA, WPA2 and RSN. These papers discuss about the advantages and disadvantages associated with the security protocols for 802.11. There are various authors who have written about the security weaknesses of the WEP and WPA. In this paper an overview and comparison of the WEP, WPA and RSA is given as a comparative chart which shows that RSA perform better than the WEP and WPA. RSN seems to be the strongest challenger among all the security protocols as it addresses all the unaddressed and previously declared vulnerabilities and drawbacks associated to WEP and WPA. RSN provides all the advantages of WPA in addition to stronger encryption through the implementation of AES, roaming support and CCM mechanism for data and header integrity.

References

1. Mishra, A., Shin, M., Arbaugh, W.A.: Your 802.11 network has no clothes. IEEE Commun. Mag. 9, 44–51 (2002)
2. Beck, M., Tews, E.: Practical attacks against WEP and WPA. In: Procedings of 2nd ACM Conference on Wireless Network Security, WiSec 2009, pp. 79–85 (2009)
3. Mishra, A., Arbaugh, W.A.: An Initial Security Analysis of IEEE 802.1X Standard, http://www.cs.umd.edu/~waa/1x.pdf
4. Reddy, S.V., Sai Ramani, K., Rijutha, K., Ali, S.M., Reddy, C.P.: Wireless Hacing-A WiFI Hack by Cracking WEP. In: IEEE Second International Conf. on Education Tech. and Computer, vol. 1, p. V1-189 – V1-193 (2010)
5. Lashkari, A.H., Danesh, M.M.S., Samadi, B.: A survey on wireless security protocols (WEP, WPA and WPA2/802.11i). In: 2nd IEEE International Conference on Computer Science and Information Technology, ICCSIT, pp. 48–52 (2009)
6. Chen, J.-C., Wang, Y.-P.: Extensible authentication protocol (EAP) and IEEE 802.1x: tutorial and empirical experience. IEEE Communication Magzine 43(12), s26–s32 (2005)
7. Liu, Y., Jin, Z., Wang, Y.: Survey on security scheme and attacking methods of WPA/WPA2. In: IEEE 6th International conf. on Wireless Communication Networking and Mobile Computing, pp. 1–4 (2010)
8. Walker, J.: Unsafe at any key size: an analysis of the WEP encapsulation. IEEE Document 802.11-00/362 (2000)

9. Borisov, N., Goldberg, I., Wagner, D.: Intercepting mobile communications: the insecurity of 802.11. In: Proc. ACM Annual International Conference on Mobile Computing and Networking (MOBICOM), pp. 180–189 (September 2002)
10. Wang, L., Srinivasan, B.: Analysis and Improvements over DoS Attacks against IEEE 802.11i Standard. In: IEEE Second International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 109–113 (2010)
11. Chiornita, A., Gheorghe, L., Rosner, D.: A Practical Analysis of EAP Authentication Methods. In: Roedunet International Conference, pp. 31–35 (2010)
12. Bachan, P., Singh, B.: Performance Evaluation of Authentication Protocols for IEEE 802.11 Standards. In: International Conference on Computer and Communication Technology, pp. 792–799 (2010)
13. Ali, H.B., Karim, M.R., Ashraf, M., Powers, D.M.W.: Modeling and Verification of EAP-TLS in Wireless LAN Environment. In: International Conference on Software Technology and Engineering, p. V2-41–V2-45 (2010)
14. He, D., Bu, J., Chan, S., Chen, C., Yin, M.: Privacy-Preserving Universal Authentication Protocol for Wireless Communications. IEEE Transactions on Wireless Communication 10(2), 431–436 (2011)
15. Fluhrer, S.R., Mantin, I., Shamir, A.: Weaknesses in the key scheduling algorithm of RC4. In: Vaudenay, S., Youssef, A.M. (eds.) SAC 2001. LNCS, vol. 2259, pp. 1–24. Springer, Heidelberg (2001)
16. Bellardo, J., Savage, S.: 802.11 denial-of-service attacks: Realvulnerabilities and practical solutions. In: 12th USENIX Security Symposium, Washington, D.C., pp. 15–27 (August 2003)
17. Zha, X., Ma, M.: Security Improvements of IEEE 802.11i 4-way Handshake Scheme. In: International Conference Communications Systems, pp. 667–671 (2010)
18. Beck, M., Tews, E.: Practical attacks against WEP and WPA. In: Proceedings of the 2nd ACM Conference on Wireless Network Security, WiSec 2009, pp. 79–85 (2009)
19. Wi-Fi Alliance, Wi-fi protected setup specification (2007)
20. IEEE 802.11a, 802.11b, 802.11g, 802.11n, 802.11i standards, `http://standards.ieee.org`
21. Zeng, K., Govindan, K., Mohapatra, P.: Non-Cryptographic Authentication and Identification in Wireless Networks. IEEE Journal on Wireless Comm. 17(5), 56–62 (2010)
22. Sharma, A., Ojha, V., Lenka, S.K.: Quaantam Key Distribution in WLAN 802.11 Networks. In: International Conference on Networking and Information Technology, pp. 402–405 (2010)
23. Trappe, W., Washington, L.C.: Introduction to Cryptography with Coding Theory. Prentice Hall, Upper Saddle River (2002)
24. Ma, Y., Cao, X.: How to use EAP-TLS Authentication in PWAN Environment. In: IEEE International Conference on Neural Network and Signal Processing, vol. 2, pp. 1677–1680 (2003)
25. Srivastava, V., Motani, M.: Cross-layer design: A survey andthe road ahead. IEEE Communications Magazine, 112–119 (2005)
26. Thamilarasu, G., Balasubramanian, A., Mishra, S., Sridhar, R.: A cross-layer based intrusion detection approach for wireless ad hoc networks. In: IEEE International Conference on Mobile Adhoc and Sensor Systems Conference, 2005, vol. 7-10 (November 2005)
27. Kawadia, V., Kumar, P.: A cautionary perspective on cross layer design. IEEE Wireless Communication Magazine 12, 3–11 (2005)

28. Xiao, L., Greenstein, L.J., Mandayam, N.B., Trappe, W.: Using the Physical layer for Wireless Authentication in Time-Variant Channels. IEEE Transactions on Wireless Communication 7(7), 2571–2579 (2008)
29. Xiao, L., Greenstein, L.J., Mandayam, N.B., Trappe, W.: A Physical-Layer Technique to Enhance Authentication for Mobile Terminals. In: Proc. IEEE International Conference on Communications, Beijing, China, pp. 1520–1524 (2008)
30. Xiao, L., Greenstein, L., Mandayam, N., Periyalwar, S.: Distributed measurements for estimating and updating cellular system performance. IEEE Transactions on Communications 56, 991–998 (2008)
31. Xiao, L., Greenstein, L., Mandayam, N., Trappe, W.: MIMO-assisted channel-based authentication in wireless networks. In: Proc. IEEE Conf. Information Sciences and Systems (CISS), pp. 642–646 (March 2008)
32. Yu, P.L., Baras, J.S., Sadler, B.M.: Multicarrier Authentication at the Physical Layer. In: Proc. IEEE International Conference on Wireless, Mobile and Multimedia Networks, 2008, pp. 1–6 (2008)
33. Mathur, S., Reznik, A., Mukharjee, R., Rahman, A., Shah, Y., Trappe, W., Mandayam, N.: Exploiting the Physical Layer for Enhanced Security. IEEE Trans. on Wireless Comm. 17(5), 71–80 (2010)
34. Zeng, K., Govindan, K., Mohapatra, P.: Non-Cryptographic Authentication and Identification in Wireless Networks. IEEE Journal on Wireless Comm. 17(5), 56–62 (2010)
35. Ren, X., Zhang, J.: A Novel Cross-Layer Architecture for Wireless Protocol Stacks. In: Proc. International Conference on Multimedia Technology, 2010, pp. 1–6 (2010)
36. Corbett, C., Beyah, R., Copeland, J.: A passive approach to wireless NIC identification. In: Proc. IEEE International Conference on Communications, vol. 5, pp. 2329–2334 (June 2006)
37. Xiao, X., Ding, L., Zhou, N.: An Improved Mechanism for Four-Way Handshake Procedure in IEEES02.11. In: IEEE International Conference on Computer Science and Information Technology, pp. 419–422 (2010)
38. Wang, L., Srinivasan, B.: Analysis and Improvements over DoS Attacks against IEEE 802.11i Standard. In: IEEE Second International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 109–113 (2010)
39. Wenju, L., Yuzhen, S., Yan, Z., Ze, W.: An Analysis of Improved EAP-AKA Protocol. In: IEEE International Conference on Computer Engineering and Technology, pp. V1-10–V-13 (2010)
40. Liu, P., Zhou, P.: Formal Analysis of EAP-AKA based Protocol Composition Logic. In: IEEE International Conference on Future Computer and Communication, pp. V3-86–V3-90 (2010)
41. Chiornita, A., Gheorghe, L., Rosner, D.: A Practical Analysis of EAP Authentication Methods. In: Roedunet International Conference, pp. 31–35 (2010)
42. Ali, H.B., Karim, M.R., Ashraf, M., Powers, D.M.W.: Modeling and Verification of EAP-TLS in Wireless LAN Environment. In: International Conference on Software Technology and Engineering, 2010, pp. V2-41–V2-45 (2010)
43. Zha, X., Ma, M.: Security Improvements in IEEE 802.11i 4-way Handshake Scheme. In: IEEE International Conference on Communication Systems, pp. 667–671 (2010)
44. Wang, L., Srinivasan, B.: Analysis and Improvements over DoS Attacks against IEEE 802.11i Standard. In: International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 109–113 (2010)

Improved Proxy Signature Scheme without Bilinear Pairings

Sahadeo Padhye and Namita Tiwari

Department of Mathematics
Motilal Nehru National Institute of Technology
Allahabad-211004, India
{sahadeomathrsu,namita.mnnit}@gmail.com

Abstract. Proxy signature is an active research area in cryptography. In order to save the running time and the size of the signature, recently a provable secure proxy signature scheme without bilinear pairings has been proposed which is based on elliptic curve discrete log problem (ECDLP). In this paper, we point out some forgery attacks and security issues on this scheme. Furthermore, we also improve the scheme to make it secure against these forgeries. Our scheme is as efficient as previous proposed scheme.

Keywords: Digital signature, Proxy signature, Bilinear pairings, Elliptic curve discrete log problem.

1 Introduction

Digital signature offers source authentication in cryptography. To handle the situations arisen in digital world related to authentication, different types of digital signatures have been developed e.g a manager want to delegate his secretaries to sign documents without giving his private signing key, while he is on vacation. Proxy signature is the solution of such problem and firstly introduced by Mambo et al [11] in 1996. Proxy signature schemes can also be used in electronic transactions and mobile agent environment [10]. Since the proxy signature appears, it attracts many researcher's great attention. Using bilinear pairings, people proposed many proxy signature schemes [6,7,9,15,16,17]. All the above schemes are very practical, but they are based on bilinear pairings and the pairing is regarded as one of the expensive cryptography primitive. Therefore, to save the running time and to reduce the size of the signature, recently a provable secure proxy signature scheme without bilinear pairings [14] has been proposed which is based on ECDLP. In this paper, we point out some forgery attacks and security issues on this scheme. We show that scheme [14] does not satisfy prevention of misuse property. It has some other drawbacks also. Furthermore, we improve the scheme against these forgeries. Our improved scheme is as efficient as previous proposed scheme [14].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 682–688, 2013.

Roadmap: The rest of this paper is organized as follows. Some preliminary works are given in the section *2*. A brief review of scheme [14] is presented in section *3*. We discuss the forgeries on scheme [14] and present it's improved version in section *4* and *5* respectively. Section *6* presents the comparative analysis. Finally, conclusions are given in Section *7*.

2 Preliminaries

2.1 Background of Elliptic Curve Group

Let the symbol E/F_p denote an elliptic curve E over a prime finite field F_p, defined by an equation

$$y^2 = x^3 + ax + b, \ a, b \in F_p, \quad \text{and}$$

discriminant $\Delta = 4a^3 + 27b^2 \neq 0$

The points on E/F_p together with an extra point O called the point at infinity form a group $G = \{(x, y) : x, y \in F_p, E(x, y) = 0\} \cup \{O\}$.

Let the order of G be n. G is a cyclic additive group under the point addition "+" defined as follows: Let $P, Q \in G$, l be the line containing P and Q (tangent line to E/F_p if $P = Q$), and R, the third point of intersection of l with E/F_p. Let $l^{'}$ be the line connecting R and O. Then $P + Q$ is the point such that $l^{'}$ intersects E/F_p at R and O and $P + Q$.

Scalar multiplication over E/F_p can be computed as follows:

$$tP = \underbrace{P + P + ... + P}(t\ times).$$

2.2 Complexity Assumption

The following problem defined over G are assumed to be intractable within polynomial time.

Elliptic curve discrete logarithm problem (ECDLP): For $x \in_R {Z_n}^*$ and P the generator of G , given $Q = x.P$ compute x.

3 Brief Review of Scheme [14]

- Setup: Takes a security parameter k, and returns system parameters $\Omega = \{F_p, E/F_p, G, P, H_1, H_2, H_3\}$ as defined in Section 2.
  $H_1 : \{0, 1\}^* \to Z_n^*$, $H_2 : \{0, 1\}^* \times G \to Z_p^*$ and $H_3 : \{0, 1\}^* \to Z_p^*$ are three cryptographic secure hash functions.
- Extract: Each signer picks at random $sk_i \in Z_n^*$ and computes $pk_i = sk_iP$. Thus $(sk_i, pk_i), i \in \{o, p\}$ is private-public key pair.
- DelGen: This algorithm takes O's secret key sk_o and a warrant m_w as input, and outputs the delegation $W_{O\to P}$ as follows:
  a. Generates a random $a \in Z_n^*$ and computes $K = aP$.
  b. Computes $h_1 = H_2(m_w, pk_p)$ and $\sigma = h_1 sk_o + a$ mod$n$.
  O sends the delegation $W_{O\to P} = \{pk_o, m_w, K, \sigma\}$ to proxy signer P.

- DelVerif: To verify the delegation $W_{O \to P}$ and message warrant m_w, proxy signer P first computes
  $h_1 = H_2(m_w, p_{k_p})$, then checks whether
  $\sigma P = h_1 pk_o + K$.
- PKGen: If P accepts the delegation $W_{O \to P}$, he computes the proxy signing key sk_{pr} as:
  $sk_{pr} = \sigma h_2 + sk_p \bmod n$, where $h_2 = H_3(m_w)$.
- PSign: Takes system parameters, the proxy signing key sk_{pr} and a message m as inputs, returns a signature of the message m. The user P does as follows.
  a. Chooses at random $b \in Z_n^*$ and computes $R = hbP$, where $h = H_1(m)$.
  b. Computes $s = hb + sk_{pr} \bmod n$.
  The resulting signature is $(pk_o, pk_p, m_w, K, m, R, s)$.
- PSVerif: To check whether the signature $(pk_o, pk_p, m_w, K, m, R, s)$ is a valid proxy signature on message m under warrant m_w, verifier V first checks if the proxy signer and the message conform to m_w and computes $h_1 = H_2(m_w, pk_p), h_2 = H_3(m_w), h = H_1(m)$ then verify whether the following equation holds.
  $sP = R + [(h_1 pk_o + K)h_2 + pk_p]$.

4 Security Analysis of Scheme [14]

In this section, we will demonstrate that the scheme [14] has some drawbacks. As the first drawback, a forgery is given by malicious signer who is not designated as a proxy signer by the original signer. However, the malicious signer can forge a valid proxy signature on any message.

4.1 Forgery by Proxy Signer

After having the delegation σ on warrant message m_w, proxy signer makes the following forgery as follows:

- Chooses another warrant $m_w{}'$, computes $h_1{}' = H_2(m_w{}', pk_p)$.
- Computes $\sigma' = (\frac{h_1{}'}{h_1})\sigma$ s.t. $\sigma' = h_1' s_{k_o} + a h_1' h_1^{-1}$ $(mod\ n)$.
- This generated σ' satisfies $\sigma' P = h_1{}' pk_o + K'$ where $K' = a h_1' h_1^{-1} P$. So $(pk_o, m_w{}', K', \sigma')$ is a valid delegation on new warrant $m_w{}'$. Using this delegation, proxy signer can sign any message of it's own choice.

Thus proxy signer can misuse the right of delegation. Our attack is possible only because public parameter K lonely exist in the delegation verification equation in the form of bases. Similarly, parameter R is also used in the proxy signature verification equation in the form of bases. As a result, some other forgeries also may be possible. To avoid such attacks, verification equations would be so complicated, that no such attacks would be possible. As a modification, we will hash K with (m_w, pk_p) as hash query $H(m_w, pk_p, K)$ and R with m as hash query

$H(m, R)$ in the improved scheme. One more thing is to observe that message m is not used in the verification equation so given proxy signature is a valid proxy signature for any chosen message. We will remove also this flaw of the scheme [14] in the improved version.

5 Improved Proxy Signature Scheme

In this section, we present the improvements on provable secure proxy signature scheme without using pairings [14].

- Setup: System parameters are generated in the same manner as in scheme [14] only with a slight change in hash functions $H_1, H_2 : \{0,1\}^* \times G \to Z_n^*$ and $H_3 : \{0,1\}^* \to Z_n^*$.
- Extract: Private-public key pair are generated in the same way as in the scheme [14].
- DelGen: This algorithm takes O's secret key sk_o and a warrant m_w as inputs, and outputs the delegation $W_{O \to P}$ as follows:
  a. Generates a random $a \in Z_n^*$ and computes $K = aP$.
  b. Computes $h_1 = H_2(m_w, pk_p, K)$ and $\sigma = (h_1 sk_o + a) \bmod n$.
  O sends the delegation $W_{O \to P} = \{pk_o, m_w, K, \sigma\}$ to proxy signer P.
- DelVerif: To verify the delegation $W_{O \to P}$ and message warrant m_w, proxy signer P first computes
  $h_1 = H_2(m_w, pk_p, K)$, then checks whether
  $\sigma P = h_1 pk_o + K$.
  Accepts if it is equal, otherwise rejects.
- PKGen: If P accepts the delegation $W_{O \to P}$, he computes the proxy signing key $sk_{pr} = (\sigma h_2 + sk_p) \bmod n$, where $h_2 = H_3(m_w)$.
- PSign: Takes system parameters, the proxy signing key sk_{pr} and a message m as inputs, returns a signature of the message m. The user P does as follows.
  a. Chooses at random $b \in Z_n^*$ and computes $R = bP$.
  b. Computes $s = hb + sk_{pr} \pmod{n}$, where $h = H_1(m, R)$.
  The resulting signature is $(pk_o, pk_p, m_w, K, m, R, s)$.
- PSVerif: To check whether the signature $(pk_o, pk_p, m_w, K, m, R, s)$ is a valid proxy signature on message m under warrant m_w, verifier V first checks if the proxy signer and the message conform to m_w and computes $h_1 = H_2(m_w, pk_p, K), h_2 = H_3(m_w), h = H_1(m, R)$ then verify whether the following equation holds.
  $sP = hR + [(h_1 pk_o + K)h_2 + pk_p]$.
  If the equality holds, Verifier V accepts the signature, otherwise rejects it.

Correctness:
Since $R = bP$, $s = (hb + sk_{pr}) \bmod n$, we have

$$\begin{aligned} sP &= (hb + sk_{pr})P \\ &= hR + [(\sigma P)h_2 + sk_p P] \\ &= hR + [(K + h_1 pk_o)h_2 + pk_p]. \end{aligned}$$

5.1 Security Analysis

We analyze the security of our scheme as follows.

Distinguishability. The proposed proxy signature $(pk_o, pk_p, m_w, K, m, R, s)$ contains the warrant m_w while the normal signature does not, so both are different in the form. Also in the verification equation, public keys pk_o, pk_p and warrant m_w are used. So anyone can distinguish the proxy signature from normal signature easily.

Verifiability. The verifier of proxy signature can check easily that the verification equation $sP = hR + [(h_1 pk_o + K)h_2 + pk_p]$ holds. In addition, this equation involves original signer's public key pk_o and warrant m_w, so anyone can be convinced of the original signer's agreement on the proxy signer.

Unforgeability. In our scheme only the designated proxy signer can create a valid proxy signature, since proxy private key $sk_{pr} = (\sigma h_2 + sk_p) \bmod n$ includes the private key sk_p of proxy signer and to compute sk_p is equivalent to solve ECDLP.

Nonrepudiation. As in the verification equation warrant m_w and public keys pk_o, pk_p are used. Also generation of proxy signature needs original and proxy signer's private key sk_o, sk_p respectively. It is already proved that neither the original signer nor the proxy signer can sign in place of other party. So the original signer can not deny his delegation and proxy signer can not deny having signed the message m on behalf of original signer to other party.

Identifiability. In the proposed scheme, it can be checked who is original signer and who is proxy signer from warrant m_w. Also seeing from the verification equation $sP = hR + [(h_1 pk_o + K)h_2 + pk_p] \bmod n$, the public keys pk_o, pk_p are asymmetrical in position. So anyone can distinguish the identity of proxy signer from proxy signature.

Prevention of Misuse. Original signer generates the delegation (pk_o, m_w, K, σ) using its private key and sends to P. So the delegation can not be modified or forged. Also it is not possible for proxy signer P to transfer his proxy power to other party unless he provides proxy private key sk_p. In addition, warrant m_w contains the limit of delegated signing capability. So it is not possible to sign the messages that have not been authorized by original signer.

6 Efficiency Comparison

Here, we compare the efficiency of our scheme with similar signature scheme [15] and show that our scheme is more efficient in computational and timing (total operation time) sense than existing scheme. We compare the total number of bilinear pairings, map-to-point hash functions (H), pairing-based scalar multiplications, elliptic curve-based scalar multiplications and consequently the total operation time in overall signature process. We also note that the operation time for one pairing computation is 20.04 milliseconds, one map-to-point hash function is 3.04 milliseconds, one pairing-based scalar multiplication 6.38 milliseconds and one ECC-based scalar multiplication 2.21 milliseconds [8]. In the following tables, we have omitted the operation time due to a general hash function, as it takes ≤ 0.001 milliseconds [8]. For the computation of operation time, we refer [8] where the operation time for various cryptographic operations have been obtained using MIRACAL [13], a standard cryptographic library, and the hardware platform is a PIV 3 GHZ processor with 512 M bytes memory and the Windows XP operating system. For the pairing-based scheme, to achieve the 1,024-bit RSA level security, Tate pairing defined over the supersingular elliptic curve $E = F_p : y^2 = x^3 + x$ with embedding degree 2 has been used, where q is a 160-bit Solinas prime $q = 2^{159} + 2^{17} + 1$ and p a 512-bit prime satisfying $p + 1 = 12qr$. For the ECC-based schemes, to achieve the same security level, the parameter secp160r1 [12], recommended by the Certicom Corporation has been employed, where $p = 2^{160} - 2^{31} - 1$.

Table 1. Computational Cost Comparison

Scheme	Extract	DelGen	DelVerif	PKgen
Scheme [15]	$1M_P$	$1M_P + 1H_M$	$1H_M + 2O_P$	$1M_p$
Our scheme	$1M_E$	$1M_E$	$2M_E$	$0M_E$

Scheme	PSign	PSVerif	Total
Scheme [15]	$3M_P$	$1M_p + 1H_M + 3O_P$	$7M_P + 3H_M + 5O_P$
Our scheme	$1M_E$	$3M_E$	$8M_E$

Table 2. Running Time Comparison(in ms)

Scheme	Extract	DelGen	DelVerif	PKGen	PSign	PSVerif	Total
Scheme [15]	6.38	9.42	43.12	6.38	19.14	69.54	153.98
Our scheme	2.21	2.21	4.41	≈ 0	2.21	6.63	17.68

According to these running time computations, the running time of our proxy signature algorithm is 11.54% of scheme [15]'s algorithm and total running time of our scheme is 11.48% of the scheme [15].

If we use the running time computation results obtained by Cao and Kou [2] in different environment then efficiency of our scheme can be improved as given in the following table.

Table 3. Running Time Comparison(in ms)

Scheme	Extract	DelGen	DelVerif	PKgen	PSign	PSVerif	Total
Scheme [15]	6.38	9.42	43.12	6.38	19.14	69.54	153.98
Our scheme	0.83	0.83	1.66	≈ 0	0.83	2.49	6.64

According to these running time computations, the running time of our proxy signature algorithm is 4.33% of scheme [15]'s algorithm and total running time of our scheme is 4.31% of the scheme [15].

7 Conclusion

In this paper, we demonstrated that previously proposed scheme [14] has some security flaws. Furthermore, we presented an improved proxy signature scheme without pairing which removes these flaws. Our improved scheme is as efficient as [14].

References

1. Chen, L., Cheng, Z., Smart, N.: Identity-based key agreement protocols from pairings. Int. J. Inf. Secur. (6), 213–241 (2007)
2. Cao, X., Kou, W.: A Pairing-free Identity-based Authenticated Key Agreement Protocol with Minimal Message Exchanges. Information Sciences (2010), doi:10.1016/j.ins.2010.04.002
3. David, P., Jacque, S.: Security arguments for digital signatures and blind signatures. J. Cryptol. 13(3), 361–396 (2000)
4. Goldwasser, S., Micali, S., Rivest, R.: A digital signature scheme secure against adaptive chosenmessage attacks. SIAM J. Comput. 17(2), 281–308 (1988)
5. Granger, R., Page, D., Smart, N.P.: High security pairing-based cryptography revisited. In: Algorit. Numb. Theo. Sympo. VII, pp. 480–494 (2006)
6. Gu, C., Zhu, Y.: Provable security of ID-based proxy signature schemes. In: Lu, X., Zhao, W. (eds.) ICCNMC 2005. LNCS, vol. 3619, pp. 1277–1286. Springer, Heidelberg (2005)
7. Gu, C., Zhu, Y.: An efficient ID-based proxy signature scheme from pairings. In: Pei, D., Yung, M., Lin, D., Wu, C. (eds.) Inscrypt 2007. LNCS, vol. 4990, pp. 40–50. Springer, Heidelberg (2008)
8. He, D., Chen, J., Hu, J.: An ID-Based proxy signature schemes without bilinear pairings. Anna Telicom (2011), doi:10.1007/s12243-011-0244-0
9. Ji, H., Han, W., Zhao, L., et al.: An identity-based proxy signature from bilinear pairings. In: 2009 WASE International Conference on Information Engineering, pp. 14–17 (2009)
10. Kim, H., Baek, J., Lee, B., Kim, K.: Secret computation with secrets for mobile agent using one-time proxy signature. In: Cryptog. and Infor. Secur., Canada, pp. 307–312 (2001)
11. Mambo, M., Usuda, K., Okamoto, E.: Proxy signatures: Delegation of the power to sign message. IEICE Transactions Fundamentals E79-A(9), 1338–1353 (1996)
12. The Certicom Corporation, SEC 2:Recommended Elliptic Curve Domain Parameters, `http://www.secg.org/collateral/sec2_final.pdf`
13. Shamus Software Ltd., Miracl library, `http://www.shamus.ie/index.php?page=home`
14. Tiwari, N., Padhye, S.: Provable secure proxy signature scheme without bilinear pairings. Int. J. Commun. Syst. (2011), doi:10.1002/dac.1367
15. Wang, A., Li, J., Wang, Z.: A provably secure proxy signature scheme from bilinear pairings. J. Electro. (china)1 27(3) (2010)
16. Wu, W., Mu, Y., Susilo, W., Seberry, J., Huang, X.: Identity-based proxy signature from pairings. In: Xiao, B., Yang, L.T., Ma, J., Muller-Schloer, C., Hua, Y. (eds.) ATC 2007. LNCS, vol. 4610, pp. 22–31. Springer, Heidelberg (2007)
17. Zhang, J., Zou, W.: Another ID-based proxy signature scheme and its extension. Wuhan Univ. J. Nat. Sci. 12, 133–136 (2007)

Enhanced Block Playfair Cipher

Arvind Kumar, Pawan Singh Mehra, Gagan Gupta, and Manika Sharma

Galgotias College of Engineering and Technology,
Greater Noida
{arvinddagur,pawansinghmehra,gagan03011987,
msmanisharma22}@gmail.com

Abstract. In this paper we will enhance the traditional Blick Playfair Cipher by encrypting the plaintext in blocks. For each block the keyword would be the same but the matrix will shift by some random value. As a result of which the diagram analysis would be very difficult which is done in the traditional Playfair Cipher to obtain the plaintext from the cipher text. The shift value will be generated using random algorithm which is very secure. Playfair Cipher method, based on polyalphabetic cipher is relatively easy to break because it still leaves much of the structure and a few hundred of letters of cipher text are sufficient. To add to its security and to make it more usable we are using 6x6 matrix instead of 5x5 which will be able to cover 26 alphabets in English and ten numerals i.e. from 0 to 9. This 6x6 matrix eliminate the case of putting of 2 alphabets (I and J) together in the matrix as it was in the 5x5 matrix. In this approach plaintext as well as key can be numeral, alphabetic or combination of both and a random number will shift the matrix every time.

Keywords: Playfair Cipher, Random number, Random algorithm, Polyalphabetic cipher.

1 Introduction

Monoalphabetic substitution ciphers are easy to break because they reflect the frequency data of the original alphabet. That is they easily reflect the statistical structure of the plaintext in the cipher text, since a particular alphabet in the plaintext is always replaced by another alphabet in the cipher text as in the case of Caesar Cipher. That is why they are prone to cryptanalysis where a cryptanalyst exploits the regularities of the language which is actually the letter frequency of the English alphabets .One principle method that is used in substitution ciphers to lessen the extent to which the structure of the plaintext survives in the cipher text is to encrypt the multiple letters of the plaintext in cipher text. The best-known multiple-letter encryption cipher is the Playfair, which treats diagrams in the plaintext as single units and translates these units into cipher text diagrams. The Playfair algorithm is based on the use of a 5 * 5 matrix of letters constructed using a keyword. The Playfair cipher is a great advance over simple monoalphabetic ciphers. For one thing, whereas there are only 26 letters, there are 26 * 26 = 676 diagrams, so that identification of individual diagrams is more difficult. Furthermore, the relative frequencies of individual letters

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 689–695, 2013.

exhibit a much greater range than that of diagrams, making frequency analysis much more difficult. For these reasons, the Playfair Cipher was for a long time considered unbreakable. Despite this level of confidence in its security, the Playfair cipher is relatively easy to break because it still leaves much of the structure of the plaintext language intact. In this case we use diagram frequencies of English letters. Other drawback of the traditional playfair cipher is that the letter I and J are considered as same in the plaintext, so overhead is created in distinguishing I and J during decryption. In modified version of the playfair cipher we have used 6x6 matrix that is capable of encrypting 26 alphabets and 10 digits(10 + 26=36).It has also addressed the problem of encrypting I and J as different alphabets as shown in matrix 1.

Matrix 1: 6 X 6 matrix of alphabets and digits

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	0	1	2	3
4	5	6	7	8	9

Also what makes it more secure is that we are encrypting the plaintext in blocks. For each block the matrix is shifted by some random value, due to which the corresponding positions of the letters are changed, which adds to lot of confusion in the mind of cryptanalyst and makes it more secure towards attacks in which the attacker tries to exploit the statistical structure of the plaintext revealed in the cipher text. This block version conceals the statistical structure to a great extend and makes it secure against cryptanalysis. To add to its security we are using random algorithm for generating random numbers that will shift the matrix for each block. The rest of the paper is organized as follows: Section II demonstrates the working of traditional Playfair Cipher and its related approaches. Section III illustrates enhanced Block Playfair approach. Section IV is concerned with the analysis of the proposed algorithm and Section V summarizes the paper.

2 Related Work

The Playfair Cipher shows a great improvement over the monoalphabetic ciphers. The identification of diagrams is more difficult than individual letters. In the monoalphabetic cipher, the attacker searches in 26 letters only. But by using the Playfair Cipher, the attacker has to search in 26 * 26 =676 diagrams [1][2][3]. The relative frequencies of individual letters exhibit a much greater range than that of diagrams, making frequency analysis much more difficult. The Playfair algorithm is based on the use of a 5 * 5 matrix of letters constructed using a keyword. In this case, the keyword is *PLAYFAIR*. The matrix is constructed by filling in the letters of the keyword (minus duplicates) from left to right and from top to bottom, and then filling in the remainder of the matrix with the remaining letters in alphabetic order as shown in the matrix 2. The letters I and J count as one letter [4][8][9].

Matrix 2 : 5x5 matrix for traditional Playfair

P	L	A	Y	F
I/J	R	B	C	D
E	G	H	K	M
N	O	Q	S	T
U	V	W	X	Z

Plaintext is encrypted two letters at a time, according to the following rules [1][2]:

1. Plaintext letters that are in the same pair are separated with a filler letter, the filler letter is taken as a character say x.
   For example, calling becomes calxling.
2. Two plaintext letters that fall in the same row of the matrix are each replaced by the letter to the right, with the first element of the row circularly following the last. For example, yf is encrypted as FP.
3. Two plaintext letters that fall in the same column are each replaced by the letter beneath, with the top element of the column circularly following the last. For example, ab is encrypted as BH.
4. Otherwise, each plaintext letter in a pair is the column occupied by the other plaintext letter. Thus, ao becomes LQ and bt becomes DQ.

In [3] the traditional Playfair Cipher has been combined with random number generator method. One of the simplest methods of random number generation called linear feedback shift register is used. Mapping is being done with random numbers to secret key of Playfair cipher method and corresponding numbers are transmitted to the recipient instead of alphabetical letter [5][6][7].

3 Enhanced Block Playfair Approach

In the conventional Playfair Cipher the same matrix is used to encrypt the entire plaintext, as a result of which it is vulnerable to cryptanalysis because it leaves the statistical traces of the plaintext in the cipher text. In our modified version we are encrypting the plaintext in blocks .For each block the keyword would be the same but the matrix will shift by some value as shown in the figure. As a result of which the diagram analysis would be very difficult which is done in the traditional Playfair Cipher to obtain the plaintext from the cipher text. The shift values will be random values, these random values can be generated through various ways, most common among them is random algorithm. It is also hard to predict the random numbers if we have the previous one. Through this technique each time we will encrypt the same plaintext the output will be different. Suppose we have a plaintext which is divided into two blocks, so we need two random numbers. Each random number will be used to shift the matrix after the previous block has been encrypted, as a result of which the corresponding position of the letters to each other will be changed for the next block and the statistical traces will be concealed in the cipher text. Also the block size can be chosen according to the requirement. If we want more security against cryptanalysis attacks, then the block size can be as small as 32 characters. In this case more overheads will be created for shifting the matrix. We can use 64 characters as a standard size of blocks which can improve performance.

In enhanced Playfair plaintext is encrypted two letters at a time, according to the following rules:

1. Plaintext letters that are in the same pair are separated with a filler letter, the filler letter is taken as a character say *.
   For example, calling becomes cal*ling.
2. Two plaintext letters that fall in the same row of the matrix are each replaced by the letter that lies in the same row by a key k as a random number, let here k=2. For example, sa is encrypted as G3.
3. Two plaintext letters that fall in the same column are each replaced by the letter In the same column by a key k as a random number, let here k=2, with the top element of the column circularly following the last. For example, m1 is encrypted as H4.
4. Otherwise, each plaintext letter in a pair is the column occupied by the other plaintext letter. Thus, ao becomes MR and bt becomes DQ.

Suppose there is a plaintext which is divided into two blocks. So two random numbers are generated. Let the key be *MESSAGE345* . The resulting matrix is shown in matrix 3.

Matrix 3 : 6x6 matrix(keyword :*MESSAGE345*)

M	E	S	A	G	3
4	5	B	C	D	F
H	I	J	K	L	N
O	P	Q	R	T	U
V	W	X	Y	Z	0
1	2	6	7	8	9

Let the first random number be 5, so the shifted matrix is matrix 1 shown in matrix 3.1

It is a left shift in which the topmost left character is shifted to the bottommost right cell in the matrix. Matrix 1 will be used to encrypt the first block of the plaintext. Thereafter the second random number is used to shift the matrix 1 to produce matrix 2.

Let the second random number be 2. Matrix 2 is shown below in matrix 3.2:

Matrix 3.1: Shifted matrix 1

3	4	5	B	C	D
F	H	I	J	K	L
N	O	P	Q	R	T
U	V	W	X	Y	Z
0	1	2	6	7	8
9	M	E	S	A	G

Matrix 3.2: Shifted matrix 2

5	B	C	D	F	H
I	J	K	L	N	O
P	Q	R	T	U	V
W	X	Y	Z	0	1
2	6	7	8	9	M
E	S	A	G	3	4

Given below are two message blocks of 16 characters each that are to be encrypted:

i) abcde12345fghijk
ii) pqrst6789zuvwxy1

Now by using primary matrix shown in matrix 4, the above two messages are encrypted as:

i) SCF4M29EBCD3JKLN
ii) RTQAQ89180O0YZV7

Matrix 4: Primary matrix

M	E	S	A	G	3
4	5	B	C	D	F
H	I	J	K	L	N
O	P	Q	R	T	U
V	W	X	Y	Z	0
1	2	6	7	8	9

Now by using Matrix 1 shown below in matrix 5.1, the above first message is encrypted as:

G3FHM205BAD3JCCQ

Now by using Matrix 2 shown in matrix 5.2, the above second message is encrypted as:

RTQAQ801GUWXYZV7

Matrix 5.1 Matrix 1

A	G	3	4	5	B
C	D	F	H	J	J
K	L	N	O	P	Q
R	T	U	V	W	X
Y	Z	0	1	2	6
7	8	9	M	E	S

Matrix 5.2 : Matrix 2

3	4	5	B	C	D
F	H	I	J	K	L
N	O	P	Q	R	T
U	V	W	X	Y	Z
0	1	2	6	7	8
9	M	E	S	A	G

As shown in the above example, we have divided 32 characters plaintext into two blocks of 16 characters each and bold letters in the plain text shows that part which is same in both the blocks. When only primary matrix is used to encrypt both the texts then it generates same encrypted character for the same pair of plaintext characters.

Let the random numbers generated are 3 and 2 using random algorithm. The primary matrix is left shifted by 3 to get Matrix1. Now Matrix1 is used to encrypt the first block of 16 characters. Matrix1 is left shifted by 2 units to get the Matrix2 which is used to encrypt the remaining 16 characters. By shifting the matrix, we are able to change the corresponding position of the characters (alphabets and numbers). Due to change in the matrix, the encrypted texts are different in both the blocks. This technique disguises the attacker and makes it highly secure against any form of cryptanalysis attack.

4 Analysis of Proposed Method

This proposed methodology increases the security of the data as compare to traditional Block Playfair approach. Cryptanalysis of this proposed method is very tedious and difficult. The random algorithm is used to generate random numbers and used to shift the matrix each time. The random numbers can be generated by a third party and provided to the communicating parties at the time of encryption and decryption at the sending and the receiving end. The proposed approach is also capable to encrypt alphabets as well as digits.

5 Conclusion

Enhanced Block Playfair Cipher proposed in this paper is highly secure as compare to traditional approaches. This approach is highly secure because of the random number generated for shifting matrix in each round. As it encrypts the plaintext in blocks uses different matrix for consecutive blocks, it is quite impossible for the cryptanalyst to generate the plaintext from the cipher text. With each random shift the relative position of characters in the matrix change, which completely conceals the statistical structure of the plaintext which is the most peculiar property of this approach. As a result of these security features, this algorithm can be extensively used for sending and receiving messages with its confidentiality intact.

References

1. Stallings, W.: Cryptography and Network Security Principles and Practice, 2nd edn. Pearson Education
2. Srivastava, S.S., Gupta, N.: Security Aspects of the Extended Playfair Cipher. In: 2011 International Conference on Communication Systems and Network Technologies (CSNT), pp. 144–147. IEEE Conferences (2011)
3. Murali, P., Senthilkumar, G.: Modified Version of Playfair Cipher Using Linear Feedback Shift Register. In: International Conference on Information Management and Engineering, ICIME 2009, pp. 488–490. IEEE Conference (2009)
4. Knudsen, J.B.: Java Cryptography, 1st edn. (May 1998) ISBN
5. Srivastava, S.S., Gupta, N.: A Novel Approach to Security using Extended Playfair Cipher. International Journal of Computer Applications (0975 – 8887), vol. 20(6) (April 2011)
6. Buchmann, J.A.: Introduction to Cryptography, 2nd edn. Springer, NY (2001)
7. Patel, D.R.: Information Security Theory and Practice, 1st edn. Prentice-Hall of India Private Limited (2008)
8. Anne-Canteaut: ”Ongoing Research Area in Symmentic Cryptography” ENCRYPT (2006)
9. Schnier, B.: Applied cryptography:protocols, algorithms and source code in C. John Wiley and sons, New York (1996)

Ciphertext-Policy Attribute-Based Encryption with User Revocation Support

A. Balu[1] and K. Kuppusamy[2]

[1] Department of Computer Science and Engg.,
Alagappa University, Karaikudi
[2] Department of Computer Science and Engg.,
Alagappa University, Karaikudi
balusuriya@yahoo.co.in, kkdiksamy@yahoo.com

Abstract. In Ciphertext-Policy Attribute-Based Encryption(CP-ABE) schemes, the encryptor may set the policy in such a way that who can decrypt the encrypted message. The policy may be formed with the help of attributes. Recent CP-ABE schemes are constructed based on Linear Secret Sharing Scheme. In this paper, we use the Linear Integer Secret Sharing Scheme (LISS) for the construction. Lewko et al.[7] proposed a direction revocation method, based on that we then present a construction of CP-ABE scheme with the ability to do the direct revocation of user. The proposed construction is selectively secure under Decision Bilinear Diffe-Hellman assumption.

Keywords: Attribute-Based Encryption, Linear Integer Secret Sharing, Revocation.

1 Introduction

Attribute-Based Encryption(ABE) has a significant advantage over the traditional PKC primitives as it achieves flexible one-to-many encryption instead of one-to-one. ABE is envisioned as an important tool for addressing the problem of secure and fine-grained data sharing and access control. In an ABE system, a user is identified by a set of attributes. In their seminal paper [10] use biometric measurements as attributes in the following way. A secret key based on a set of attributes ω, can decrypt a ciphertext encrypted with a public key based on a set of attributes $\omega^{'}$,only if the sets ω and $\omega^{'}$ overlap sufficiently as determined by a threshold value t. There are two variants of ABE: Key-Policy Based ABE (KP-ABE) and Ciphertext Policy Based ABE(CP-ABE)[4]. In KP-ABE, the ciphertext is associated with a set of attributes and the secret key is associated with the access policy. The encryptor defines the set of descriptive attributes necessary to decrypt the ciphertext. The trusted authority who generates the user's secret key defines the combination of attributes for which the secret key can be used. In CP-ABE, the idea is reversed: now the ciphertext is associated with the access policy and the encrypting party determines the policy under which the data can be decrypted, while the secret key is associated with a set of attributes.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 696–705, 2013.

1.1 Motivation

Up to date, in most of CP-ABE schemes, the secret exponent s is shared by Linear Secret sharing scheme. Waters [3] proposed three CP-ABE schemes, which are based on Linear Secret Sharing Scheme(LSSS). In 2006, Damgard et al. [6] introduced the notion of Linear Integer Secret Sharing (LISS) scheme. In LISS, the access structure is expressed using AND, OR operators. The following is the advantages of LISS over LSSS.

1. The computations in LISS are done directly over the Integer, while LSSS is done over a finite field.
2. In LISS, the secret sharing cost is less.

In LISS, it is possible to represent the threshold access policy. Using these advantages, it is very easy to express the access policy effectively and share the secret exponent s efficiently in our CP-ABE construction.

1.2 Our Contribution

We present a new scheme for constructing a CP-ABE based on Linear Integer Secret Sharing Scheme (LISS), which allows to do the direct revocation. In this scheme, the access policy $\mathscr{P}$will be expressed using AND, OR operators. We assign a unique id for each user. A ciphertext will be encrypted such that a certain set S = $\{Id_1, ..., Id_r\}$ will be revoked from decrypting it. If the user's private key satisfies the access structure $\mathscr{P}$ and the $ID \notin S$ then the algorithm will decrypt the ciphertext and return the original message. We prove that the proposed scheme is selectively secure under the decisional bilinear Diffie-Hellman assumption.

1.3 Related Work

CP-ABE. The first ciphertext policy ABE was proposed by Bethencourt et al. [4] uses a threshold secret sharing to enforce the policy in the encryption phase. This method requires polynomial interpolation to reconstruct the secret and secure in the generic group model. The CP-ABE proposed by Cheung and Newport [5], in which decryption policies are restricted to a single AND gate, but attributes are allowed to be either positive or negative. In this method, the size of the ciphertext and secret key increases linearly with the total number of attributes in the system. Water's [3] presented three constructions, which are based Linear Secret Sharing Scheme (LSSS) and secure under various difficulty assumptions. Ciphertext and public parameters size were increased in the DBDH assumption [3]. The scheme in [8] is fully secure in the standard model.

Direct Revocation scheme. Direct revocation enforces revocation directly by the sender who specifies these revocation lists while encrypting. An advantage of the direct method is that it does not require key the update phase for all non revoked users interacting with the key authority. Ostrovsky et al. [9] showed a connection between revocation schemes and achieving non-monotonic access formulas in ABE, to negate an attribute in an access formula one applies a revocation scheme using the attribute as an identity to be revoked. Lewko et al. [7] proposed the direct revocation method based on a new "two equation " technique for revoking users. Attrapadung et al. [2] proposed Broadcast ABE for Key-Policy and Ciphertext-Policy to support direct revocation mechanism. Attrapadung et al. [1] proposed another construction, which supports direct and indirect revocation in a single scheme.

2 Preliminaries

2.1 Access Structures

Definition 1. (Access structure) *Let $\{1, 2, ...n\}$ be a set of parties. A collection $\Gamma \subseteq 2^{\{1,2,...,n\}}$ is monotone if $\forall B, C$: if $B \in \Gamma$ and $B \subseteq C$ then $C \in \Gamma$. An access structure(respectively, monotone access structure) is a collection (respectively, monotone collection) A of non-empty subsets of $\{1, 2..., n\}$ i.e $\Gamma \subseteq 2^{\{1,2,..,n\}} \setminus \phi$. The sets in Γ are called the authorized sets, and the sets not in Γ are called the unauthorized sets.*

2.2 Linear Integer Secret Sharing

In the LISS scheme, the secret is an integer chosen from a (publically known) interval, and each share is computed as an integer linear combination of the secret and some random numbers chosen by the dealer. Reconstruction of the secret is done by computing a linear combination with integer coefficients of the shares in a qualified set. Let $P = \{1, 2, ..n\}$ denote the n share holders and D the dealer. Let Γ be a monotone access structure on P. Let ℓ be an integer constant. The dealer D wants to share a secret s from the publically known interval $\left[-2^{\ell}, 2^{\ell}\right]$ to the shareholders P over Γ, such that every set of shareholders $A \in \Gamma$ can reconstruct s, but a set of shareholders $A \notin \Gamma$ gets no or little information on s.

We say that a subset $A \subseteq P$ is qualified if the parties in A jointly are allowed to reconstruct the secret s. In a LISS scheme, the shares consist of a collection of integers,$\{s_i\}_{i \in I}$, where for each $i \in I$, the integer s_i belongs to exactly one party and s_i is computed by a linear integer combination of s and some randomness chosen by the dealer. Given a qualified subset of shares $\{s_i\}_{i \in I'}$, then the secret can be reconstructed by a linear combination $s = \sum\limits_{i \in I'} \lambda_i s_i$, where $\{\lambda_i\}_{i \in I'}$ are integer coefficients that are determined by the index $I^{'}$. We use a distribution matrix $M \in Z^{d \times e}$ and a corresponding surjective function $\Psi : \{1, ..d\} \to P$. We say that the i-th row is labeled by $\Psi(i)$ or owned by party $P_{\Psi(i)}$. We use a distribution vector $\rho = (s, \rho_2, ..., \rho_e)$ where s is the secret, and the $\rho_i^{'} s$ are uniformly random chosen integers in $\left[-2^{\ell_0+k}, 2^{\ell_0+k}\right]$, where k is the security parameter and ℓ_0 is a constant. The dealer D calculates shares by

$$M.\rho = (s_1, ..., s_d)^T \qquad (1)$$

where we denote each s_i as a share unit for $1 \leq i \leq d$. The i'th share unit is then given to the $\Psi(i)'th$ shareholder. If $A \subseteq P$ is a set of shareholders, then M_A denotes the restriction of M rows jointly owned by A.

2.3 Integer Span Program

Definition 2. *$\mathscr{M} = (M, \Psi, \xi)$ is called an Integer Span Program(ISP) if $M \in Z^{d \times e}$ and the d rows of M are labeled by a surjective function $\Psi : \{1, ..d\} \to P$. Finally $\xi = (1, 0, 0, ...0)^T \in Z^e$ is called the target vector. We define size($\mathscr{M}$) = d, where d is the number of rows of M.*

Definition 3. *Let Γ be a monotone access and let$\mathscr{M} = (M, \Psi, \xi)$ be a integer span program. Then $\mathscr{M}$ is an ISP, if for all $A \subseteq \{1, .., n\}$ the following holds.*
1. If $A \in \Gamma$, then there is a vector $\lambda \in Z^d$ such that $M_A^T \lambda = \xi$.
2. If $A \notin \Gamma$, then there exists $\mathbf{k} = (k_1, ..., k_e)^T \in Z^e$ such that $M_A \cdot \mathbf{k} = \mathbf{0} \in Z^d$ with $k_1 = 1$, which is called the sweeping vector for A.

If we have an ISP $\mathscr{M} = (M, \Psi, \xi)$ which computes Γ, we build a LISS scheme for Γ as follows: We use M as the distribution matrix and
$\ell_0 = \ell + \lceil \log_2(k_{max}(e-1)) \rceil + 1$, where ℓ is the length of the secret and
$k_{max} = max\{|a|/a$ is an entry in some sweeping vector $\}$.

Secret Reconstruction. An authorized set A can computes the secret by taking a linear combination of their values, since there exists $\lambda_A \in Z^{d_A}$ such that $M_A^T \cdot \lambda_A = \xi$ (as per Definition 5).
With the secret shares of s_A it is justified to reconstruct the secret s by the following way

$$s_A^T \cdot \lambda_A = (M_A \cdot \rho)^T \cdot \lambda_A (\text{by eqn } (1))$$

$$= \rho^T \cdot (M_A^T \cdot \lambda_A) = \rho^T \cdot \xi = s$$

2.4 Bilinear Maps

Let G_0, G_1, be two multiplicative cyclic groups of prime order p. Let g be a generator of G_0 . Let e be a bilinear map, $e : G_0 \times G_0 \to G_1$. The bilinear map e has the following properties:

1. Bilinearity: for all $u, v \in G_0$, and $a, b \in \mathbb{Z}_p^*$,
we have $e\left(u^a, v^b\right) = e\left(u, v\right)^{ab}$

2. Non-degeneracy: $\hat{e}\left(g, g\right) \neq 1$. -
The map e is symmetric since $e\left(g^a, g^b\right) = e\left(g, g\right)^{ab} = e\left(g^b, g^a\right)$

2.5 Decisional Bilinear Diffie- Hellman Assumption

We define the Decisional Bilinear Diffie-Hellman problem as follows. A challenger chooses a group G_0 is of prime order p according to the security parameter. Let

$a, b, s \in \mathbb{Z}_p^*$ be chosen at random and g be a generator G_0 . The adversary given $\left(g, g^a, g^b, g^s\right)$ must distinguish a valid tuple $e\left(g, g\right)^{abs} \in G_1$ from a random element R in G_1. An algorithm A that outputs {0,1} has the advantage ϵ in solving decisional BDH in G_0 if

$$\begin{aligned} Pr\left[\mathcal{A}\left(g, g^a, g^b, g^s, D = e\left(g, g\right)^{abs}\right)\right) = 0\right] - \\ Pr\left[\mathcal{A}\left(g, g^a, g^b, g^s, D = R\right) = 0\right] \geq \epsilon \end{aligned}$$

2.6 Ciphertext Policy Attribute Based Encryption with User Revocation(CP-ABE-UR)

A cipher text policy attribute based encryption scheme consists of four fundamental algorithms: Setup, Key Generation, Encryption and Decryption.

Setup. The setup algorithm takes no input other than the implicit security parameter. It outputs the public parameters PK and a master key MK.

KeyGen (MK,PK, L, ID). The key generation algorithm takes as input the master key MK, public key PK, an identity ID and the attribute list L . It outputs a private key SK_L for the attribute list L .

Encrypt (S, PK, $\mathscr{P}$, m). The encryption algorithm takes as input a revocation set S of identities, public parameters PK, the message m, and an access policy $\mathscr{P}$ over the universe of attributes. The algorithm will encrypt m and produce a ciphertext CT such that any user with a key for an identity $ID \notin S$ and the attribute list L satisfies the access policy can decrypt.

Decrypt(CT,SK_L, ID,S). The decryption algorithm takes as input the ciphertext CT that was generated for the revoked set S, as well as an identity and a private key SK_L for the attribute list L. If the list L of attributes satisfies the access policy $\mathscr{P}$ and the $ID \notin S$ then the algorithm will decrypt the ciphertext and return a message M.

2.7 Security Model for CP-ABE-UR

The selective security notion for CP-ABE-UR is defined in the following game.

Init. The adversary chooses the target set S^* and the challenge access policy $\mathscr{P}^*$, gives it to the challenger.

Setup. The challenger runs the Setup algorithm and gives the public parameters, PK to the adversary.

Phase1. The adversary makes a secret key request to the Keygen oracle for any attribute list L , user index ID, with the restriction that $ID \notin S^*, L \nvDash \tau^*$. The Challenger returns Keygen(L, MK, ID, PK).

Challenge. The adversary submits two equal length messages M_0 and M_1. The Challenger flips a random coin d, and encrypts M_d under $(\mathscr{P}^*, S^*)$. The ciphertext CT^* is given to the adversary.

Phase 2. The adversary can continue querying Keygen with the same restriction as during Phase1.

Guess. The adversary outputs a guess $d^{'}$ of d.

Definition 4. *A ciphertext-policy attribute based encryption scheme is said to be secure against chosen-plaintext attack(CPA) if any polynomial time adversaries have only a negligible advantage in the IND-CPA game, where the advantage is defined to be* $\epsilon = \left|Pr[d^{'} = d] - \frac{1}{2}\right|$.

3 Main Construction

In this section, first we specify the method to form the access policy matrix M and then the construction of the CP-ABE-UR scheme.

3.1 Formation of Access Policy Matrix M

The access policy specified by the encryptor can be represented by a LISS matrix M by the following procedure.

Let $M_u \in Z^{1\times 1}$ be the matrix with single entry which is one, i.e $M_u = 1$.

If we have a matrix $M_a \in Z^{d_a \times e_a}$ then we can form $c_a \in Z^e$ to represent the first column in M_a and $R_a \in Z^{(d_a-1)\times e_a}$ to represent all but the first column in M_a.

Given any access policy $\mathscr{P}$, we can construct the distribution matrix M by using the following rules.

Rule 1. Each variable a_i in the access policy $\mathscr{P}$can be expressed by M_u.

Rule 2. For any OR-term $\mathscr{P} = \mathscr{P}_a \bigvee \mathscr{P}_b$. Let $M_a \in Z^{d_a \times e_a}$ and $M_b \in Z^{d_b \times e_b}$ be the matrices which expresses the formulas $\mathscr{P}_a$ and $\mathscr{P}_b$ respectively. We can construct a matrix $M_{OR} \in Z^{(d_a+d_b)(e_a+e_b-1)}$ expressing $\mathscr{P}$, which is defined by letting the first column of M_{OR} be the concatenation of the two column vectors c_a and c_b, then letting the following $d_a - 1$ columns be the columns of R_a expanded with e_b succeeding zero entries, and the last $d_b - 1$ columns be the columns of R_b expanded with e_a leading zero entries. This is visualized by

$$M_{OR} = \begin{array}{|c|r|r|}\hline c_a & R_a & 0 \\ \hline c_b & 0 & R_b \\ \hline\end{array}$$

Rule 3. For any AND-term $\mathscr{P} = \mathscr{P}_a \bigwedge \mathscr{P}_b$. Let $M_a \in Z^{d_a \times e_a}$ and $M_b \in Z^{d_b \times e_b}$ be the matrices which expresses the formulas $\mathscr{P}_a and \mathscr{P}_b$ respectively. We can construct a matrix $M_{AND} \in Z^{(d_a+d_b)(e_a+e_b)}$ which expresses the access policy $\mathscr{P}$. It is defined by letting the first column of M_{AND} be the column vector c_a expanded with e_b succeeding zero entries, the next column to be the concatenation of c_a and c_b the following $d_a - 1$ columns be the columns of R_a expanded with e_b succeeding zero entries, and the last $d_b - 1$ columns be the columns of R_b expanded with e_a leading zero entries. This is visualized by

$$M_{AND} = \begin{array}{|l|l|l|l|}\hline c_a & c_a & R_a & \quad 0 \\ \hline 0 & c_b & \quad 0 & R_b \\ \hline \end{array}$$

3.2 CP-ABE-UR Scheme

Setup ($\mathbf{1^k}$). The setup algorithm chooses a group G_0 of prime order p and a generator g.
Let $A = \{a_1, a_2, ..., a_n\}$ be the set of attributes.
For each attribute a_i, it chooses random element $t_i \in Z_p$, and computes $T_i = g^{t_i}\{1 \leq i \leq n\}$
Let $y = e(g, g)^\alpha$ where $\alpha \in Z_p$. Let b be a random element $\in Z_p$.
The Public Key is PK $= \big(g, g^b, y, \{T_i; 1 \leq i \leq n\}\big)$ and the Master Secret Key is MK $= (\alpha, b, t_i\{1 \leq i \leq n\})$.

KeyGen (ID,MK,PK,L). This algorithm takes as input the user index ID, master secret key, public key and the attribute list of the user and performs the following:

a) Select random values $a, r, \omega \in Z_p$
$d_0 = g^\alpha g^{ar} g^{b\omega}$; $d_2 = g^\omega$; $d_3 = (g^{bID} g)^\omega$
For each attribute in the attribute list L, $d_i^* = g^{art_i^{-1}}$
The secret key is $SK_L = (d_0, d_2, d_3, \forall a_i \in L : d_i^*)$

Encrypt(S,PK, $\mathscr{P}$, m). The encryption algorithm takes as input a user index set $S = \{ID_1, .., ID_r\}$, the public key PK , a message $m \in G_1$ to encrypt and the access policy $\mathscr{P}$.

Step 1. Select a random element $s \in \left[-2^\ell, 2^\ell\right]$ and compute $C_0 = g^s$.
M is the distribution matrix constructed by the above method for the access policy $\mathscr{P}$. Choose $\rho = (s, \rho_2, ..., \rho_e)^T$, where $\rho_i' s$ are uniformly random chosen integers in $\left[-2^{\ell_0+k}, 2^{\ell_0+k}\right]$.

Step 2.

a) Computes $M \cdot \rho = (s_1, ..., s_d)^T$

b) $C_1 = m \cdot y^s = m \cdot e(g,g)^{\alpha s}$; $C_k^{'} = g^{s_k}$; $C_k^{+} = (g^{bID_k} g)^{s_k}$;k= 1 to r

c) For each attribute in $\mathscr{P}$, compute $C_i^{*} = T_i^{s_i}$ using the corresponding shares of the attribute a_i.

The ciphertext is published as $CT = \left(C_0, C_1, C_i^{*}, C_k^{'}, C_k^{+}\right)$.

Decrypt(CT,$\boldsymbol{SK_L}$,ID,S). The decryption algorithm takes as input the ciphertext CT that was generated for the revoked set S, as well as an identity and a private key SK_L for the attribute list L. If the list L of attributes satisfies the access policy $\mathscr{P}$ and the $ID \notin S$ then there is a vector $\lambda_L \in Z^{d_L}$ such that $M_L^T \lambda_L = \xi$ (as per definition 3). With this, it is possible to reconstruct the secret using $\sum\limits_{i \in L} \lambda_i s_i = s$.

The decryption algorithm computes E= $\frac{e(C_0,d_0)}{\prod\limits_{i\in L} e(C_i^{*},(d_i^{*})^{\lambda_i})} \prod\limits_{k=1}^{r} \left[\frac{e(d_2,C_k^{+})}{e(d_3,C_k^{'})}\right]^{\frac{\lambda_k}{ID-ID_k}}$

$$= e(g,g)^{\alpha s}$$

where it can compute since $ID \neq ID_k$ for k = 1,..,r , then it computes

$$m = C_1/E.$$

3.3 Security Analysis

Theorem 1. Suppose the decisional BDH assumption holds, then no polynomial adversary can selectively break our system.

Proof: Suppose we have an adversary $\mathcal{A}$ with non-negligible advantage ϵ in the selective security game against our construction. We show how to use the adversary $\mathcal{A}$ to build a simulator $\mathcal{B}$ that is able to solve the DBDH assumption. The Challenger gives the simulator $\mathcal{B}$ the DBDH challenge : $(g,A,B,C,D) = \left(g, g^a, g^b, g^s, D\right)$.

Init. The adversary chooses the challenge access policy $(M^{'}, \mathscr{P}^{*})$, a revocation set $S^{*} = \{Id_1, Id_2, .., Id_r\}$ and gives it to the simulator.

Setup. The simulator selects at random $a^{'} \in Z_p$ and implicitly sets $\alpha = ab + a^{'}$ by letting $e(g,g)^{\alpha} = e(g^a, g^b) e(g,g)^{a^{'}}$. For all $a_j \in U$ it chooses a random $q_j \in Z_p$ and set $T_j = g^{\frac{1}{\left(M_{i,j}^{'} q_j\right)}}$ if $a_j \notin \mathscr{P}^{*}$, otherwise $T_j = g^{q_j}$. The simulator $\mathcal{B}$ sends the public parameters to $\mathcal{A}$.

Phase 1. In this phase the simulator answers private key queries. Suppose the simulator is given a private key for a list L where L does not satisfy $\mathscr{P}^{*}$ and $ID \notin S^{*}$. On each request $\mathcal{B}$ chooses a random variable $v, \delta \in Z_p$, and finds

a vector $\mathbf{k} = (k_1, k_2, .., k_e)^T \in Z^e$ such that $M^{'} \cdot \mathbf{k} = \mathbf{0}$ with $k_1 = 1$. By the definition of Sweeping vector such a vector must exist. The simulator sets $r = v - k_j b$ and computes
$d_0 = g^{\alpha} g^{ar} g^{b\omega} = g^{ab+a^{'}} g^{a(v-k_j b)} g^{b\delta} = g^{a^{'}} A^v g^{b\delta}$
In calculating d_i^* we have the term $M^{'}_{i,j} a \cdot k_j b$ get canceled because of $M^{'} \cdot \mathbf{k} = \mathbf{0}$
$d_i^* = g^{(v-k_j b) a q_j M^{'}_{i,j}} = A^{v M^{'}_{i,j} q_j}$
$d_2 = g^{\delta}$ $d_3 = (g^{bID} g)^{\delta}$

Challenge. $\mathcal{A}$ submits two messages $m_0, m_1 \in G_1$. The simulator flips a fair binary coin d, and returns the encryption of m_d. The encryption of m_d can be done as follows:
$C_0 = g^s, C_1 = m_d De(g^s, g^{a^{'}})$
The simulator will choose uniformly random integers $z_2, ..., z_h \in \left[-2^{\ell_0+k}, 2^{\ell_0+k}\right]$ and share the secret $x$using the vector $\Phi = (x, z_2, ..., z_h)$.
Create the distribution matrix M, for the access policy $\mathscr{P}^*$. Compute $M \cdot \Phi$ and use the shares to encrypt the access policy with corresponding q_j for the attributes present in the access policy $\mathscr{P}^*$, $C_i^* = T_i^{x_i}$, $C_i^{'} = g^{x_i}, C_i^{+} = (g^{bID_i} g)^{x_i}$

Phase 2. Same as Phase 1.

Guess. $\mathcal{A}$ outputs a guess $d^{'}$of d. The simulator then outputs 0 to the guesses that $D = e(g, g)^{abs}$ if $d^{'} = d$; otherwise, it outputs 1 to indicate that it believes D is random group element in G_1.

When D is a tuple the simulator $\mathcal{B}$ gives a perfect simulation, so we have that $Pr\left[\mathcal{B}\left(\rho, D = e(g, g)^{abs}\right) = 0\right] = \frac{1}{2} + \epsilon$.

When D is a random group element the message m_d is completely hidden from the adversary, and we have $Pr\left[\mathcal{B}\left(\rho, D = R\right) = 0\right] = \frac{1}{2}$.

4 Conclusion

We constructed the CP-ABE scheme based on a new secret sharing scheme called Linear Integer Secret Sharing Scheme. Our scheme has the ability to do the direct revocation of users. The security of our scheme is provided in the standard model under DBDH assumption.

References

1. Attrapadung, N., Imai, H.: Attribute-Based Encryption Supporting Direct/Indirect Revocation Modes. In: Parker, M.G. (ed.) Cryptography and Coding 2009. LNCS, vol. 5921, pp. 278–300. Springer, Heidelberg (2009)
2. Attrapadung, N., Imai, H.: Conjunctive broadcast and attribute-based encryption. In: Shacham, H., Waters, B. (eds.) Pairing 2009. LNCS, vol. 5671, pp. 248–265. Springer, Heidelberg (2009)
3. Waters, B.: Ciphertext policy attribute based encryption: An expressive, efficient, and provably secure realization. In: Cryptology eprint report 2008/290 (2008)

4. Bethencourt, J., Sahai, A., Waters, B.: Ciphertext policy attribute based encryption. In: IEEE Symposium on Security and Privacy, pp, pp. 321–334 (2007)
5. Cheung, L., Newport, C.: Provably secure Ciphertext police ABE. In: CCS 2007: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 456–465. ACM Press, New York (2007)
6. Damgård, I.B., Thorbek, R.: Linear Integer Secret Sharing and Distributed Exponentiation. In: Yung, M., Dodis, Y., Kiayias, A., Malkin, T. (eds.) PKC 2006. LNCS, vol. 3958, pp. 75–90. Springer, Heidelberg (2006)
7. Lewko, A., Sahai, A., Waters, B.: Revocation systems with very small private keys. Cryptology eprint report 2009/309 (2009)
8. Lewko, A., Okamoto, T., Sahai, A., Takashima, K., Waters, B.: Fully secure functional encryption: Attribute-based encryption and (Hierarchical) inner product encryption. In: Gilbert, H. (ed.) EUROCRYPT 2010. LNCS, vol. 6110, pp. 62–91. Springer, Heidelberg (2010)
9. Ostrovsky, R., Sahai, A., Waters, B.: Attribute-based encryption with non-monotonic access structures. In: ACM Conference on Computer and Communications Security, pp. 195–203 (2007)
10. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 457–473. Springer, Heidelberg (2005)

A Proposal for SMS Security Using NTRU Cryptosystem

Ashok Kumar Nanda and Lalit Kumar Awasthi

Computer Science & Engineering Department,
National Institute of Technology, Hamirpur, Himachal Pradesh, India - 177005
ashokkumarnanda@yahoo.com, lalit@nith.ac.in

Abstract. Short Message Service (SMS) is getting more popular now-a-days. It will play a very important role in the future business areas of mobile commerce (M-Commerce). Presently many business organizations are using SMS for their business purposes. SMS's security has become a major concern for both business organizations and customers. There is a need for an end to end SMS Encryption in order to provide a secure medium for communication. Security is main concern for any business company such as banks who will provide these mobile banking services. Till now there is no such scheme that provides complete SMSs security. The transmission of an SMS in GSM network is not secure at all. Therefore it is desirable to provide SMS security for business purposes. In this paper, we have analyzed Number Theory Research Unit (NTRU) Crypto algorithm and NTRUSign (NTRU Signature) algorithm. We have compared theoretically the performance metrics like key size, key generation time, encryption time, decryption time, CPU computational power, speed, efficiency, memory space and security strength between NTRU and RSA. This theoretical results encouraged us to simulate the NTRU cryptosystem and NTRUSign algorithm on mobile phones using full size of SMS as future work.

Keywords: M-Commerce, NTRU Cryptosystem, NTRUSign, Performance analysis, RSA, SMS Encryption, SMS Security.

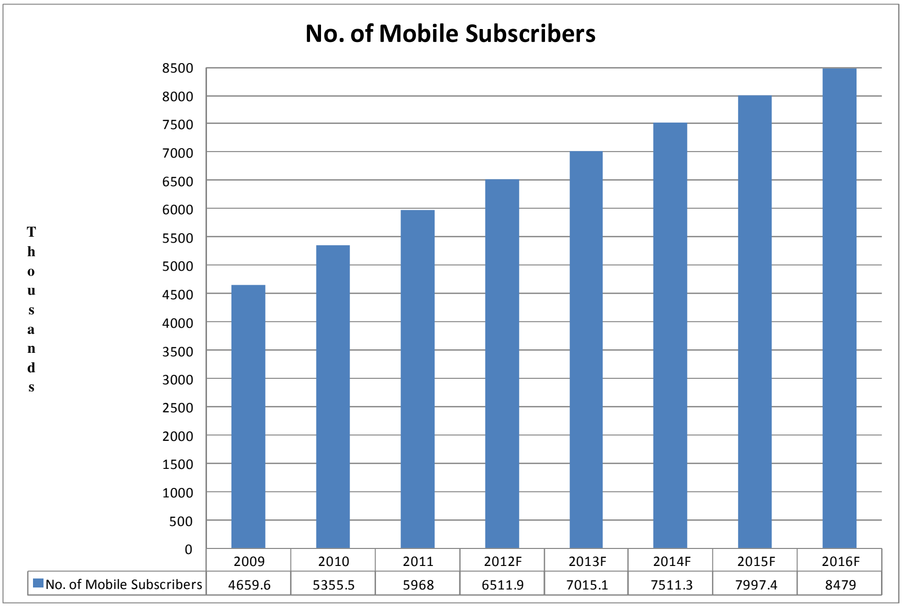
1 Introduction

In earlier times mobile phones used to be a craze, symbol of money and success but nowadays every common people finds its necessity of their individual life. Mobile phones are having a great influence in every one's live and are very convenient to keep with us. Mobile phones are a faster and more effective way to transfer information. Indeed, it is a resource that gives its user's great advantages. These days mobile phones are not just used for phone calls but they are about messaging, videos, songs, games, alarm clock, notes, calendar, reminder, etc. So one equipment, lots' of uses! Day by day mobile subscribers are increasing. The Fig. 1 shows the growth of mobile subscribers of world during 2009 to 2016F. 'F' stands for forecast value.

In 1992, the first SMS technology enables the sending and receiving of messages between two mobile phones. SMS message contains at most 140 bytes (1120 bits) of

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 706–718, 2013.

data, so one SMS message can contain up to 160 characters (if 7-bit character encoding is used) and 70 characters (if 16-bit Unicode UCS2 character encoding is used). SMS provides more convenient for mobile phone users to communicate with each other using text messages via mobile phones either from mobile phones or Internet connected computers. One major advantage of SMS is that it is supported by 100% GSM mobile phones. Almost all subscription plans provided by wireless carriers include inexpensive SMS messaging service. The mobile messaging market is growing rapidly and is a very profitable business for mobile operators. It can be seen from Fig. 2 that the growth rate of SMS in world during 2000 to 2016F. 'F' stands for forecast value.

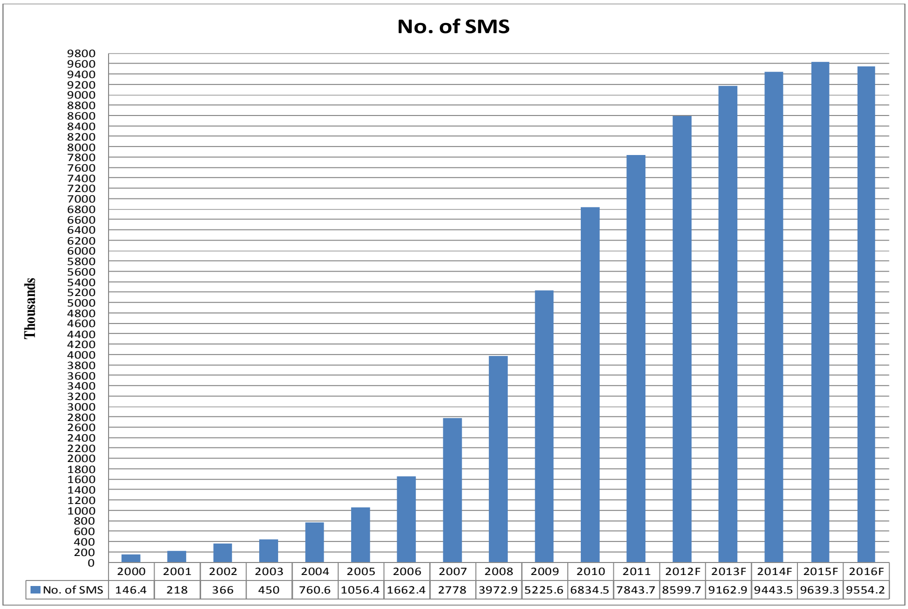


	2009	2010	2011	2012F	2013F	2014F	2015F	2016F
■ No. of Mobile Subscribers	4659.6	5355.5	5968	6511.9	7015.1	7511.3	7997.4	8479

Fig. 1. Growth of Mobile Subscribers – World from 2009 to 2016F (F stands for Forecast) Source: Portio Research Ltd

SMS is getting more popular now-a-days. It will play a very important role in the future business areas of mobile commerce (M-Commerce). Many financial and business organizations are using SMS more for their business purposes. SMS has a variety of advantages and disadvantages for M-Commerce purpose. SMS's security has become a major concern for both business organizations and customers. There is a need fast key generation, encryption and decryption with optimization of memory size, CPU energy consumption. Security is main concern for any business company such as banks who will provide these mobile banking services to their customer. Currently there is no such scheme that provides complete SMSs security.

Many people of United States, EU5 (UK, Germany, France, Spain and Italy) and Japan prefer information exchange as text message (SMS) as compared to instant message by mobiles is shown in below mentioned Table 1. The major advantages of SMS are: i) SMS is a personal like phone call but a person can read at any time without any disturbance to the work ii) Messages are instantly recorded so that one can refer at any time iii) It is relatively less SPAM free iv) SMS is discreet in nature v) SMS bills are considered as negligible vi) SMS is more convenient for deaf and hearing-impaired people to communicate vii) SMS is a store-and-forward service viii) SMS doesn't overload the network as much as phone calls ix) It is possible to send SMS many people at a time x) easy to use xi) common messaging tool among consumers xii) works across all wireless operators xiii) no specific software required to installation. There are very few and negligible disadvantages are : i) Consumes more time to type as compared to phone call ii) No proper authentication of SMS sender iii) Length of SMS is maximum 140 - 160 characters iv) Reliability and versatility can be compromised when using SMS v) does not support sending media, including videos, pictures, melodies or animations vi) does not offer a secure environment for confidential data during transmission. The same Table - 1 indicates that few people access financial services such as bank account information and financial news or stock quotes using SMS because SMS are not fully secure in wireless environment due to its broadcast nature.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012F	2013F	2014F	2015F	2016F
■ No. of SMS	146.4	218	366	450	760.6	1056.4	1662.4	2778	3972.9	5225.6	6834.5	7843.7	8599.7	9162.9	9443.5	9639.3	9554.2

Fig. 2. Growth of SMS – World from 2000 to 2016F (F stands for Forecast) Source: Portio Research Ltd

Presently researchers proposed some security concepts regarding SMS security. Most of the proposals are software frames to be installed on mobile device and /or on the SIM cards to implement security. When SMS used for M-Commerce the following services are required [1]:

Table 1. Mobile behavior in United States, EU5 (UK, Germany, France, Spain and Italy) and Japan – October, November, December 2010 Percent of total mobile audience (Age 13+)

	US	**Europe**	**Japan**
Used Messaging			
Sent Text Message	68%	82.7%	41.6%
Instant Messaging	17.2%	14.2%	3.6%
Accessed Financial Services			
Bank Accounts	11.4%	8%	7%
Financial news or stock quotes	10.2%	8%	16.5%

Source: comScore MobiLens (Feb 2011)

a) Confidentiality: only the valid communicating users can view the SMS.
b) Integrity: the SMS can’t be tampered by the intruders. The system should be able to find out such alteration.
c) Non-repudiation: no party can deny the receiving or transmitting the data communicating between them.
d) Authentication: each party has to have the ability to authenticate the other party.
e) Authorization: it has to be ensured that, a party performing the transaction is entitled to perform that transaction or not.

We realized that security is most essential for mobile phone users and network operators to avoid different threats at different levels. The transmission of an SMS in GSM network is not secure at all. Therefore it is desirable to secure SMS for business purposes by additional encryption.

Rest of paper is organized as follows. Section 2 provides an overview of related work. Section 3 brief explaining about NTRU cryptosystems. Section 4 analysis and compares the theoretically results. Section 5 discusses about NTRUSign Scheme and followed by discussion and future work.

2 Related Work

Challa and Pradhan in 2007 [2] compared RSA and NTRU using ‘C’ language for measuring encryption, decryption speeds. They performed test for encryption and decryption using key size as 128 bits, 256 bits, 512 bits, 1Kb, 2Kb, 5Kb and 10Kb for both RSA & NTRU respectively. They used data size 22 bits and 10 bits for encryption and decryption methods respectively for RSA and 51 bits and 20 bits for encryption and decryption methods respectively for NTRU. Xiaoyu Shen and his team in 2009 [3] enhanced NTRU by changing forms of random polynomial 'f' and coefficient of polynomial integer 'p' and using low hamming weight products to improve efficiency of NTRU for mobile java systems. Their programs written in Java ME and used device emulator in the WTK (Sun Java Wireless Toolkit) 2.5.2. They considered NTRU-251 and RSA-1024 have same security level. They had compared key generation time, encryption time and decryption time between NTRU-251 and RSA-1024 using equivalent security key strength. Sameer Hasan Al-bakri and M.L. Mat Kiah in 2010 [4] used hybrid NTRU and AES-Rijndael for peer-to-peer

SMS security. They implemented in J2ME using mobile information device application (MIDlet). They performed test on Symbian OS of Nokia N70, Nokia N73, Nokia N93 and Nokia 5800 Express mobile phones with data size 1block (20Byte) with key-size of NTRU - 251.

3 NTRU Cryptosystem

The NTRU public key cryptosystem was developed in 1996 at Brown University by three mathematicians J. Hoffstein, J.Pipher and J.H. Silverman. NTRU encryption algorithm is a lattice-based alternative to RSA and ECC and is based on the shortest vector problem in a lattice. NTRU can be used in mobile devices and other mobile applications because of its features of easy generation of keys, high speed and low memory use [3]. This is based on shortest vector problem in a lattice and operations based on objects in a truncated polynomial ring $R = z[X]/(X^N - 1)$.

All polynomials in the ring have integer coefficients and degree at most N-1:

$$a = a_0 + a_1x + a_2x^2 + \dots\dots + a_{N-2}x^{N-2} + a_{N-1}x^{N-1} = \sum_{i=0}^{N-1} a_i x^i$$

it can represented as vector: $a = (a_0, a_1, a_{2,\dots\dots,} a_{N-2,} a_{N-1})$

$$b = b_0 + b_1x + b_2x^2 + \dots\dots + b_{N-2}x^{N-2} + b_{N-1}x^{N-1} = \sum_{i=0}^{N-1} b_i x^i$$

it can represented as vector: $b = (b_0, b_1, b_{2,\dots\dots,} b_{N-2,} b_{N-1})$

$$a + b = \sum_{i=0}^{N-1} a_i x^i + \sum_{i=0}^{N-1} b_i x^i = \sum_{i=0}^{N-1} (a_i + b_i) x^i$$

$$a * b = \sum_{i=0}^{N-1} a_i x^i * \sum_{i=0}^{N-1} b_i x^i = \sum_{k=0}^{N-1} [\sum_{i+j=k(\bmod N)} a_i * b_j] x^k$$

Another operation we should know is modular arithmetic in which: and $a = b(\bmod c)$ means a and b have the same reminder when they are divided by c.

When we do modular arithmetic to a polynomial in the ring with the integer modulus, it just means to divided each coefficient of the polynomial by the modulus and keep the reminders as the new coefficients.

NTRU has 3 integer parameters: *N*, *p*, *q*. *N* represents the degree of the polynomials at most N-1; *p* is smaller than *q*. *p* and *q* are small moduli used to reduce the coefficients of the polynomials. They do not have common divisor. We briefly describe the NTRU algorithm [3] as follows.

3.1 Key Generation

We have to choose two random polynomials *f* and *g* in the ring with the restriction that their coefficients are small, usually in {-1, 0, 1}. We import another symbol here: $L(d_1, d_2)$, which means a set of polynomials with d_1 coefficients are 1, d_2 coefficients are -1 and the rest are 0. Usually we choose f from $L_f(d_f, d_{f-1})$ and g from $L_g(d_g, d_{g-1})$. Then we compute f_p (the inverse of f modulo p) and f_q (the inverse of f modulo q) with the property that $f * f_p = 1$ (mod p) and $f * f_q = 1$ (mod q).

If f doesn't have these inverses, another f should be chosen. The pair of polynomials f and f_p should be kept as the private key, and the public key h can be computed by $h = p * f_q * g(\bmod q)$

Both f and f_p are used for private key and h is used for public key.

Example: The parameters (N, p, q) have the values $N = 11$, $p = 3$ and $q = 32$ and therefore the polynomials f and g are of degree at most 10. The system parameters (N, p, q) are known to everybody. The polynomials are randomly chosen, so suppose they are represented by

$$f = -1 + x + x^2 - x^4 + x^6 + x^9 - x^{10} \text{ and}$$

$$g = -1 + x^2 + x^3 + x^5 - x^8 - x^{10}$$

Using the Euclidean algorithm the inverse of f modulo p and modulo q, respectively, is computed so

$$f_p = 1 + 2x + 2x^3 + 2x^4 + x^5 + 2x^7 + x^8 + 2x^9 (\bmod 3) \text{ and}$$

$$f_q = 5 + 9x + 6x^2 + 16x^3 + 4x^4 + 15x^5 + 16x^6 + 22x^7 + 20x^8 + 18x^9 + 30x^{10} (\bmod 32)$$

Which creates the public key h computing the product $h = p * f_q * g(\bmod q)$

$$= 8 + 25x + 22x^2 + 20x^3 + 12x^4 + 24x^5 + 15x^6 + 19x^7 + 12x^8 + 19x^9 + 16x^{10} (\bmod 32)$$

3.2 Encryption

The message to be sent can be put into a form of a polynomial m ε $L_m(d_m, d_m)$ whose degree is at most N-1. Then we randomly choose a blinding polynomial r ε $L_r(d_r, d_r)$ in the ring. So the encrypted message e should be computed by $e = r * h + m(\bmod q)$

Example:

Let $m = -1 + x^3 + x^4 - x^8 + x^9 + x^{10}$ *and* $r = -1 + x^2 + x^3 + x^4 - x^5 - x^7$

$$e = r * h + m(\bmod q)$$

$$= 14 + 11x + 26x^2 + 24x^3 + 14x^4 + 16x^5 + 30x^6 + 7x^7 + 25x^8 + 6x^9 + 19x^{10} (\bmod 32)$$

3.3 Decryption

First, use a part of the private key f to compute polynomial $a = f * e(\bmod q)$, then $b = a(\bmod p)$, and then We use the other part of the private key f_p to compute polynomial $c = f_p * b(\bmod p) = -1 + x^3 + x^4 - x^8 + x^9 + x^{10} = m$.

If this procedure is successful, c will be the original message m. Actually, for appropriate parameter values, this probability is extremely high. The polynomial satisfies

$$a = f * e(\bmod q) = f * (r * h + m)(\bmod q) = f * (r * (p * f_q * g + m)(\bmod q)$$
$$= p * r * g + f * m(\bmod q) \qquad [f * f_p = 1(\bmod q)]$$

The coefficients of *r, g, f, m* and the prime *p* are all much smaller than *q*, and for appropriate parameter values, the coefficients of *a* can be ensured lie in *[-q/2, q/2*], so after reduced modulo *q*, these coefficients are not changed. Then

$$b = a(\bmod p) = (p * r * g + f * m)(\bmod p) = f * m(\bmod p)$$
$$c = f_p * b(\bmod p) = f_p * (f * m)(\bmod p) = m(\bmod p)$$
$$[f * f_p = 1(\bmod p)]$$

so polynomial c is just the original message m.

Example:

$$a = f * e(\bmod q)$$
$$= 3 - 7x - 10x^2 - 11x^3 + 10x^4 + 7x^5 + 6x^6 + 7x^7 + 5x^8 - 9x^9 - 7x^{10}(\bmod 32)$$
$$b = a(\bmod p) = -x - x^2 + x^3 + x^4 + x^5 + x^7 - x^8 - x^{10}(\bmod 3)$$
$$c = f_p * b(\bmod p) = -1 + x^3 + x^4 - x^8 + x^9 + x^{10} = m \quad \text{(Proved)}$$

The below figure 4 is representing overall analysis of NTRU algorithm for SMS using between two mobile phones.

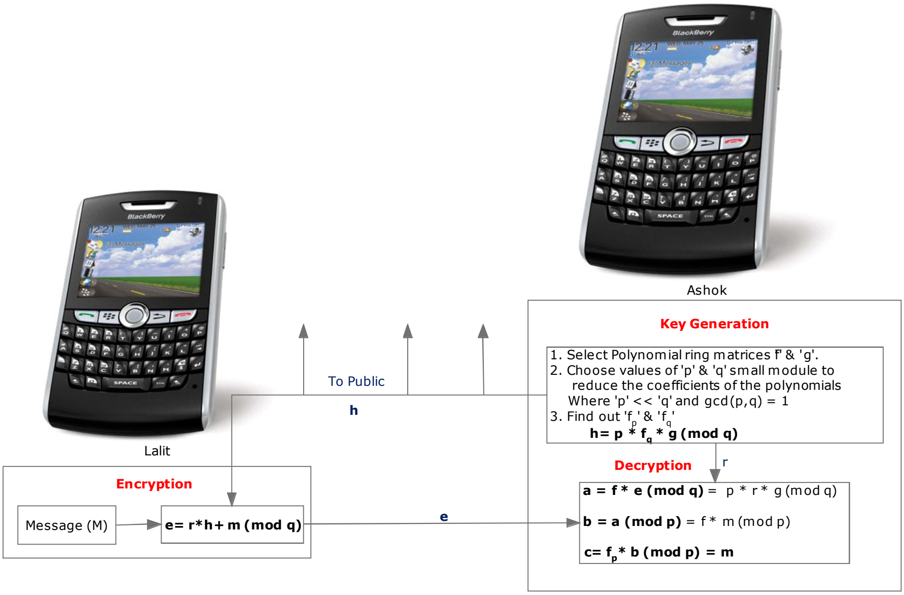

Ashok
Key Generation
1. Select Polynomial ring matrices f' & 'g'.
2. Choose values of 'p' & 'q' small module to reduce the coefficients of the polynomials Where 'p' << 'q' and gcd(p,q) = 1
3. Find out 'f_p' & 'f_q'
h= p * f_q * g (mod q)
r
Decryption
a = f * e (mod q) = p * r * g (mod q)
b = a (mod p) = f * m (mod p)
c= f_p * b (mod p) = m
To Public
h
Lalit
Encryption
Message (M)
e= r*h+ m (mod q)
e

Fig. 3. Analysis of NTRU Crypto Algorithm applied for two mobile users

4 Theoretical Result Analysis and Comparison

RSA and ECC cryptosystems are considered as the most popular traditional public cryptography algorithms. In the literature, many authors presented many weaknesses on RSA and ECC. They stated that RSA is slow [5] Hastad stated in [6] that low exponent RSA is not secure if the same message is encrypted to several receivers. In practice, RSA has proved to be quite slow. Furthermore, RSA is not well suited for limited environments like mobile phones and smart cards without RSA co-processors [7]. RSA also requires longer keys in order to be secure compared to some other cryptosystems like ECC. ECC is faster than RSA [7], ECC-160 has 6× smaller key-size than RSA-1024 and can generate a signature 12 times faster than RSA and ECC is faster, it occupies less memory space than an equivalent RSA system, ECC generates asymmetry keys pair faster than RSA, ECC is more efficient than the ubiquitous RSA based schemes because ECC utilizes smaller key sizes for equivalent security, Security wise, ECC is stronger than RSA [8]. The NTRU crypto system is a new public key cryptography approved in 2009. The table 2 gives the total comparison between NTRU and RSA.

The company www.securityinnovation.com has built a cryptographic toolkit called NERI that is based on the NTRU algorithm. It provides data that compare the performance of NTRU with that of RSA and ECC on both servers and PDAs in Table 3. It implies that NTRU to have a performance advantage that ranges from 9:1 in the case of NTRU:ECC decryption on PDA to over 333:1 in the case of NTRU:RSA decryption on PDA.

In [4], they performed test on NTRU pair keys generation only for Nokia mobile phones like Nokia N70, N73, N93 and Nokia 5800 Xpress Music using Java emulator. The table 4 has shown its comparison. From the results, they noticed that NTRU algorithm performed very well on the mobile devices and there were no negative effects on the mobile devices' performance due to the small time required for the key generation. NTRU does not require high computing power, which makes it the best alternatives for mobile devices with providing either same or more security facility. Table 4 shows the proposed public key cryptography implementation in non-server architecture based on NTRU algorithm.

NTRU cryptosystem is gaining more popularity slowly because it's key size is very small, key generation, encryption speed, decryption speed are much faster and computation power requires very less, Operation speed is very fast, more efficient, consuming less space and more suitable for mobile devices shown in table 2. It is not free (as per our knowledge). It is standardized in IEEE 1363.1-2008 and X9.98-2010. Unlike RSA and ECC, NTRU is resistant to quantum computing based on crypto attacks. It is the smallest public key crypto available on market (8 kb). Some constraints are i) no support for NTRU in the leading browsers and ii) it is necessary required to implement NTRU at both ends of the SSL tunnel. www.securityinnovation.com provides SSL libraries and software development toolkits in C/C++ and Java. Unlike RSA and ECC, no successful attack has been recorded to break the security of this algorithm [4]. From the above, we hope that NTRU crypto algorithm will be more suitable and easy to implement in mobile devices for our proposed scheme.

Table 2. comparison among ntru, and rsa cryptosystems [3, 4]

Factors	NTRU	RSA
Key size	Very small (1/4 of RSA of same size)	slow
Key generation	200 times faster than RSA	slow
Encryption/sec	1113 times faster than 2048 – RSA	slow
Decryption/sec*	1132 ms for NTRU – 251 (More than 30 times faster)	35102 ms for RSA – 1024
Computation power	Too less than compared to both in mobile and smart cards.	Much more compared to NTRU
Speed	1300 times faster than 2048 – RSA and 117 times faster than ECC NIST – 224	Quite slow
Efficiency	Fastest	slow
Applicable to mobile device	Forefront on mobile environment	Not well suited without RSA coprocessors
Memory Space	Least than RSA	More compared both ECC and NTRU
Security	Strongest	Not secure if message is encrypted to several receivers. Needs longer key size.

Table 3. RSA, ECC and NTRU performance on servers (800 MHz Pentium III) and PDAs (Palm) [4, 7 & 8]

Key Size	Server		PDA	
	Encryption (blocks/sec)	**Decryption (blocks/sec)**	**Encryption (blocks/sec)**	**Decrypt on (blocks/sec)**
1024-bit RSA	1280	110	0.5	0.036
163-bit ECC	458	702	0.4	1.3
N=251 NTRU	22727	10869	21	12

5 NTRUSign Scheme

In this section, we briefly describe NTRUSign digital signature scheme. For NTRU encryption scheme, please refer above NTRU Cryptosystems in section 3.

In some steps, NTRUSign uses the quotient ring $R_q = Z_q[x] / (x^N - 1)$, where the coefficients are reduced modulo q, where q is typically a power of 2, for example 128.

The multiplicative group of units in R_q is denoted by R_q^*. The inverse polynomial of a a ε R_q^* is denoted by a^{-1}. If a polynomial '*a*' has all coefficients chosen from the set {0, 1}, we call this a binary polynomial.

Table 4. NTRU pair keys generation operation test [4]

	Nokia N70	**Nokia N73**	**Nokia N93**	**Nokia 5800 Xpress music**
generation	2G	3G	3G	5G
Operating System	Symbian OS v8.1a	Symbian OS v9.1	Symbian OS v9.1	Symbian OS v9.4
CPU	ARM9	Dual ARM 9	Dual ARM 11	ARM 11
Clock rate	220MHz	220MHz	332 MHz	434 MHz
Internal memory	22MB	42MB	50MB	81MB
External Memory	MMC type	2 GB Mini SD	2 GB Mini SD	16GB Mini SD
pair keys generation operation	142 ms	77 ms	53 ms	29 ms

The security of NTRUSign scheme is based on the approximately closest vector problem in a certain lattice, called NTRU lattice. In this scheme, the signer can sign a message by demonstrating the ability to solve the approximately closest vector problem reasonably well for the point generated from a hashed message in a given space.

The basic idea is as follows: The signer's private key is a short basis for an NTRU lattice and his public key is a much longer basis for the same lattice. The signature on a digital document is a vector in the lattice with two properties:

- The signature is attached to the document being signed.
- The signature demonstrates an ability to solve a general closest vector problem in the lattice.

NTRUSign digital signature scheme works as follows [9]:

5.1 System Parameters

- *N*: a (prime) dimension.
- *q*: a natural number used as a modulus.
- d_f, d_g: are nonnegative integers in the interval [0, N] used as a key size parameters; .
- NormBound: a bound parameter *of verification.*

5.2 Key Generation

A signer creates his public key *h* and the corresponding private key *{(f, g), (F, G)}* as follows:

- Choose binary polynomials *f* and *g* with d_f 1's and d_g 1's, respectively.
- Compute the public key $h \equiv f^{-1} * g(\bmod q)$.
- Compute small polynomials (*F, G*) satisfying $f * G - g * F = q$.

5.3 Signing Step

A signer generates his signature *s* on the digital document *D* as follows:

- Obtain the polynomials (m_1, m_2) mod *q* for the document *D* by using the public hash function.
- Write $G * m_1 - F * m_2 = A + q * B$; and - $g * m_1 + f * m_2 = a + q * b$;

where *A* and *a* have coefficients between – q/2 and q/2.

- Compute polynomials *s* and *t* as

  $s \equiv f * B + F * b \ (mod\ q),$

  $t \equiv g * B + G * b \ (mod\ q).$

  Here, a vector $(s,t) \in L_h^{NT}$ is very close to m = (m_1,m_2).
- The polynomial *s* is the signature on the digital document *D* for the public key *h*.

5.4 Verification Step

For a given signature *s* and document *D*, a verifier should do the following:

- Hash the document *D* to recreate (m_1, m_2) mod q.
- Using the signature *s* and public key *h*, compute the corresponding polynomial $t \equiv s * h \ (mod\ q),$
- Which becomes exactly the same as the polynomial $g * B + G * b \ (mod\ q)$. (Note that (*s, t*) is a point in the NTRU lattice L_h^{NT})
- Compute the distance from (*s, t*) to (m_1, m_2) and verify that this value is smaller than the NormBound parameter. In other words, check that $\| s - m_1\|^2 + \| t - m_2 \|^2 \leq$ NormBound^2, where the norm ($\| \cdot \|$) is a centered norm.
- NTRUSign algorithm uses the centered norm concept instead of Euclidean norm in verification step to measure the size of an element a ϵ R.

6 Discussion and Future Work

Now-a-days SMS is more popular for different applications in our daily real life. Errorless data transmission with secured is important in wireless environment. In this paper we have discussed about NTRU Cryptosystem and NTRUSign Scheme. From the table 2 - 4, we concluded that NTRU cryptosystem is much faster and providing stronger security than other traditional (example RSA and ECC in both server and PDA) cryptosystems. We are expecting that it will be more efficient scheme and will provide better result for our proposed scheme to implement for SMS's security at any mobile devices. So it may improve the current security level, fastest speed and provide reliable message at receiver end with respect to key generation, encryption decryption, CPU power consumption and memory size with smaller key size.

Our future work is to implement NTRU crypto algorithm and NTRUSign scheme for any mobile device to provide security of SMS and compare it with traditional cryptosystems with respect to all performance parameters like key generation time, encryption time & decryption, CPU power consumption, memory space and security strength.

Acknowledgment. This research work is supported by Ministry of Human Resource Development (MHRD), Government of India. The authors would like to thank the anonymous reviewers for their insightful comments.

References

1. Hossain, A., Jahan, S., Hussain, M.M., Amin, M.R., Shah Newaz, S.H.: A Proposal For Enhancing The Security System of Short Message Service In GSM. In: 2nd International Conference on Anti-Counterfeiting, Security and Identification, ASID 2008 (2008), doi:10.1109/IWASID.2008.4688386
2. Challa, N., Pradhan, J.: Performance Analysis of Public key Cryptographic Systems RSA and NTRU. International Journal of Computer Science and Network Security 7(8), 87–96 (2007)
3. Shen, X., Du, Z., Chen, R.: Research on NTRU Algorithm for Mobile Java Security. In: International Conference Scalable Computing and Communications; Eighth International Conference on Embedded Computing (SCALCOM-EMBEDDEDCOM 2009), pp. 366–369 (2009)
4. Al-Bakri, S.H., Mat Kiah, M.L., Zaidan, A.A., Zaidan, B.B., Alam, G.M.: Securing peer-to-peer mobile communications using public key cryptography: New security strategy. International Journal of the Physical Sciences 6(4), 930–938 (2011)
5. RSA description and algorithm, http://www.oocities.org/hmaxf_urlcr/rsa.htm
6. Kurosawa, K., Okada, K., Tsujii, S.: Low exponent attack against elliptic curve RSA, http://www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.44.2453
7. Karu, P., Loikkanen, J.: Practical Comparison of Fast Public-key Cryptosystems, http://www.tml.tkk.fi/Opinnot/Tik-110.501/2000/papers.html

8. Gupta, V., Gupta, S., Chang, S.: Performance Analysis of Elliptic Curve Cryptography for SSL, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.3368
9. Min, S., Yamamoto, G., Kim, K.: Finding Malleability in NTRUSign, http://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&sqi=2&ved=0CEwQFjAA&url=http%3A%2F%2Fwww.autoidlabs.org%2Fuploads%2Fmedia%2FAUTOIDLABS-WP-HARDWARE-033.pdf&ei=IS3aT4ZLieutB7O-kZUB&usg=AFQjCNE_5aqVHyvfFisC7GwheselLkgsfA

Analysis and Improvement of an Authentication Scheme Using Smart Cards

Sonam Devgan Kaul and Amit K. Awasthi

Department of Applied Mathematics,
Gautam Buddha University, Greater Noida,201308, UP, India
{sonamdevgan11,awasthi.amitk}@gmail.com

Abstract. In 2010, Sood et al [16] proposed a secure dynamic identity based authentication scheme using smart cards. They claimed that their scheme is secure against various attacks. In this paper, we demonstrate that their scheme is completely insecure and vulnerable to outsider attack as well as insider attack. An outsider attacker can obtain the common session key between the user and the server, while an insider attacker can get not only the session key but also the secret key of the server. Therefore, the entire system collapses. To remedy these security flaws, an improved scheme is proposed to withstand these attacks.

Keywords: cryptanalysis, authentication protocol, smart cards, dynamic identity, password.

1 Introduction

With the rapid increasing need of remote digital services and electronic transactions; authentication schemes that ensure secure communication through an insecure channel are gaining popularity and have been studied widely in recent years. In 1981, Lamport [9] proposed first remote user password based authentication scheme by employing a one way hash chain, in an insecure and untrusted network, but this scheme has a major drawback of its dependency on verification table. Smart cards implementation solved this problem of dependency on verification tables and ensures secure communication. That is why, Smart cards based authentication schemes are becoming day by day more popular. In 2001, Hwang et al [6] proposed first smart cards based authentication scheme. As Security and efficiency are the main factors for any authentication scheme from the user's perspective. In view of the fact, several smart cards based remote user authentication schemes [1,2,3,5,8,11,12,15] have been proposed.

In 2004, Das et al [4] proposed a dynamic identity based remote user authentication scheme using smart cards that preserves user's anonymity. However, their scheme is vulnerable to various attacks. In 2005, Liao et al [10] proposed an improved scheme that achieves mutual authentication. In 2006, Yoon and Yoo [17] cryptanlyse the mutual authentication of Liao et al's scheme. In the same direction in 2010, Sood et al [16] proposed an improved protocol of Liao

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 719–728, 2013.

et al's scheme and demonstrated that improved protocol is secure against various attacks like malicious user attack, impersonation attack, offline and online dictionary attack, denial of service attack and so on.

Recently, Pelaez and Novella [14] demonstrated that Sood et al's scheme is vulnerable to malicious user attack, man-in-the-middle attack, stolen smart card attack, off-line ID guessing attack, impersonation attack and server spoofing attack. In this paper, we also pointed out few weaknesses of the scheme. This paper shows that an insider attacker who has access to server can obtain the secret key of the server, which makes this scheme totally insecure. To remedy these security flaws, we proposed an upgraded authentication scheme, that preserves some properties of Sood et al's scheme, resolves all the identified weaknesses of their scheme and makes it more secure and efficient for practical applications.

The rest of the paper is organized as follows: Section 2 briefly reviews Sood et al's authentication scheme. Section 3 describes the weaknesses of Sood et al's scheme. Our proposed scheme is presented in Section 4, followed by security analysis in Section 5. Finally, we conclude the paper in Section 6.

2 Review of Sood et al's Scheme

In this section, we examine the dynamic identity based authentication scheme proposed by Sood et al in 2010. This scheme consists of four phases: registration phase, login phase, verification and session key agreement phase and password changing phase. The notations used throughout the paper are summarized in table 1.

Table 1. Notations and Symbols used in paper

U_i	Legitimate i^{th} user
ID_i	Identifier of U_i
PW_i	Password of U_i
S	The Server
x	Secret key of the server S
y_i	Server's random value
sk_i	Session Key
T	Current date and time of input device
T'	Current date and time of the server S
δT	Expected time interval for a transmission delay
$H(.)$	Secure one way Hash Function
$\oplus$	Bitwise Exclusively or (XOR) operation
$\|\|$	Bitwise concatenation operation

2.1 Registration Phase

To register itself to the server S, the user U_i chooses his identity ID_i and password PW_i and sends it to server S via a secure communication channel. Then the server S chooses random value y_i for i^{th} user and computes:

$N_i = H(PW_i) \oplus H(y_i \| ID_i) \oplus H(x)$,
$B_i = y_i \oplus H(PW_i)$,
$V_i = H(ID_i \| PW_i) \oplus PW_i$,
$D_i = H(y_i \| ID_i)$

The server S stores $y_i \oplus x$ and $ID_i \oplus H(x)$ corresponding to D_i in its database. Then S stores $(N_i, B_i, V_i, H(.))$ into smart card and sends it to U_i via a secure communication channel.

2.2 Login Phase

When the user U_i wants to login, he simply inserts the smart card into the card reader and keys in ID_i^* and PW_i^*. The smart card computes
$V_i^* = H(ID_i^* \| PW_i^*) \oplus PW_i^*$
and verifies it with the stored V_i. After verifying the legality of the user, the smart card computes:
$y_i = B_i \oplus H(PW_i)$,
$H(x) = N_i \oplus H(PW_i) \oplus H(y_i \| ID_i)$,
$CID_i = H(y_i \| ID_i) \oplus H(H(x) \| T)$,
$M_i = H(H(x) \| H(y_i) \| T)$
and sends the login request message (CID_i, M_i, T) to the server S.

2.3 Verification and Session Key Agreement Phase

Upon receiving the login request, S first check the validity of time stamp T by checking $(T' - T) \leq \delta T$ to accept/reject the login request. If login request is accepted, the server S computes $D_i^* = CID_i \oplus H(H(x) \| T)$ and extract $y_i \oplus x$ and $ID_i \oplus H(x)$ corresponding to D_i^* from its database to obtain y_i and ID_i. Then the server computes $M_i^* = H(H(x) \| H(y_i) \| T)$ and verifies it with the received M_i . If it finds true, then U_i is authenticated. Finally, S and U_i computes common session key $sk_i = H(ID_i \| y_i \| H(x) \| T)$ for further communication.

2.4 Password Change Phase

Whenever U_i wants to update his password, he inserts his smart card into card reader and presents the credentials such as identifier ID_i and password PW_i. After verifying the legality of the user by verifying V_i, the smart card ask U_i to input the new password PW_i^{new} to replace the value of N_i, B_i and V_i with the N_i^{new}, B_i^{new} and V_i^{new} where $N_i^{new} = N_i \oplus H(PW_i) \oplus H(PW_i^{new})$, $B_i^{new} = B_i \oplus H(PW_i) \oplus H(PW_i^{new})$ and $V_i^{new} = H(ID_i \| PW_i^{new}) \oplus PW_i^{new}$.

3 Security Flaws in Sood et al's Scheme

Here, we consider an outside attacker is one who has no direct access to the server. An user with valid identity and password also comes in outside attacker category. On the other side, Insider attacker is one who is having administrative access of the server. It is the basic requirement of the authentication scheme that any insider can not get the secret key of the server or any attacker can not compute the common session key between the user and the server. Sood et al's scheme is highly insecure as the basic requirement is not fulfilled. In this section, we demonstrate that Sood et al's scheme is vulnerable to outsider attack and insider attack.

3.1 Outsider Attack

The secret information stored in the smart card can be extracted by some means, such as monitoring the power consumption [7] or analyzing the leaked information [13]. So, any outsider U_a, who is the legal user and owns a smart card, can get information $(N_a, B_a, V_a, H(.))$, that is stored on his smart card, where

$N_a = H(PW_a) \oplus H(y_a \| ID_a) \oplus H(x)$,

$B_a = y_a \oplus H(PW_a)$,

$V_a = H(ID_a \| PW_a) \oplus PW_a$,

then he compute: $y_a = B_a \oplus H(PW_a)$ and $H(x) = N_a \oplus H(PW_a) \oplus H(y_a \| ID_a)$. Thus, an outsider can get $H(x)$ which is same for each legal user and is very sensitive information, the hash value of secret key of the server.

An the attacker extracts security parameters $(N_i, B_i, V_i, H(.))$ from other legitimate user $U_i's$ smart card. During the login transaction between U_i and the server, the attacker intercepts login request message (CID_i, M_i, T) that the user U_i sends to the server S. The attacker uses his knowledge of $H(x)$ and computes the following

$$y_i = CID_i \oplus N_i \oplus B_i \oplus H(H(x) \| T) \oplus H(x)$$

In such a way an outsider U_a obtains $H(x)$ as well as y_i, just by using his own smart card and the legitimate user's smart card. Then, an outsider attacker (the user U_a) can easily compute the session key for the transmission between server and the user U_i, as,

$$sk_i = H(ID_i \| y_i \| H(x) \| T)$$

and thus, he can get the unauthorized access to the services provided by the server to the user U_i.

3.2 Insider Attack

The system manager or a privileged insider user, who has direct access to the server, simply apply for registration and gets a valid smart card. Now he may adopt the procedure like outsider attacker to get $H(x)$ as well as y_i. He computes $D_i = H(y_i \| ID_i)$ and extracts the information $y_i \oplus x$ and $ID_i \oplus H(x)$ from server's database. With this information he can easily compute the secret key of the server as

$$x = y_i \oplus (y_i \oplus x)$$

Thus any privileged insider user of the system, after getting the secret key of the server can purposely leak the information or impersonate the legitimate user or may modify the information. He can also issue an illegal smart card to some fake user. Thus, Sood et al's proposed scheme is insecure and vulnerable to various attacks and is not secure and efficient for practical applications.

4 Our Proposed Scheme

In this section, we propose an upgraded authentication scheme, that preserves the properties of Sood et al's scheme and resolves all the identified weaknesses of their scheme and make it secure and efficient for practical applications. The scheme consists of four phases: registration phase, login phase, verification & session key agreement phase and password changing phase.

4.1 Registration Phase

When the user U_i wants to register, he chooses his identity ID_i and password PW_i, and send it to the server S via a secure communication channel. Then, the server S chooses random value y_i for i^{th} user and computes:

$N_i = H(y_i\|PW_i) \oplus H(y_i\|ID_i) \oplus H(x)$,
$B_i = y_i \oplus H(PW_i)$,
$V_i = H(ID_i\|PW_i) \oplus PW_i$,
$D_i = H(y_i\|ID_i)$

S chooses the value of y_i in such a way that the value of D_i must be unique for each user. The server S stores $y_i \oplus H(x||ID_i)$ and $ID_i \oplus H(x)$ corresponding to D_i in its database. Then, S stores $(N_i, B_i, V_i, H(.))$ into smart card and sends it to U_i via a secure channel.

4.2 Login Phase

The user U_i inserts the smart card in to the card reader and keys in ID_i^* and PW_i^*, then the smart card computes

$V_i^* = H(ID_i^*\|PW_i^*) \oplus PW_i^*$

and checks whether computed V_i^* is equal to the stored V_i or not. If they are equal, the requested user is the legitimate bearer of the smart card otherwise rejects the login request. To resist offline password guessing attack, the card reader locks the card if U_i enters either wrong identifier or wrong password more than limited number of times. After verifying the legality of the user, the smart card computes:

$y_i = B_i \oplus H(PW_i)$,
$H(x) = N_i \oplus H(y_i\|PW_i) \oplus H(y_i\|ID_i)$,
$CID_i = H(y_i\|ID_i) \oplus H(H(x)\|T)$,
$M_i = H(H(x)\|H(y_i)\|T)$

and sends the login request message (CID_i, M_i, T) to the server S.

Table 2. Registration Phase

U_i		S
Choose ID_i		
and PW_i	$\xrightarrow[Secure]{ID_i,\ PW_i}$	
		Choose random value y_i
		$N_i = H(y_i \| PW_i) \oplus H(y_i \| ID_i) \oplus H(x)$
		$B_i = y_i \oplus H(PW_i)$
		$V_i = H(ID_i \| PW_i) \oplus PW_i$
		$D_i = H(y_i \| ID_i)$
		Store $y_i \oplus H(x \| ID_i)$ and $ID_i \oplus H(x)$ for each D_i
		Store $(N_i, B_i, V_i, H(.))$ into smart card
	$\xleftarrow[SmartCard]{(N_i,B_i,V_i,H(.))}$	

Table 3. Login Phase

U_i	**Smart card**	**S**
Input ID_i^* and PW_i^*		
	Compute $V_i^* = H(ID_i^* \| PW_i^*) \oplus PW_i^*$	
	Verifies $V_i^*\ ?=\ V_i$	
	Compute $y_i = B_i \oplus H(PW_i)$	
	$H(x) = N_i \oplus H(y_i \| PW_i) \oplus H(y_i \| ID_i)$	
	$CID_i = H(y_i \| ID_i) \oplus H(H(x) \| T)$	
	$M_i = H(H(x) \| H(y_i) \| T)$	
	$\xrightarrow{(CID_i,M_i,T)}$	

4.3 Verification and Session Key Agreement Phase

Upon receiving the login request, S first check the validity of time stamp T by checking $(T' - T) \leq \delta T$ to accept/reject the login request. If it finds incorrect, the login request is rejected else the server S computes

$$D_i^* = CID_i \oplus H(H(x) \| T)$$

and extract $y_i \oplus H(x \| ID_i)$ and $ID_i \oplus H(x)$ corresponding to D_i^* from its database and recompute ID_i and y_i using its secret information x. Then the server computes

$$M_i^* = H(H(x) \| H(y_i) \| T)$$

and verifies computed M_i^* with the received M_i. If it finds true, then U_i is authenticated and the login request is accepted else the connection is interrupted. Finally, S and U_i computes the common session key $sk_i = H(ID_i \| y_i \| H(x) \| T)$ of the transmission.

4.4 Password Change Phase

Whenever U_i wants to update his password, he inserts his smart card into the card reader and presents the credentials such as identifier ID_i and current password PW_i. After verifying the legality of the user by verifying V_i, the smart card

Table 4. Verification and Session Key Agreement Phase

U_i		**S**
	$\xrightarrow{(CID_i, M_i, T)}$	
		Verifies $(T' - T) \leq \delta T$
		Compute $D_i^* = CID_i \oplus H(H(x) \| T)$
		Extract D_i^*, $y_i \oplus H(x \| ID_i)$ and $ID_i \oplus H(x)$
		Obtain ID_i and y_i
		Compute $M_i^* = H(H(x) \| H(y_i) \| T)$
		Verifies M_i^* ? = M_i
		Session Key $sk_i = H(ID_i \| y_i \| H(x) \| T)$
Session Key		
$sk_i = H(ID_i \| y_i \| H(x) \| T)$		

ask U_i to input the new password PW_i^{new} to replace the value of N_i, B_i and V_i with the N_i^{new}, B_i^{new} and V_i^{new} where $N_i^{new} = N_i \oplus H(y_i \| PW_i) \oplus H(y_i \| PW_i^{new})$, $B_i^{new} = B_i \oplus H(PW_i) \oplus H(PW_i^{new})$ and $V_i^{new} = H(ID_i \| PW_i^{new}) \oplus PW_i^{new}$. To resist offline password guessing attack, the card reader locks the card if U_i enters either wrong identifier or wrong password more than limited number of times.

5 Security Analysis

In this section, we analyze the security of our scheme under the assumption that the secret information stored in the smart card could be extracted by some means, such as monitoring the power consumption [7] or analyzing the leaked information [13].

5.1 Denial of Service Attack

To resist password guessing attack, the card reader locks the card if someone enters either wrong identifier or wrong password more than limited number of times, So even if an adversary got the legitimate user smart card, but he is unable to create valid login request by guessing identity ID_i and password PW_i correctly at the same time. Thus, the proposed protocol is secure against denial of service attack.

5.2 Malicious User Attack

A legal but malicious user U_a can get the value of $H(x)$ from his own card, which is same for each user. But from $H(x)$, U_a may not be able to compute y_i, which makes the proposed protocol secure against malicious user attack.

5.3 Impersonation Attack

As both CID_i and M_i are protected by secure one way hash function, any modification in login request message (CID_i, M_i, T) will be detected by the server by verifying M_i. So, because the attacker has no way to find PW_i and y_i of the legitimate user U_i, he can not modify login request message, which makes this protocol secure against impersonation attack.

5.4 Offline Password Guessing Attack

After gathering the information on legitimate user U_i's smart card, an attacker can intercept the login request message (CID_i, M_i, T) during the login transaction, and try to guess out ID_i, PW_i, y_i and x, but it is not possible to guess out all the parameters correctly at the same time, which makes this protocol secure against offline guessing attacks.

5.5 Stolen Smart Card Attack

An attacker can extract security parameters $(N_i, B_i, V_i, H(.))$ from legitimate user U_i's smart card. But, this information does not help him to find out the value of server's secret x, or user's secret parameter y_i corresponding to the i^{th} legitimate user. He can not use this information to generate fake login request. He is also not able to play as man in middle by using any information on card. Thus, the proposed protocol is secure against stolen smart card attack.

5.6 Insider Attack

Any privileged insider user (a system manager as an attacker) can obtain $H(x)$ from his registered legal smart card, but without knowing the password of i^{th} user, he cannot compute y_i and secret key x of the server and can not use the secret information for personal benefit. Thus, this protocol is secure against an insider attack.

5.7 Online Password Guessing Attack

As the card reader locks the card after limited number of wrong login attempts, So it is impossible for an attacker, to pretend to be the legitimate user U_i and try to login the server by online guessing different words as identity ID_i and password PW_i of the user U_i.

5.8 Server Spoofing Attack

An adversary may not be able to masquerade server by modifying login transaction message because of verification of M_i. So, he may not be able to compute the common session key $sk_i = H(ID_i \| y_i \| H(x) \| T)$. Moreover, the session key is session variant for the same user. Thus, the proposed protocol is secure against server spoofing or masquerading attack.

5.9 Stolen Verifier Attack

If an attacker may be able to steal the verification table from the server, then he can obtain $y_i \oplus H(x \| ID_i)$ and $ID_i \oplus H(x)$ corresponding to D_i from its database but from this, he is unable to compute the secret key x and y_i of the legitimate user.

5.10 Replay Attack and Parallel Session Attack

Our proposed protocol, can withstand Replay attack and Parallel Session Attack because replaying a login request message (CID_i, M_i, T) of one session into another session is useless as the authenticity of the login request is verified by checking the freshness of the time stamp T and also by replaying a login request message within the valid time frame window, can not give an attacker, the common session key between the user U_i and the server S.

5.11 Man-in-the-Middle Attack

Even if an adversary can intercept the session, get the login request message (CID_i, M_i, T) and authenticate himself to the server S, but he can not compute the session key $sk_i = H(ID_i \| y_i \| H(x) \| T)$ between the user and the server as there are two secret parameters $H(x)$ and y_i are included in the session key. Only registered users are able to compute $H(x)$, but they can not compute y_i for any other user until they know the password of the user U_i. Thus, either the valid user who initiated the session or the server can retrieve the original message during transmission.

6 Conclusion

In this paper, we analysed Sood et al's dynamic identity based authentication scheme using smart cards and its immunity against various attacks. We got that their scheme is insecure for practical applications and vulnerable to outsider and insider attacks. To remedy these security flaws, we proposed an upgraded protocol for authentication scheme that preserves the similar properties of their scheme and resolves all the identified weaknesses of their scheme and make it more secure and efficient for practical purpose.

References

1. Awasthi, A.K.: Comment on a dynamic id-based remote user authentication scheme. Transaction on Cryptology 1(2), 15–16 (2004)
2. Chien, H.-Y., Chen, C.H.: A remote authentication scheme preserving user anonymity. Proc. Advanced Information Networking and Applications 2, 245–248 (2005)
3. Chien, H.-Y., Jan, J.-K., Tseng, Y.-M.: An efficient and practical solution to remote authentication: smart card. Computers and Security 21(4), 372–375 (2002)

4. Das, M.L., Saxena, A., Gulati, V.P.: A dynamic id-based remote user authentication scheme. IEEE Transactions on Consumer Electronics 50(2), 629–631 (2004)
5. He, D., Wu, S.: Security flaws in smart card based authentication scheme for multi server environment. Wireless Personal Communications (2012) (0929-6212)
6. Hwang, M.-S., Li, L.H.: A new remote user authentication scheme using smart cards. IEEE Transactions on Consumer Electronics 46(1), 28–30 (2000)
7. Kocher, P.C., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
8. Ku, W.-C., Chang, S.-T.: Impersonation attack on dynamic id-based remote user authentication scheme using smart cards. IEICE, Transactions on Communications E88-B(5), 2165–2167 (2005)
9. Lamport, L.: Password authentication with insecure communication. Communication of the ACM 24(11), 770–772 (1981)
10. Liao, I.-E., Lee, C.-C., Hwang, M.-S.: Security enhancement for a dynamic id-based remote user authentication scheme. Proc. Conference on Next Generation Web Services Practice, 437–440 (2005)
11. Liou, Y., Lin, J., Wang, S.: A new dynamic id-based remote user authentication scheme using smart cards. In: Proc. 16th Information Security Conference, Taiwan, pp. 198–205 (July 2006)
12. Liu, J., Zhong, S.: Analysis of kim-jeon-yoo password authentication scheme. Cryptologia 33(2), 183–187 (2009)
13. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examining smart card security under the threat of power analysis attacks. IEEE Transactions on Computers 51(5), 541–552 (2002)
14. Pelaez, R.M., Novella, F.R.: Cryptanalysis of sood et al's authentication scheme using smart cards. IACR Cryptology ePrint Archive (386) (July 2012)
15. Snih, H.C.: Cryptanalysis on two password authentication schemes. In: Labortary of Cryptography and Information Security. National Central University, Taiwan (2008)
16. Sood, S.K., Sarje, A.K., Singh, K.: An improvement of liao et al's authentication scheme using smard cards. In: Proc. IEEE 2nd International Advance Computing Conference, pp. 240–245 (February 2010)
17. Yoon, E.-J., Yoo, K.-Y.: Improving the dynamic ID-based remote mutual authentication scheme. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4277, pp. 499–507. Springer, Heidelberg (2006)

A Review on Remote User Authentication Schemes Using Smart Cards

Keerti Srivastava, Amit K. Awasthi, and R.C. Mittal

[1] Department of Applied Mathematics, Gautam Buddh University,
Greater Noida, 201310, UP, India
keert.cipet@gmail.com
[2] Department of Applied Mathematics, Gautam Buddh University,
Greater Noida, 201310, UP, India
awasthi.amitk@gmail.com
[3] Indian Institute of Technology, Roorkee, UA, India
rcmmmfma@iitr.ernet.in

Abstract. Remote user authentication is a mechanism in which the remote server verifies the legitimacy of a user over an insecure communication channel. Password based authentication schemes have been widely deployed to verify the legitimacy of remote users as password authentication is one of the simplest and the most convenient authentication mechanism over insecure networks. In remote user authentication scheme, the user is assigned a smart card, which is being personalized by some parameters and provide the legal users to use the resources of the remote system. Until now, there have been ample of remote user authentication schemes published in the literature and each published schemes have its own merits and demerits. Recently, many schemes proposed are based on the one-way hash function. The computational complexity of their schemes is superior to the discrete logarithm-problem-based schemes. In our paper, we have defined all the security requirements and the goals. An ideal password authentication schemes should satisfy and achieve all of these. We have presented the results of our survey through five of the currently available secure one way hash function based remote user authentication schemes. We hope an ideal smart card (not storing (ID_i, PW_i)), which meets all the security requirements and achieves all the goals can be developed.

1 Introduction

With large scale development of network technology, remote user authentication in e-commerce and m-commerce has become an indispensable part to access the precious resources. It provides the legal users to use the resources of the remote system. To access resources at remote systems, users should have proper access rights like in Remote Login Systems, Automated Teller Machines (ATM's), Personal Digital Assistants (PDA) and Database Management Systems, etc and to access these resources, each user should have an identity and

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 729–749, 2013.

password (ID_i, PW_i). Traditionally, the (ID_i, PW_i) are maintained by the remote system and when a user wants to login to a remote server, he simply submits his (ID_i, PW_i) to the server. On receiving the login message, the remote server compares the submitted corresponding pair (ID_i, PW_i) with stored one in password table. If match found, user will be granted to access the server resources.

Due to their efficiency and one-way property, one way hash functions have been used and as the basis on which more and more cryptosystems including password authentication systems, are being deployed. Now two factor authentication is very common in practice. In most cases, user ID with password and smart card are component of two factor authentication. In an open network environment, the remote authentication scheme using smart card is a very practical solution to validate the legitimacy of a remote user. In 2003, Wu and Chieu presented a user friendly remote authentication scheme using smart card [22], but Wang et al [33] found that Wu-Chieu scheme is vulnerable to forged login attack. He also presented an improved remote authentication scheme using smart card which eliminate this vulnerability. In 2005, D.Z. Sun found Wang et al.'s scheme was not secure under the smart card loss assumption, and proposed a new improved user friendly remote authentication scheme using smart card [7]. In continuation, many remote user authentication schemes have been proposed. These all authentication schemes having a common features in Registration phase. In each case, server issues a smart card which is storing ID and by which an intruder can impersonate a legal user by stealing the user's ID from stolen smart card. To overcome this problem, various proposal of improved remote user authentication schemes have been proposed, which can withstand the stolen smart card attack, impersonation attack by not storing the ID (or any secret information) in smart card. So that, if an intruder gets the stolen smart card, he could not get any access to the server as a legal user. Keeping in this view, an ample of authentication schemes as Sun et al' scheme [11], Chein et al [12], Ku et al [30], Yoon et al [8], X.M.Wang et al's [31] have been proposed.

In this paper we have reviewed above mentioned authentication schemes and proposed a new set of security requirements and goals for remote user authentication scheme with smart card. Every security requirement and goal is clearly defined. The separation of security requirements set and goals set allows us to establish a systematic approach for proving security of password authentication scheme with smart card (not storing the secret information). We reviewed few remote user authentication scheme using smart card to capture the work done and also to recognize the security challenges in this area. We presented our results based on the proposed security requirements, goals and communication and computation costs.

The remainder of this paper is organized as follows: In Section 2, related research work in this field is presented. Security requirements and the goals are presented in Section 3. Review of remote user authentication schemes and their cryptanalysis is presented in Section 4. Performance comparison of the schemes is given in Section 5. Finally, we conclude the paper in Section 6.

2 Related Research Work

In 1981, Lamport [16] proposed a novel password authentication scheme using cryptography hash function. However, high hash overhead and the necessity for password resetting decrease its suitability for practical use. Since then, many improved password authentication schemes [19],[18],[1],[29] have been proposed. One of the common feature of these schemes is that the server has to securely store a verification table, which contains the verifiers of user's passwords. If the verification table is stolen by an adversary, the system may be partially or totally broken. To resist such a stolen verifier attack, Hwang et al[28] in 1990, proposed a non interactive password authentication scheme and its enhanced version, which additionally uses smart cards. In Hwang et al scheme, the server requires neither storing the verifiers of user's passwords nor keeping any secret of the user. In 2000, Hwang and Li [10] proposed a verifier-free password authentication scheme using smart cards based on ElGamal's public-key technique. However, Hwang-Li's scheme does't allow users to freely choose and change their passwords. Furthermore, Hwang-Li's scheme had found to be vulnerable to impersonation attack. To improve their efficiency, Sun [11] proposed a light-weight verifier-free password authentication scheme using smart cards based on cryptographic hash functions. The major drawbacks of Sun's scheme was that the password is not easily memorizable and the user can't freely choose/change his password. Later, in 2002 H.Y.Chien pointed out that Sun's scheme [11] achieves unilateral user authentication and he also proposed a protocol to achieve mutual authentication[12]. In addition, the user can freely choose his password and the smart card not containing user's ID_i can avoid the risk of ID_i-theft or impersonation. Unfortunately, Chien'et al scheme can't withstand a parallel session attack [30],[6],[27],[26].

Further Ku's et al [30] pointed out the Chien et al [12] scheme is vulnerable to reflection attack, insider attack, guessing attack and is not repairable. However, Yoon et al [8] showed that Ku et al [30] scheme was susceptible to parallel session attack and was insecure for changing the user's password, and also proposed an enhancement to Ku et al's scheme to overcome such problems. Due to the power constraints of smart cards, the cost of implementation should be low as the lower the cost, the great chance of success in practical realization. Among those smart card based schemes, Ku et al's and Yoon et al's schemes require only several hash operations instead of the costly modular exponentiations. Therefore, their schemes exhibits great application potentiality in smart card field, regardless of their security.

In continuation process 2005, Yoon et al [9] pointed out Lee et al scheme [27],[26] is also vulnerable to some insidious attacks, reflection attack, stolen verifier attack, parallel session attack, replay attack, etc. To remedy these pitfalls, In 2007, X-M Wang pointed out that Ku et al's and Yoon et al's scheme are still vulnerable to the guessing attack, forgery attack and denial of service attack. As a result, only requiring few additional hash operations, X-M-Wang scheme can withstand the previously proposed attacks. In addition, wrong passwords input by the users can be deducted immediately and session key is also provided after authentication phase. The computational cost and efficiency of the improved scheme are encouraging for the practical implementation in the resource-constraints environment.

3 Security Requirements

In this section, we define and list out the security attacks that are required for an ideal password authentication scheme to withstand.

3.1 SR1.Denial of Service Attack

This attack rejects all or specific users by means of an offensive action on the server or by means of a falsification of user's password-verifier. In this attack, an attacker can inconvenience the user but cannot imitate the user.

3.2 SR2.Forgery Attack(Impersonation Attack)

An attacker attempts to modify intercepted communication and masquerade as the legal user so that he can access the resources of a remote system. To manipulate the sensitive data of the legal users, an attacker can also masquerade as the legal server.

3.3 SR3.Parallel Session Attack

Without knowing the user's password, an attacker by masquerade as the legal user, can create a valid login message out of some eavesdropped communication between the user and the server. The attacker may launch a parallel attack by replaying the server's response message as the user's login message at a later time.

3.4 SR4.Password Guessing Attack

Most passwords have such low entropy that it is vulnerable to password guessing attack, where an attacker intercepts authentication messages and store it locally and then attempts to use a guessed password to verify the correctness of his guess using these authentication messages.

3.5 SR5.Replay Attack

Having intercepted previous communications, an attacker can impersonate as the legal user and login to the system. The attacker can replay the intercepted messages. An attack in which a valid data transmission is maliciously or fraudulently repeated either by the originator or by an adversary who intercepts the data and retransmits it possibly as part of a masquerade attack.

3.6 SR6.Smart Card Loss Attack

When the smart card is lost or stolen then an unauthorized users can easily change the password of the smart card or can guess the password of the user using password guessing attack or can impersonate the user to login to the system.

3.7 SR7.Stolen-Verifier Attack

In most of the application, the server stores hashed passwords instead of clear text passwords. In the stolen verifier attack, an adversary who steals the password verifier(e.g.hashed password) from the server can use it directly to masquerade as a legitimate user during the user authentication phase.

3.8 SR8.Reflection Attack

A reflection attack is a method of attacking a challenge-response authentication system that uses the same protocol in both directions. That is the same challenge-response protocol is used by each side to authenticate the other side. The essential idea of the attack is to trick the target in to providing the answer to its own challenge.

3.9 SR9.Insider Attack

An insider attack is intentional misuse by individuals who are authorized to use the servers and the networks. Insider of the server can perform an off-line guessing attack to obtain password. If succeeds, the insider of the server can try to use password to impersonate users to login other servers employing normal password authentication methods.

4 Goals

An ideal password authentication scheme should withstand all of the above attacks. Besides, it should achieve the following goals:

4.1 G1.No Verification Table

The remote system should not have a dictionary of verification tables such as clear text passwords or hashed passwords to authenticate users.

4.2 G2.Freely Chosen Password by the Users

If the password is chosen by the remote server without the consent of the user, then the user has no choice to choose his own password, which is not a case in the real-life applications, e.g. email subscription and online banking, etc. Secondly, password chosen by the server could be long and random (for example, 1024 or 2048 bits), which might be difficult for a registered user to remember easily and it is most likely that user may forged this long and random password, if he is not frequently using the system. So,users should be able to choose their password freely.

4.3 G3.No Password Reveal

If the user's password is revealed to the server during registration, then it is likely that user uses the same password to login several servers for his convenience. In this case, the insider, e.g.,the administrator of the server can try to use the same password to impersonate user to login other servers that adopt normal remote user password authentication schemes. Therefore, the passwords should not be revealed by the administrator of the server.

4.4 G4.Password Dependent

The password independent scheme means that the scheme is equivalent to no password scheme, because user with any random password may access the server. Suppose an intruder theft the smart card for a short duration and makes a duplicate of it, now he has no need to crack the password because he may insert any random password, server will authenticate the intruder as a valid user. So, the authentication scheme should be password dependent.

4.5 G5.Mutual Authentication

Mutual Authentication should be provided between the user and remote system. Not only the server verify the legal users, but the users should be able to verify the legal server. Mutual Authentication can help withstand the server impersonation attack, where an attacker pretends to be the server to manipulate the sensitive data of the legal users.

4.6 G6.Session Key Agreement

A Session key should be established during the password authentication process. It is pertinent that after the successful authentication process, both parties will communicate some secret message, which should be encrypted to provide the confidentiality and secrecy of transmitted data.

4.7 G7.Forward Secrecy

Suppose the server's secret key is revealed and if the attacker tries to get passwords or other login information from the stolen smart card, he can easily impersonate the user and login to the system. Therefore, the scheme should be capable to provide forward secrecy even if the smart card is lost or stolen.

4.8 G8. User Anonymity

In some authentication scenarios, it is very important to preserve the privacy of a user because an adversary sniffing the communication channel can eavesdrops the communication parties involved in the authentication process and can easily analyze the transaction being performed by user.

4.9 G9.Smart Card Revocation

It is one of the requirements of smart card-based authentication schemes that in case of lost of cards, there should be provision in the system for invalidating the further use of lost smart card, otherwise an adversary can impersonate valid registered user.

4.10 G10.Efficiency for Wrong Password Login

Even if the user inputs wrong password by mistake in login phase, without any delay client should notify the user with error message, instead of sending the user's login request unconditionally to the server. If the client sends the information to the server for password verification, then the authentication will be delayed.

To be called an ideal scheme, a password authentication scheme should be able to withstand all of the above attacks and achieve all of the above goals. Unfortunately, none of the existing password authentication schemes can withstand all the above attacks and achieve all the goals. So, still there are opportunities to develop an ideal remote user password authentication scheme, which satisfies all security requirements and which meets all the goals.

5 Review of Five Remote User Authentication Schemes Based on Smart Cards

In this section, we review five smart card based password authentication schemes, which are based on hash function. Each password authentication scheme is composed of four phases. They are Registration phase, Login phase, Authentication phase and Password change phase. In the Registration phase, the user U registers with the remote server S and obtains a smart card through secure channel for future use. In the Login phase, When U wants to login to S for using resources of S, he inserts his smart card in to card reader and keys in his identity ID and password PW to access services. In the Authentication phase, S verifies the validity of the login request. Password change phase is invoked, whenever the user U wants to change his password. He can easily change his password without taking any assistance from the remote system. Now,we review some smart card based password authentication scheme.

The notations used throughout this paper are described as in the following.

U :	An User
(ID_u, PW_u) :	User $U's$ identifier and password respectively.
$CARD$:	$U's$ Smart Card.
S :	A Remote Server.
x :	Sserver's Secret Key.
T_U, T_S :	User's and Server's Current Time stamp respectively.
$h(.)$:	A Hash Function.
$\oplus$:	Bitwise XOR Operation.
$X \rightarrow Y\{M\}$:	X Send a message M to Y over an insecure channel.

5.1 Review of Sun's Scheme(Sun's Scheme[11])

A Registration Phase:
The user submits his ID_i to the remote system upon receiving the registration request, the remote system performs the following steps:

R1 Compute $PW_i = h(ID_i, x)$, where x be a secret key maintained by Remote system and h is a one-way function.

R2 Personalizes the smart card with the parameter $h(.)$.

R3 $S \Rightarrow U_i$:Smart card.

R4 $S \Rightarrow U_i$:PW_i through a secure channel.

B Login Phase:
The userU_i inserts his smart card to the card reader of a terminal, and keys his (ID_i, PW_i). Then Smart Card will perform the following operations:

L1 Compute $C_1 = h(T \oplus PW_i)$ where T is the current date and time of the input device.

L2 $U_i \Rightarrow S$:$C = (ID_i, C_1, T)$.

C Authentication Phase:
Upon receiving the login message $C = (ID_i, C_1, T)$ at time T' the remote system authenticates the userU_i with the following steps:

A1 Check the validity of ID_i.

A2 Verify the validity of the time interval between T and T'.if$(T' - T) \geq \Delta T$, then the remote system rejects the login request.

A3 Computes $PW_i = h(ID_i, x)$ and $C_1' = h(T \oplus PW_i)$.If $(C_1' = C_1)$ then the system accepts the login request. Otherwise, it rejects the login request.

Cryptanalysis of H.M.Sun'Scheme(Sun,2000)[11]

1. Mutual Authentication(breaks): Sun's scheme only achieves unilateral user authentication that, only authentication Server can authenticate the legitimacy of the remote user while the user cannot authenticate the legitimacy of Authentication Server. An attacker pretends to be the server to manipulate sensitive data of the legal users.
2. Replay Attack(Supports):Replay attack is not possible since uses time stamps. The idea behind the use of time stamps is to generate a synchronization mechanism between the client and the server. Replaying attacks(replaying an old (ID_i, C_1, T) in login phase)can not work because this will make A2 of Authentication phase fail.
3. Forward Secrecy(breaks):Suppose that an intruder has stolen the remote systems secret key x.It is obvious in this scheme that an intruder can easily compute each user's secret hash value as $PW_i = h(ID_i, x)$ and can impersonate any legitimate user. Therefore, in future, the server $S's$ secret key x may be changed to prevent an intruder's malicious activity. However, it would be much costs at a time and too expensive to re-compute all secret hash values at a time and communicate them to the users. Therefore, Sun et al's scheme does not guarantee a system's secret key forward secrecy.
4. Efficiency for wrong password login(breaks):If the user U_i inputs a wrong password in login phase by mistake,this wrong password will not be detected by the smart card at login phase. It is transferred unconditionally to the server. Server will check whether entered PW_i is wrong or right at step A3 of the Authentication phase. So the authentication will be delayed and inefficient.
5. Denial of Service Attack(breaks):Due to the unchangebility of $h(ID \oplus x)$ in Sun et al.'s scheme [11],a forged login request can not be prohibited even when U detected that his C_1 has been compromised. Accordingly, Ku et al. extended ID with $EID = (ID.n)$ and replace $C_1 = h(ID \oplus x)$ with $C_1 = h(EID \oplus x)$ in their improved scheme, so that C_1 can be changed by EID with different n when C_1 has been compromised. Unfortunately, the number n is stored in an entry table in server side, which is somewhat equivalent with using verification table, and suffers from the risk of modified entry table and the cost of protecting and maintaining the entry table. Once the intruder modifies n in entry table, the user's login message C_2 keeps $h(h(EID \oplus x) \oplus T_u)$ as before while the authentication message C_2 computed by system will change to $h(h(EID' \oplus x) \oplus T_u)$.
6. Impersonation attack(breaks):In Sun et al.'s[11] scheme, an adversary can obtain the corresponding password PW_i by performing a password guessing attack. The adversary intercepts the login request $C = (ID_i, C_1, T)$. First, he guesses a password PW_i^* and then computes $C_1^* = h(PW_i^* \oplus T)*$. If $C_1^* = C_1$, then the adversary has correctly guessed the password $(PW_i^* = PW_i)$. Once the adversary has correctly obtain PW_i, then he can impersonate the legal user.

7. Smart Card Lost Attack(supports):In Sun et al.'s[11] scheme only $h(.)$ is stored in smart card, which is one way function, the Smart Card Lost Attack is not possible in Sun et al.'s[11] scheme.
8. Password Guessing attack(breaks):In Sun et al.'s[11] scheme, an adversary can obtain the corresponding password PW_i by performing a password guessing attack. The adversary intercepts the login request $C = (ID_i, C_1, T)$. First, he guesses a password PW_i^* and then computes $C_1^* = h(PW_i^* \oplus T)^*$. If $(C_1^* = C_1)$, then the adversary has correctly guessed the password $(PW_i^* = PW_i)$. Once the adversary has correctly obtain PW_i, then he can impersonate the legal user.
9. Stolen Verifier attack(breaks):In the Registration phase,ID is passed to the server and PW is passed to the user. We assume that an adversary A can obtain the one way hash function $h(.)$ from stolen smart card. An adversary A can exhaustively examine all possible random number x until $PW_i = h(ID_i \oplus x)$. So, the scheme is vulnerable to stolen verifier attack.
10. Reflection Attack(supports):Reflection attack is possible only when the adversary gets both the messages(user to server and server to user). In Sun et al.'s[11],Reflection Attack is not possible since this scheme does not support mutual authentication. In this scheme, only the server authenticates the user, but the user does not authenticate the server. An adversary can get the message $C = (ID_i, C_1, T)$, which was transferred from the client to the server, but he cannot get the message, which was transferred from the server to the user. Reflection attack is possible only when the adversary gets both the messages.
11. Insider Attack(supports):The scheme is not vulnerable to insider attack. In the Registration Phase, PW_i is not in plain text form and calculated by $PW_i = h(ID_i, x)$ using secret key x which is known to server only, so any insider of S can not calculate the password $PW_i = h(ID_i, x)$.
12. Parallel Session Attack(supports):Parallel Session attack is possible only when the message structures between the user and the server are same.In Sun et al.'s[11],Parallel Session Attack is not possible since this scheme does not support mutual authentication. In this scheme, only the server authenticates the user, but the user does not authenticate the server. An adversary can get the message $C = (ID_i, C_1, T)$, which has transferred from the client to the server, but he cannot get the message which has transferred from the server to the user. .

5.2 Review of Chien et al.'s Remote User Authentication Scheme(Chien et al.'s scheme[12])

A.Registration Phase:
The user submits hisID_i and the password PW_i to the remote system through a secure channel. Upon receiving the registration request, the remote system performs the following steps:
R1.Computes a secret number $(R = h(ID_i \oplus x) \oplus PW_i)$ and checks an entry for the user U_i in his account database. Here x is a secret key of Remote Server and

$h(.)$ is a one-way function
R2.Personalizes the smart card with the parameters $h(.)$ and the secret number R
R3.$S \Rightarrow U_i$:Smart card.

B.Login Phase:
The userU_i inserts his smart card to the card reader of a terminal, and keys his ID_i, PW_i. Then Smart Card will perform the following operations:
L1.Compute $C_1 = (R \oplus PW_i)$ and $C_2 = h(C_1 \oplus T)$where T is the current date and time of the input device.
L2.$U_i \Rightarrow S$:$C = (ID_i, C_2, T)$.

C.Authentication Phase:
Upon receiving the login message $C = (ID_i, C_2, T)$ at time T' the remote system authenticates the user U_i with the following steps:
A1.Check the validity of ID_i.
A2.Verify the validity of the time interval between T and T'. If $(T' - T) \geq \Delta T$, then the remote system rejects the login request.
A3.Computes $C_2' = h(h(ID_i \oplus x) \oplus T)$. If $C_2' = C_2$, then the system accepts the login request. Otherwise, it rejects the login request.
A4.Computes $C_3 = h(h(ID_i \oplus x) \oplus T'')$ where T'' is the current time stamp.
A5.$S \Rightarrow U_i$:$D = (C_3, T'')$ for Mutual Authentication.
A6.On Receiving the message D from Remote Server, the user U_i computes $C_3' = h(C_1 \oplus T'')$.If $C_3' = C_3$ holds,the legitimacy of AS is verified.

Cryptanalysis of Chien et al.'s Remote User Authentication Scheme (Chien et al.'s scheme[12])

1. Parellel Session Attack(breaks):Consider the scenario of the parallel session attack [17] that an intruder U_a without knowing user's password wants to masquerade as a legal user U_i by creating a valid login message from the eavesdropped communication between AS and U_i when U_i wants to login the Authentication server AS, U_i sends the login message $C = (ID_i, C_2, T)$ to AS, if valid then the identification of U_i is authenticated and AS responses $D = (C_3, T'')$ to U_i. Once U_a intercepts this message, he masquerades as the legal user U_i to start a new session with AS by sending $C^* = (ID_i, C_2^*, T'')$ back to AS, where $(C_2^* = C_3)$. The login message $C^* = (ID_i, C_2^*, T'')$ will pass the user authentication of Chien et al.'s scheme[12] due to the fact that $C_2^* = C_3 = h(h(ID_i \oplus x) \oplus T'')$. Finally, AS responses the message (T''', C_3^*) to U_i, where$C_3^* = h(C_1' \oplus T''')$ and T''' is the current timestamp. The intruder intercepts and drops this message.
2. Reflection Attack(breaks):A malicious user intercepts the login request $C = (ID_i, C_2, T)$ and replaces the pair (C_3, T'') with (C_2, T) in verification phase. When the user U receives the pair (C_2, T), he verifies $C_2 = h(C_1 \oplus T)$, which holds truly. In this way, a malicious user reflects AS and U will be

fooled. Thus, Chien et al.'s scheme fails to provide mutual authentication and vulnerable to the reflection attack.

3. Impersonation attack(breaks):In Chein et al.'s scheme an adversary can obtain the corresponding password PW_i by performing a password guessing attack. The adversary intercepts the login request $C = (ID_i, C_2, T)$. First, he guesses a password PW_i^* and then computes $C_1^* = (R \oplus PW_i^*) = h(ID_i \oplus x)^*$ and $C_2^* = h(C_1^* \oplus T)$.if $(C_2^* = C_2)$, then the adversary has correctly guessed the password $(PW_i^* = PW_i)$ and $(C_1^* = C_1)$. Once the adversary has correctly obtain C_1, then he can impersonate the legal user.
4. Reparability of password(breaks):Since the password PW_i is the function of the identity ID_i of the user and the secret key x of AS. Therefore, to change the password PW_i for U_i, AS has to change ID_i or x. However, since x is commonly used for all users rather than specifically used for only U_i. It is not reasonable and efficient to change the secret key x for the security of a single user. Additionally, it is also impractical to change identity of the user. Thus, they claimed that the Chien et al.'s scheme[12] is non reparable.

5. Insider Attack(breaks):The password of the user U_i will be reveal to AS in the registration phase. If user U_i uses the same password to access other servers for convenience, the insider of AS can impersonate the user U_i to access other services.
6. Replay Attack(supports):Replay attack is not possible since this scheme uses time stamps. The idea behind the use of time stamps is to generate a synchronization mechanism between the client and the server. Neither the replay of an old login message $C = (ID_i, C_2, T)$ in the login phase nor the replay of the remote server's response message $D = (C_3, T'')$ in the verification phase.
7. No Password Reveal(supports):Since the user's password is not revealed to the server during registration, therefore impersonation attack by Authentication server is not possible in Chein et al.'s scheme.
8. Mutual Authentication(supports):The Chein et al.'s scheme can mutually authenticate each other between user and server by A1,A2,A3,A6 step in Authentication Phase.
9. Efficiency for wrong password login(breaks):If user U inputs a wrong password by mistake, this wrong password will not be detected by the client, instead it transfers login information unconditionally to the server. Even though server checks for valid login at step A2 of Authentication Phase.
10. Denial of Service Attack(breaks):Due to the unchangebility of $h(ID, x)$ in Chien et al.'s scheme [12], a forged login request can not be prohibited even when U detected that his C_1 has been compromised. Accordingly, Ku et al. extended ID with $EID = (ID.n)$ and replace $C_1 = h(ID \oplus x)$ with $C_1 = h(EID \oplus x)$ in their improved scheme, so that C_1 can be changed by EID with different n when C_1 has been compromised. Unfortunately, the number n is stored in an entry table in server side, which is somewhat equivalent with using verification table, and suffers from the risk of modified entry table and the cost of protecting and maintaining the entry table. Once the intruder modifies n in entry table, the user's login message C_2 keeps

$h(h(EID \oplus x) \oplus T_u)$ as before while the authentication message C_2 computed by system will change to $h(h(EID' \oplus x) \oplus T_u)$.

11. Password Guessing attack(supports):In Chien et al.'s[12] scheme, the password guessing attack is not possible because if an adversary intercepts the login request $C = (ID_i, C_2, T)$, he could not guess a correct password PW_i from C_2 because $C_2 = h(C_1 \oplus T)$, $C_1 = (R \oplus PW_i)$ and $R = h(ID_i \oplus x) \oplus PW_i$.
12. Smart Card Lost Attack(breaks):A user U_i may lose this smart card, which is found by an attacker. then he could extract the stored values through some technique such as by monitoring their power consumption and reverse engineering techniques as pointed out in Kocher et al[21]. He can extract the stored message $(R, h(.))$ from smart card then by intercepting login message $C = (ID_i, C_2, T)$ he can compute and $C_2^* = h(R \oplus PW_i^* \oplus T)$ by guessing password PW_i^* and if $(C_2^* = C_2)$ then the guessed password will be correct.
13. Stolen Verifier attack(breaks):In the Registration phase, (ID_i, PW_i) is passed to the server. We assume that an adversary A can obtain the secret value $(R, h(.))$ from stolen smart card. A can exhaustively examine all possible random number x until $R = h(ID_i \oplus x)$. So, the scheme is vulnerable to stolen verifier attack.

5.3 Review of Ku and Chen's Scheme[30]

This scheme has four phase:the Registration Phase,Login Phase,Verification Phase and the Password change phase.

A.Registration Phase:This phase is invoked whenever U initially registers to S. Let n denote the number of times U registers to S.
R1.U selects a random number b and computes $h(b \oplus PW_i)$
R2.$U \Rightarrow S$:$ID, h(b \oplus PW_i)$
R3.If it is $U's$ initial registration, S create an entry for U in the account database and stores $n = 0$ in this entry. Otherwise, S sets as $n = n + 1$ in the existing entry for U. Next, S performs the following computations:$R = h(EID \oplus x) \oplus h(b \oplus PW_i)$, where $EID = (ID.n)$.
R4.$S \Rightarrow U$:a smart card containing $(R, h(.), b)$.

B.Login Phase:The userU_i inserts his smart card to the card reader of a terminal, and keys his (ID_i, PW_i). Then Smart Card will perform the following operations:
L1.Compute $C_1 = R \oplus h(b \oplus PW_i)$,$C_2 = h(C_1 \oplus T)$ where T is the current date and time of the input device.
L2.$U_i \Rightarrow S$:$C = (ID_i, C_2, T)$.

C.Authentication Phase:
Upon receiving the login message $C = (ID_i, C_2, T)$ at time T' the remote system authenticates the user U_i with the following steps:
A1.Check the validity of ID_i.
A2.Verify the validity of the time interval between T and T'.if$(T' - T) \geq \Delta T$,

then the remote system rejects the login request.
A3.S computes $C_2' = h(h(EID \oplus x) \oplus T)$ If $(C_2' = C_2)$, then Server S accepts $U's$ login request. Otherwise, it rejects the login request.
A4.Computes $C_3 = h(h(EID \oplus x) \oplus T'')$ where T'' denotes $S's$ current time stamp.
A5.$S \Rightarrow U_i$:$D = (C_3, T'')$ for Mutual Authentication.
A6.On Receiving the message D from Remote Server, checks the validity of T_s
A7.The user U_i computes $C_3' = h(C_1 \oplus T'')$. If $(C_3' = C_3)$ holds, the legitimacy of AS is verified.

D.Password change Phase:
This phase is invoked whenever U wants to change his password PW with a new one, say PW_{new}.
P1.Uinserts his smart card in to card reader, enters ID and PW, and requests to change password. Next, U enters PW_{new}.
P2.Smart Card computes$R_{new} = R \oplus h(b \oplus PW_i) \oplus h(b \oplus PW_{new})$, which yields $h(EID \oplus x) \oplus h(b \oplus PW_{new})$ and then replaces R with R_{new}.

Cryptanalysis of Ku et al.'s Scheme[30]. 1.Smart card loss Attack(breaks): In Ku et al's scheme, $U's$ smart card contains $(R, b, h(.))$. Due to the fact that adversary could have extracted the secret information stored in the smart card by monitoring the power consumption[21] or by analysing the leaked information, the adversary can obtain $R = h(EID \oplus x) \oplus h(b \oplus PW)$ as well as b. Suppose that the adversary also has intercepted one of $U's$ past login messages, i.e,$C = (ID_i, C_2, T)$ he can perform a guessing attack to obtain PW_i by guessing a password PW_i^* and comparing $C_2' = h(R \oplus h(b \oplus PW_i^*) \oplus T_u)$ with the received C_2if $(C_2' = C_2)$, the adversary has correctly guessed $(PW_i^* = PW_i)$, otherwise, the adversary tries another candidate password. Since password PW_i are selected by users, they are usually short and simple for catchiness. Hence, PW_i could be obtained by off-line guessing attack.

2.Forgery Attack (breaks):Once the adversary obtain (PW, R, b) by guessing attack then he can compute $C_1 = R \oplus h(b \oplus PW_i)$ and then impersonates U by forging $U's$ login message $(ID, h(C_1 \oplus T_u^{'}), T_u^{'})$ at time $T_u^{'}$.

3.Denial of service attack(breaks):Due to the unchangebility of $h(ID \oplus x)$ in Chein et al.'s scheme [12],a forged login request can not be prohibited even when U detected that his C_1 has been compromised. Accordingly, Ku et al. extended ID with $EID = (ID.n)$ and replace $C_1 = h(ID \oplus x)$ with $C_1 = h(EID \oplus x)$ in their improved scheme, so that C_1 can be changed by EID with different n when C_1 has been compromised. Unfortunately, the number n is stored in an entry table in server side, which is somewhat equivalent with using verification table, and suffers from the risk of modified entry table and the cost of protecting and maintaining the entry table. Once the intruder modifies n in entry table, the user's login message C_2 keeps $h(h(EID \oplus x) \oplus T_u)$ as before while the authentication message C_2 computed by system will change to $h(h(EID' \oplus x) \oplus T_u)$.

4.Forward Secerecy Attack(supports):Suppose that an intruder has stolen the remote systems secret keys x.However in this scheme that an intruder can not compute each user's secret hash value$h(EID_u \oplus x)$, as $EID_u = h(ID_u \oplus n)$ and S sets as $(n = n + 1)$. Therefore, Hsiang et al.'s Scheme is not vulnerable to Forward secerecy.

5.Inefficiency for error password login(breaks):Even if Uinputs an error password in login phase, the smart card still sends $U's$ login request unconditionally to server. This error is not detected until the server checks $C_2 ? = h(h(EID \oplus x) \oplus T_u)$ at authentication phase. Therefore, the password authentication is delayed and inefficient.

6.Parellel Session Attack(breaks):Parellel Session Attack is possible since the message structures between the user and the server are same.

7.Replay Attack(supports):Replay attack is not possible since this scheme uses time stamps. The idea behind the use of time stamps is to generate a synchronization mechanism between the client and the server. Neither the replay of an old login message $C = (ID_i, C_2, T)$ in the login phase nor the replay of the remote server's response message $D = (C_3, T'')$ in the verification phase.

8.Stolen verifier attack(supports):In Registration phase,ID is passed in plain text form to the server, We assume that an adversary A intercepts and gets ID and $h(b \oplus PW_i)$. He can obtain R from lost smart card. Adversary cannot compute $h(EID \oplus x)$ as $h(EID \oplus x) = h(b \oplus PW_i) \oplus R$, $EID = (ID.n)$ and S set as $n = n + 1$.

9.Reflection attack(breaks):A malicious user intercepts the login request ID_i, C_2, T_u and replaces the pair (C_3, T_s) with (C_2, T_u) in verification phase. When the user U receives the pair (C_2, T_u), he verifies $C_2 = h(C_1 \oplus T)$, which holds truly. In this way, a malicious user reflects AS and U will be fooled. Thus, Chien et al.'s scheme fails to provide mutual authentication and vulnerable to the reflection attack.

10.Insider Attack(supports):The scheme is not vulnerable to insider attack because In the Registration Phase, PW_i is not in plain text form and calculated by $h(b \oplus PW_i)$ using secret key b which is known to user only and need not to remember after this step. So any insider of S can not calculate the password $h(b \oplus PW_i)$.

11.Password guessing attack(supports):The scheme is not vulnerable to insider attack because an adversary can not guess the password from the login message (ID_i, C_2, T_u).

5.4 Review of X-M.Wang et al.' Scheme [31]

A.Registration Phase:
The user selects a random number b and computes $h(b \oplus PW)$. An User submit his/her $ID_i, h(b \oplus PW)$ to the remote system. Upon receiving the registration request, the remote system performs the following steps:
R1.Compute $p = h(ID_i, x), R = p \oplus h(b \oplus PW), V = h_p(h(b \oplus PW))$, where x be a secret key maintained by Remote system and h is a one-way function.
R2.Personalizes the smart card with the parameters $R, V, h(.), h_p(.)$.

R3.$S \Rightarrow U_i$:Smart card.
R3.U enters b into his smart card so that he does not need to remember b anyone.

B.Login Phase:
The userU_i inserts his smart card to the card reader of a terminal and keys his (ID_i, PW_i). Then Smart Card will perform the following operations:
L1.Compute $p = R \oplus (b \oplus PW)$ and checks whether $h_p(h(b \oplus PW)) = V$, if not hold then reject the login request.
L2.Smart card generates a random number r, and performs the following computations:
$C_1 = p \oplus h(r \oplus b)$,$C_2 = h_p(h(r \oplus b) \oplus T_u)$
L3.$U_i \Rightarrow S$:$M = (ID_i, C_1, C_2, T_u)$ where T_u denotes $U's$ current time stamp.

C.Authentication Phase:
Upon receiving the login message $M = (ID_i, C_1, C_2, T_u)$ at time T_s the remote system authenticates the user U_i with the following steps:
A1.Check the validity of ID_i.
A2.Verify the validity of the time interval between T_u and T_s. If$(T_s - T_u) \geq \Delta T$, then the remote system rejects the login request.
A3.S Computes $p = h(ID_i \oplus x)$ and $C_1' = p \oplus C_1$.If $C_1' = C_1$ then check whether equation $h_p(C_1' \oplus T_u) = C_2$ holds or not. If holds, it means user is authentic and S accepts the login request, and performs step 4. Otherwise, S rejects login request.
A4.For the mutual authentication, S computes $C_3 = h_p(C_1' \oplus T_s)$ and then sends mutual authentication message C_3, T_s to user U.
A5.Upon receiving the message (C_3, T_s),U verifies either T_s is invalid or $(T_s = T_u)$, U terminates this session, otherwise performs step6.
A6.U computes $C_3' = h_p(h(r \oplus b) \oplus T_s)$ and computes $C_3' = C_3$ holds then user believes that the remote party is authentic system and the mutual authentication between U and S is completed, otherwise U terminates the operation. In addition, since r is randomly generated in each login phase, $C_1' = h(r \oplus b)$ shared between $U$and S can be used as the session key for the subsequent private communication.

D.Password Change Phase:U inserts his smart card in to card reader, enters ID, PW and requests to change password, then the smart card performs the following steps without any help of server.
P1.Compute $p^* = R \oplus h(b \oplus PW)$ and $V^* = h_p^*(h(b \oplus PW))$
P2.Check whether V^* equals to the stored V or not. If not, rejects the password change request. Otherwise U chooses a new password PW_{new}.
P3.Compute $R_{new} = p^* \oplus h(b \oplus PW_{new})$ and $V_{new} = h_p^*(h(b \oplus PW_{new}))$, and then stores R_{new}, V_{new} in to the user's smart card and replaces the old values R, V respectively. Now, new password is successfully updated and this phase is terminated.

Cryptanalysis of X-M-Wang et al.'s Remote User Authentication Scheme[31]:

1. Guessing attack resistance(supports):Firstly,R is stored in smart card with $R = h(ID \oplus x) \oplus h(b \oplus PW)$ since x and PW are unknown to adversary, one can get neither $h(ID \oplus x)$ nor $h(b \oplus PW)$ even if R, b are extracted from the smart card. Similarly, even if the stored information $V = h_p(h(b \oplus PW))$ is revealed, both $p = h(ID_i \oplus x), h(b \oplus PW)$ are still secure. Next, suppose the login message $M = (ID_i, C_1, C_2, T_u)$ sent by U be eavesdropped in a common channel. However, even under the advanced hash collision attack proposed by Wang et al's [32], the secret information$h(ID \oplus x)$ is still secure due to the fact that $C_1 = p \oplus h(r \oplus b)$,$C_2 = h_p(h(r \oplus b) \oplus T_u)$ are combined with $h(r \oplus b)$, which is randomized in each login request and one has no way to get it. Moreover,C_1, C_2, R, V are all combined with two random items.
2. Forgery/Impersonation attack resistance(supports):Impersonation Attack is not possible in this scheme, if an adversary attempts to modify $U's$ login message $M = (ID_i, C_1, C_2, T_u)$ in to $M = (ID_i, C_1^*, C_2^*, T_u*)$. However, this impersonation attempt will fail in step A3 of the Authentication phase, because there is no way to obtain the values of $h(ID_i \oplus x), h(r \oplus b)$ to compute the valid value of C_2.
3. Replay Attack Resistance(supports): Neither the replay of an old login message (ID_i, C_1, T_u) nor replay of the remote system's response C_3, T_s will work. It would have failed in step A2 and A5 of authentication phase, because of the time interval validation respectively.
4. Denial Of Service attack resistance(supports):In this scheme secret hash value i.e.$h(ID \oplus x)$ is not stored directly into smart card but is combined with the other hash values, such as $h(b \oplus PW)$ in R or $h(r \oplus b)$ in C_1, or act as the secret key of keyed hash function in V and C_2. Clearly, $h(ID \oplus x)$ can't be derived from any revealed value (R, V, C_1, C_2) or their combined values. So, the assumption of $h(ID \oplus x)$ being revealed is impractical or impossible for this scheme. That is, no entry table is necessary any more in this scheme.
5. Server spoofing attack resistance(supports) :The spoofing attack is completely solved by providing mutual authentication between user and remote system. Remote system S sends mutual authentication message C_3, T_s to the user. If an attacker intercepts it and re-sends the forge message i.e.(C_3^*, T_s^*) to user U, it will be verified in steps A5 and A6 of the authentication phase because the value of $C_3'$is computed by $C_3' = h_p(h(r \oplus b) \oplus T_s)$. In addition, replay of this message can be exposed because of the time stamp.
6. High efficiency in password authentication(supports):In login phase, If U inputs an error password PW', the smart card computes $p' = R \oplus h(b \oplus PW')$ and checks equation $h_p'(h(b \oplus PW'))? = V$ in 2 step. Obviously, the result is negative when $(PW \neq \ PW')$ and smart card terminates the login session. Thus, the validity of input password can be immediately detected by smart card yet need not wait for server authentication.
7. Forward secrecy(breaks):Suppose that an intruder has stolen the remote systems secret keys x. It is obvious in this scheme that an intruder can

easily compute each user's secret hash value as Compute $p = h(ID_i, x), R = p \oplus h(b \oplus PW), V = h_p(h(b \oplus PW))$, where x be a secret key maintained by Remote system and h is a one-way function and can impersonate any legitimate user. Therefore, in the future, the server's secret key x may be changed to prevent an intruder's malicious activity. However,it would be much too expensive to re-compute all secret hash values at a time and communicate them to the users. Therefore, this scheme does not guarantee a system's secret key forward secrecy.

8. Parallel Session Attack(supports):Without knowing a user's password, an attacker cannot masquerade as the legal user due to the typical structure of $V = h_p(h(b \oplus PW))$and (R, b).
9. Smart card lost attack(supports):Firstly,R is stored in smart card with $R = h(ID \oplus x) \oplus h(b \oplus PW)$. Since x and PW are unknown to adversary, one can get neither $h(ID \oplus x)$ nor $h(b \oplus PW)$ even if R, b are extracted from the smart card. Similarly, even if the stored information $V = h_p(h(b \oplus PW))$ is revealed, both $p = h(ID_i \oplus x), h(b \oplus PW)$ are still secure. Next, suppose the login message$M = (ID_i, C_1, C_2, T_u)$ sent by U be eavesdropped in a common channel. However, even under the advanced hash collision attack proposed by Wang et al's [32],the secret information $h(ID \oplus x)$ is still secure due to the fact that $C_1 = p \oplus h(r \oplus b)$,$C_2 = h_p(h(r \oplus b) \oplus T_u)$ are combined with $h(r \oplus b)$, which is randomized in each login request and one has no way to get it. Moreover, (C_1, C_2, R, V) are all combined with two random items.
10. Stolen verifier attack(supports):In this scheme all the secret values are stored in hashed way. So an adversary can not steal the secret information during the transaction of information.
11. Reflection Attack(supports):Reflection attack is not possible since message structures between the user and the server are different. Here the user computes $C_1 = p \oplus h(r \oplus b)$,$C_2 = h_p(h(r \oplus b) \oplus T_u)$ by step L3,L4 and sends (C_1, C_2) to the server. The server intern sends mutual authentication message (C_3, T_s) by step A3,A4 of Authentication phase. Both the messages have different structures. So the adversary will not be able to perform this attack.
12. Insider Attack(supports):The scheme is not vulnerable to insider attack because In the Registration Phase, PW_i is not in plain text form and calculated by $h(b \oplus PW_i)$ using secret key b which is known to user only and need not to remember after this step. So any insider of S can not calculate the password $h(b \oplus PW_i)$.

6 Performance Comparison

In This section, we compare the schemes in terms of security requirements.

By checking out all the security requirements that are listed in the section 3, we can judge if a scheme deserves the title of an ideal password authentication scheme. Comparison of security requirements:

Security Requirements Schemes	SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8	SR9
Sun et al.[11]	N*	N*	Y*	N*	Y	Y*	N*	Y*	Y*
Chien et al.[12]	N*	N*	N[2]	Y*	Y*	N*	N*	N*	N*
Ku et al.[30]	N[31]	N[31]	N*	Y*	Y*	N[31]	Y*	N*	Y*
Wang et al.[31]	Y[31]	Y[31]	Y*	Y[31]	Y[31]	Y*	Y*	Y*	Y*

6.1 Security Requirements Analysis

In table 1,a comparison of security requirements is shown in which the following notations are used.
SRi:Proposed security requirements are in Section 3.
Y:Meets the security requirement,cryptanalysis done by the corresponding authors.
N[n]:Not meets the security requirements,cryptanalysis done by the [n] authors.
Y*:Meets the security requirement,cryptanalysis done by us.
N*:Not meets the security requirement,cryptanalysis done by us.

7 Conclusion

In this paper, the survey of the five smart card based authentication schemes over insecure networks has been done. We have defined the security requirements and An ideal password authentication scheme should satisfy and achieve it. Survey results are based on the cryptanalysis done by other researchers and also done by us. We have done the security and functionality comparison of schemes based on the 9 security requirements. Except one of them, Wang et al.'s[31], all the schemes do not meet all the security requirements , Wang et al.'s[31] scheme satisfies all the security requirements but do not achieve all the goals, So we can not say that this is an ideal password authentication scheme. Therefore, there is a need to look into these goals in future research work. Unfortunately, none of the schemes can satisfy all the security requirements and all the goals. We hope our work will provide a better understanding of the security challenges of smart card based remote user authentication and pave the way for further research in this area.

References

1. Shimizu, A.: A dynamic password authentication method by one way function. IEICE Transactions 173-D-1(7), 630–636 (1990)
2. Hsu, C.-L.: Security of chien et al.'s remote user authentication scheme using smart cards. Computer Standards and Interfaces 26, 167–169 (2004)
3. Hsu, C.L.: A user friendly Remote User Authentication scheme with smart cards against impersonation attacks. Applied Mathematical and Computer 170, 135–143 (2005)
4. Chan, C.K., Chang, L.M.: Cryptanalysis of a Remote user authentication scheme using smart card using smart cards. IEEE Trans, Consumer Electron 46, 992–993 (2000)

5. Li, C.T., Hwang, M.S.: An efficient biometric based remote user authentication scheme using smart cards. Journal of Network and Computer Applications 33(1), (5) (January 2010)
6. Hsu, C.L.: Security of Chien et al's remote user authentication scheme using smart cards. Computer Standard and Interfaces 26(3), 167–169 (2004)
7. Sun, D.-Z., et al.: Weakness and improvement on wang-Li-Tie's user friendly remote authentication scheme. Applied Mathematics and Computation 170, 1185–1193 (2005)
8. Yoon, E.J., Ryu, E.K., Yoo, K.Y.: Further improvement of an efficient password based remote user authentication scheme using smart cards. IEEE Trans on Consumer Electronics 50(2), 612–614 (2004)
9. Yoon, E., Yoo, K.: More efficient and secure remote user authentication scheme using smart card. In: Proceeding of 11th International Conference on Parallel and Distributed System, vol. 2, pp. 73–77 (2005)
10. Hwang, Hwang, L.M.S., Li, L.H.: A new remote user authentication scheme using smart card. IEEE Transactions on Consumer Electronics 46(1), 28–30 (2000)
11. Sun, H.M.: An efficient remote user authentication scheme using smart card. IEEE Trans on Consumer Electronic 46(4) (2000)
12. Chien, H.Y., Jan, J.K., Tseng, Y.M.: An efficient and practical solution to remote authentication: smart cards. Computer and Security 21(4), 372–375 (2002)
13. I-En-Liao, C.-C., Lee, N.-S., Hwang, N.-S.: A password authentication scheme over insecure networks. Journals of Computer and System Sciences 72, 727–740 (2006)
14. Shen, J.J., Lin, C.W., Hwang, M.S.: A modified remote user authentication scheme using smart cards. IEEE Trans, Consumer Electron 49(2), 414–416 (2003)
15. Xu, J., Zhu, W.-T., Feng, D.-G.: An improvement smart card based Password Authentication scheme with provable security. Computer Standard and Interfaces 31, 723–728 (2009)
16. Lampot: Password authentication with insecure communication. ACM 24(11), 770–772 (1981)
17. Gong, L.: A security risk of depending on synchronized clocks. Operating Systems Review 26(1), 49–53 (1992)
18. Sandirigama, M., Shimiz, A., Noda, M.T.: Simple and secure password authentication protocol. (SAS), IEICE Transactions on Communication E83-B(6), 1363–1365 (2000)
19. Haller, N.H.: The S/KEY(TM) one time password system. proc. In: Proc. Internet Society Symposium on Network and Distributed System Security, pp. 151–158 (1994)
20. Lee, N.Y., Chin, Y.C.: Improved RAS with smart cards. Computer Standards and Interface 27(2), 177–180 (2005)
21. Kocher, P.C., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, p. 388. Springer, Heidelberg (1999)
22. Wu, S.T., Chieu, B.C.: A user friendly remote user authentication scheme with smart cards. Computers and Security 22(6), 547–550 (2003)
23. Lee, S.W., Kim, H.S., Yoo, K.Y.: Improvement of chien etal's remote user authentication scheme using smart card. Computer Standards and Interface 27(2), 181–183 (2005)
24. Kim, S.K., Chung, M.G.: More secure remote user authentication scheme using smart cards. Journal of Computer and Communications, doi:10.10161-1 coman 2008.11.026
25. Kim, S.K., Chung, M.G.: More secure remote user authentication scheme. Computer Communication (2009)

26. Lee, S., Kim, H., Yoo, K.: Improvement of chen et's remote user authentication scheme using smart cards. Computer Standards and Interface 27, 181–183 (2004)
27. Lee, S., Kim, H., Yoo, K.: Improved efficient remote user authentication scheme using smart card. IEEE Trans on Communication Electronics 50(2), 565–567 (2004)
28. Hwang, T., Chen, Y., Laih, C.S.: Non interactive password authentication without password tables. In: Proc. IEEE Region 10 Conference on Computer and Communication Systems, Hong Kong, pp. 429–431 (September 1990)
29. Chen, T.H., Lee, W.B.: A new method for using hash functions to solve remote user authentication. Computers and Electicals Engineering 34, 53–62 (2008)
30. Ku, W.C., Chen, S.N.: weakness and improvement of an efficient password based remote user authentication scheme using smart cards. IEEE Trans on Consumer Electronics 50(1), 204–207 (2004)
31. Wang, X.M., Zhang, W.F., Zhang, J.S., Khan, M.K.: Cryptanalysis and improvement on two efficient remote user authentications scheme using smart cards. Computer Standards and Interfaces 29(5), 507–512 (2007)
32. Wang, X., Yin, Y.L., Yu, H.: Finding Collisons in the full SHA1 (February 2005), `http://www.infosec.sdu.edu.en/paper/sha1`
33. Wang, Y.J., Li, J.H., Tie, L.: Security analysis and improvement of a user friendly remote authentication protocol. Applied Mathematics and Computer (in press)

Outlier Detection and Treatment for Lightweight Mobile Ad Hoc Networks

Adarsh Kumar[1,2], Krishna Gopal[2], and Alok Aggarwal[1,3]

[1] Computer Science Engineering and Information Technology Department,
[2] Jaypee Institute Of Information Technology, Noida, India
[3] JP Institute Of Engineering and Technology, Meerut, India
{adarsh.kumar,krishna.gopal}@jiit.ac.in, alok289@yahoo.com

Abstract. This work is to detect and prevent unprecedented data identified from lightweight resource constraint mobile sensor devices. In this work, event or error detection technique of Traag et. al., local-global outlier algorithm of Branch et. al., Teo and Tan's protocol of group key management and Cerpa et. al protocol of Frisbee construction are integrated and modified for lightweight resource constraint devices [20][22]-[24]. The proposed technique in this work is better than other techniques because of: (a) scalability, (b) optimization of resources, (c) energy efficient and (d) secure in terms of collision resistant, compression, backward and forward secrecy. The deviations in modified form of proposed mechanism are corrected using virtual programmable nodes and results show that proposed scheme work with zero probability of error and attack.

Keywords: lightweight, outlier, anomalies, security, key management, MANET.

1 Introduction

Mobile Ad Hoc Networks (MANETs) consist of self configuration, infrastructure less, short range wireless technology, dynamic topology and mobile or semi-mobile devices. Various applications of MANETs are: Vehicular Ad-Hoc Networks (VANETs), house-hold appliances, military purposes, commercial security devices, peer to peer applications, mobile game programming etc. Major challenges of these types of networks are: security constraints, scarcity of resources, limited bandwidth availability, small subnets, traffic overhead, high processing cost etc. Since MANETs frequently and dynamically changes subnets thus these low capacity devices demand lightweight or ultra lightweight cryptographic implementation. According to Moore's law, only 30% resources are available for cryptographic primitives. Various security primitives need to be integrated within available resources for resource constraint mobile nodes are [1]:

- Availability: ensures that nodes should be available for communication despite of any worst conditions.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 750–763, 2013.

- Confidentiality: ensures the security breach of information during communication should not be compensated at any cost.
- Integrity: guarantees that message or user authentication information is never corrupted.
- Authentication: ensures that impersonation, masquerading and interference of resources, user identities and sensitive information should not be tolerated.
- Authorization: ensures that resource or information is trusted and collision resistant.
- Key Management: promises that key generation, transportation, confirmation and renewing is proper, secure and fast.
- Non-repudiation: convince the source node from not betray from sending information and other nodes about compromised source node.

Other security factors that need to be taken care of are: frequent key contributiveness, pre-image resistant, information distortion, message replay, active or passive attacks etc. This work is in continuation of work done to secure the MANET with respect to confidentiality, integrity, authentication & authorization and key management for resource constraint devices [23]. In this work, concentration is drawn towards availability of nodes for communication despite of attacks or corruption. Intrusion is an important security breach and is meant to compromise the cryptographic primitives like: availability, confidentiality, integrity, key management. Non-availability of nodes is mainly due to outliers or anomalies created inside the network [2]. The outliers or anomalies are the deviations of data as compared to normal data in order to gain some advantage.

The rest of the paper is organized as follows. Section 2 provides introduction to anomalies and classification of various outlier detection techniques. Section 3 describes the proposed approach to distinguish between an error or an event based on Markov chain and proposed local-global outlier detection algorithm. Section 4 describes the experimental setup, performance analysis of proposed algorithm, verification and validation of results and algorithm correction. Section 5 presents the conclusion.

2 Related Work

2.1 Outliers in Sensor Based Networks

Various sources of outlier in sensor networks are: (a) Fault detection, due to hardware, software or environmental anomalies [3-4], (b) automatic event detection, due to uncertainty in data [5-6], [11-12] and (c) intrusion detection, due to deviation from regular system usage in order to compromise security primitives [7-9]. These anomalies can occur at data, node or network levels [10].

2.2 Outlier Detection Techniques

In the literature, outlier detection techniques can be classified into various categories:

First classification is based on node, network or data based outliers. Node based outliers occur from internal system calls with sequential data [13]. Network based outliers occur from network generated socket calls and data based outliers are because of calculation errors. Various node, network and data level detection techniques are: statistical techniques based methods, models based methods, state machine based methods, neural network based, rule based systems etc.

Second classification is based on: (a) data attributes and its correlation, (b) local or global views of outliers, (c) error, event or attack based outliers, (d) degree of deviation from normal data and (e) supervised, semi-supervised or unsupervised data. Various detection techniques used to analyze outliers based on above classification can be categorized as: (a) statistical based techniques, (b) nearest neighbor based, (c) clustering based, (d) bayesian network based, (e) spectral decomposition based etc. Statistical based techniques can have the knowledge about data. For example, Gaussian based techniques. Statistical techniques without prior data information are: kernel based or histogram based [14]. Well known nearest neighbor based technique is single hop Frisbee construction technique [24].Other non-statistical techniques are: network intrusion detection, neural network based etc.

Third classification is based on supervised or unsupervised detection mode. Supervised data techniques have prior knowledge about data sets consisting of information about anomalies and normal data. Unsupervised techniques do not have any prior information about data sets. For example, supervised techniques are: Bayesian network based, SVM based and unsupervised techniques are: statistical based, knowledge based, neural network based, fuzzy logic based, Markov or Hidden Markov Model (HMM) based, nearest neighbor based, clustering algorithm based etc [15][16][30].

Fourth classification is based on: (a) distance based, (b) density based, (c) machine learning or soft computing based. Distance based outlier detection are based on distance between selective node's attributes from the data set taken into consideration. For example, Hawkin outlier [17] and DB outlier technique [18]. Popular density based outlier techniques are: LOF, RDF, natural outlier based etc [31]-[33] and machine learning based technique is: SVM.

Fifth classification is based on: (a) local outlier, (b) global outliers, (c) semi global outliers, (d) distributed global outliers and (e) semi-global distributed outlier detection mechanism. In these outlier detection techniques local views of neighbors are collected to form a local view and then these views are further broadcasted to global nodes [20].

In this work, hybrid approach is developed for lightweight devices. Lightweight protocol are identified and integrated in order to get energy efficient, optimized and scalable solution for resources constraint mobile sensor devices. First an approach of distinguishing between an event and an error for a mobile node is proposed using Markov chain, which is based on Traag et. al. technique [25]. A lightweight local-global outlier detection mechanism is integrated with modified Teo and Tan's protocol for anomaly score calculation [20]. In order to validate the results, automated security tools are studied and two tools are used for experimental evaluation [34].

3 Proposed Approach

3.1 Assumptions and Premises

Let 'R' be the region selected for observation at starting time T_S to ending time T_E during a week of observation w_o. Let T_{WIN} is the time window $[T_S, T_E]$ of a complex event. T_{WIN}^{o} be time window during week 1 to W. Furthermore, a node made message communication or acting as router 'ROU'. Let node has started message communication MC_1…….MC_n during time TC_1….TC_n and routing ROU_1^M……… ROU_n^M at time $TROU_1$….$TROU_n$ in time window T_{WIN}. $MOBC_1^{((x_1^i,y_1^i)......(x_1^n,y_1^n)}$ ……….. $MOBC_1^{((x_1^i,y_1^i)......(x_1^n,y_1^n)}$ be the mobility of nodes during message communication and $MOBROU_1^{((x_1^i,y_1^i)......(x_1^n,y_1^n))}$ ……….. $MOBROU_1^{((x_1^i,y_1^i)......(x_1^n,y_1^n))}$ be the mobility of nodes during routing. Following are the steps to be followed in order to calculate anomaly score.

1. *Find the probability that a mobile node is following a particular path.*

Let $P_{(i,\ j)}$ be the probability of any mobile node MN_x to move from $MN_z^{(x_i,y_i)}$ *to* $MN_z^{(x_z,y_z)}$, where z∊[1…n].

According to Markov chain, a probability of following a path through states $s_1^{(x_1,y_1)}$ to $s_n^{(x_n,y_n)}$ is calculated as:
$\mathrm{P}(s_1^{(x_1,y_1)}, s_1^{(x_2,y_2)}\ldots..\ s_n^{(x_n,y_n)}) = s_1^{(x_1,y_1)}, s_1^{(x_2,y_2)}\ldots..\ s_n^{(x_n,y_n)} = \mathrm{P}(s_1^{(x_1,y_1)} = s_1^{(x_1,y_1)})p_{x_1x_2}p_{x_2x_3}\ldots..p_{x_{n-1}x_n}=\mathrm{P_S}$

With integration of communication states and routing states, probability can be calculated as:

$\mathrm{P_S} = \mathrm{P}((S_{MOBC_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}||S_{MOBROU_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}), \ldots\ldots\ldots\ldots(S_{MOBC_n}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}||S_{MOBROU_n}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))})) = \mathrm{P}(s_1^{(x_1,y_1)} = (S_{MOBC_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}||S_{MOBROU_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}))\ p_{x_1x_2}p_{x_2x_3}..p_{x_{n-1}x_n}.$

2. *Find the probability that node is attending regular event in a region 'R'.*

In order to find this probability, average probability of presence in a regular region 'R' by mobile node 'MN' using T_{WIN} is calculated as:

$$P_S^{AVG} = (1/(W-1)\sum_{v=1,v=w}^{W} P_s(MN,R,T_{WIN}^{v})$$

According to Markov chain, every next sequence is dependent upon previous states. Thus

$$P_S^{AVG} = (\frac{1}{W-1})(\sum_{v=1,v=w}^{W}(S_{MOBC_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))}||S_{MOBROU_1}^{((x_1^i,y_1^i)....((x_1^n,y_1^n))})$$
$$.\ p_{x_1x_2}p_{x_2x_3}..p_{x_{n-1}x_n}, \mathrm{R}, (T_{WIN.........}^{Ts}T_{WIN}^{Ts})_{\mathrm{v}})$$

3. *Detecting an event*

In order to find that whether an event has occurred or not, anomaly score is calculated as:

Anomaly Score = ($MN_{Active}^{Attendee}$- ($AVG_{(MN_{ACTIVE}+MN_{SLEEP})}^{Attendee}$)) / STDEV

Higher event range values than threshold (>4) are considered as anomalies.

3.2 Distributed Local-Global Outlier Detection Mechanism

After deciding the method to distinguish between an event and an error in subsection 3.1, strategy of how to deploy detection method is proposed in this subsection. Detection methods can be deployed (a) centrally or (b) distributed. In centralized outlier detection deployment, it is required to collect all data at one central node and test it by single or group of nodes. Such a centralized mechanism has several disadvantages [19]: (a) expose central point of failure for system, (b) data collection and processing at some central point can cause end to end delays, (c) power consumption overhead on centralized and intermediate nodes, (d) scalability and robustness of network make it imperative to deploy the strategy distributed.

3.2.1 Distributed System Setup

The distributed system architecture consists of local view formation and global view formation strategies. In local view formation, a group of nodes in close vicinity form the view about anomaly in the data sets. These views collectively help in formation of weighted score for global view formation. Global view will instruct the active nearby nodes of Markov chain trajectory's sensor nodes to update anomaly score. Based on this anomaly score, the misbehaving nodes $\sum_{i=1}^{n} MN_i$ are charged for power and communication loss until they prove their authenticity. Neighboring nodes will change the view about a particular node MN_i if k-neighboring nodes agree to authenticate the node MN_i. The factor 'k' is calculated using distributed algorithm [20] and Shamir's threshold secret sharing scheme [21].

As shown in figure 1, in order to deploy distributed approach J. C. M. Teo and C. H. Tan's approach of group formation is modified for mobile nodes[22][23]. Each subgroup will form a local view in terms of anomaly score and this anomaly score is transmitted to main group controller through subgroup controller during group key updating process. If some critical updating is required then it can initiate Critical Updating Process (CUP) prior to group key updating process. Examples of critical situations are: sensor malfunctioning due to tsunami or earthquake, power failure etc. Algorithm for CUP is discussed in next subsection. Top layer of hierarchy consist of single main group and every other subgroup is controlled by subgroup controller in its parent directory. Virtual nodes help in formation of optimized subgroups in close vicinity. Each subgroup runs an algorithm at its local level called Local View Formation algorithm (LVFA) and main group runs Global View Formation Algorithm (GVFA) for anomaly score calculation. These algorithms are described in next subsections.

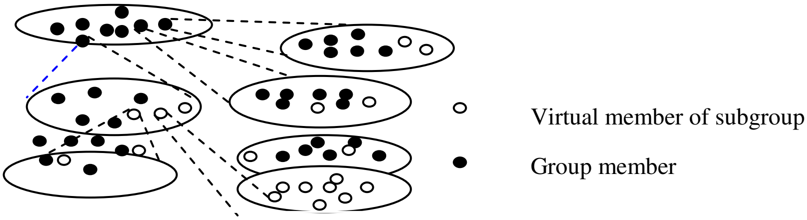


Fig. 1. Virtual group/subgroup hierarchy

3.2.2 Local View Formation algorithm

In this subsection, LVFA is proposed through which mobile sensor nodes in a subgroup form a generalized view about an event in a close vicinity. If anomaly score of an event increases a threshold limit then the error is reported to main group controller. In the process of error or anomaly formation, it should be taken into consideration that sensor nodes have scarcity of power resources and thus its losses should be minimized. In order to minimize the losses, "Frisbee Model" is integrated with Markov chain to trace the path of a mobile node and Frisbees [24]. These Frisbees will help to form a view about an event or calculate anomaly score at local and global levels. In order to implement optimal Frisbee, Teo and Tan's group key management protocol is used for secure message exchange and one hop nearest neighbor will construct the Frisbee periphery. Figure 2 shows the construction of Frisbees with integration of Teo and Tan's protocol of subgroup construction. Figure 2a shows the possible trajectory according to Markov chain. Figure 2b shows the construction of Frisbee periphery using single hop neighbor based shamir's threshold scheme. Figure 2c shows the sequence of Frisbees constructed during node's mobility or trajectory. LVFA is developed as follows:

Protocol: Local event or anomaly detection.

Premises: Let HL_i is the hierarchy of subgroups $SG_j^{HL_i}$, where each subgroup consists of 'n' number of elements and i∊{1,2,3…..s}, j∊{1.2…r}. 'h' is the height of hierarchical structure such that $m=n^h$, j^{th} subgroup at i^{th} layer for j∊{0….. n^i-1} is represented by $SG_j^{HL_i}$, subgroup controller of j^{th} subgroup at hierarchical layer HL_i is represented by $SG_{SC_j}^{HL_i}$. k^{th} member of j^{th} subgroup at hierarchical layer HL_i is represented as $SM_{(j,k)}^{HL_i}$, where k=jn+l for l∊{0,…..n-1}. Data compression, collision resistance, forward and backward secrecy is achieve through a hash function 'H'. $SMO_{(j,k)}^{HL_i}$ represents the outlier node in i^{th} hierarchy.

Goal: Form subgroups and calculate anomaly score.

Step 1: Form initial 1-hop nearest neighbor Frisbee

a. $SM_{(j,k)}^{HL_i}$ broadcasts it's group key updating request to other nodes in the subgroup $SG_{SC_j}^{HL_i}$.

b. Like $SM_{(j,k)}^{HL_i}$'s contribution request to update group key, all subgroup members will send their primitive contributions also.

c. $SG_{SC_j}^{HL_i}$ will decide the neighbor nodes of $SM_{(j,k)}^{HL_i}$ based on 1-hop criteria and form initial Frisbee.

Step 2: Apply Markov chain model & found the possible Frisbee trajectory.

a. Markov chain will give an approximation of trajectories to be followed by mobile node in order to attend an event using formula:

$P_S = P(s_1^{(x_1,y_1)} = s_1^{(x_1,y_1)}) p_{x_1x_2} p_{x_2x_3} \ldots\ldots p_{x_{n-1}x_n}$

The best path is selected (i.e. P_S =1).

b. $SG_{SC_j}^{HL_i}$ calculate anomaly score based on scheme mentioned in subsection 3.1 and update $SG_{SC_j}^{HL_{i+1}}$ with anomaly score and trajectory followed in jth subgroup at ith layer.

Step 3: Outlier node can later put a request to nearby nodes for change of their views based on new anomaly score.

a. If some $SM_{(j,k)}^{HL_i}$is found as sending data anomaly source then using previous two steps, it can be easily detected.

b. After some time interval, if same mobile node $SM_{(j,k)}^{HL_i}$ want to attend an event then it will send a request to it's subgroup controller.

c. $SG_{SC_j}^{HL_i}$will use Burmester & Demesdt protocol (BD protocol) to prove the authenticity and shamir's threshold mechanism to recalculate it's anomaly score [21][29].

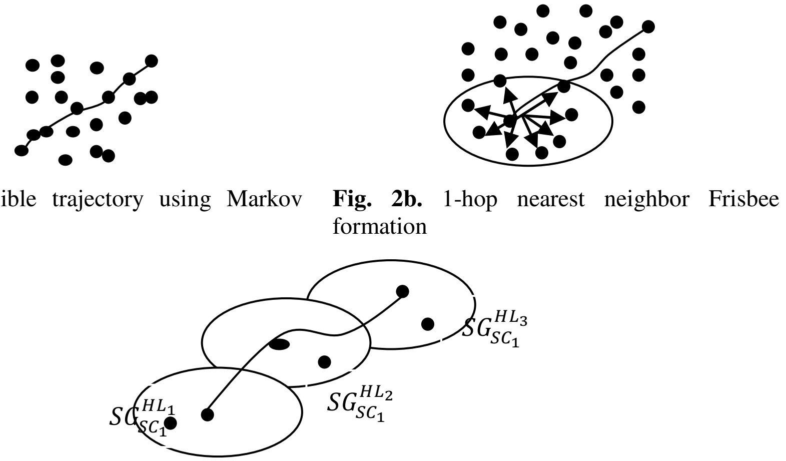


Fig. 2a. Possible trajectory using Markov model

Fig. 2b. 1-hop nearest neighbor Frisbee formation

Fig. 2c. Sequence of Frisbees formed during trajectory

Fig. 2. Frisbee formation during LVFA

3.2.3 Global View Formation algorithm

Protocol: Global event or anomaly detection

Premises: Same as local event or anomaly detection protocol.

Goal: Collect anomaly scores from subgroups and broadcast opinion about outliers.

Step 1: Collecting anomaly scores from all layers.

a. $SG_{SC_j}^{HL_i}$ of each layer will send encrypted anomaly score to its parent subgroup controller.

b. This subgroup controller will send an integrated encrypted report to its parent subgroup controller.

c. This process will continue until all anomaly score are collected by primary subgroup controller.

Step 2: Form a global view of outlier nodes.

a. As outlier node can attend some other events in different subgroup thus a generalized option about outlier nodes should be communicated to each subgroup controller.

b. All outlier mobile nodes $SMO_{(j,k)}^{HL_i}$ are identified. A report of outliers is formed and sends to subgroup controllers. $SG_{SC_j}^{HL_i}$sends the report to every subgroup controller at HL_{i+1} layer.

c. At local level, views can be updated using same process as in LVFA's step 3.

3.2.4 CUP Algorithm

Protocol: Anomaly updating before renewing the group key.

Premises: Same as local event or anomaly detection protocol.

Goal: Collect anomaly scores from subgroups and broadcast opinion about outliers.

Step 1:- $SG_{SC_j}^{HL_i}$ sends E_K("Anomaly") to $SG_{SC_j}^{HL_{i-1}}$.

Step 2:- Step 1's process continues until it reaches to main subgroup.

Step 3:- Top layer hierarchy subgroup controller $SG_{SC_j}^{HL_i}$initiate the process of group key formation.

Strengths of proposed mechanism are: (a) solution is optimized and scalable, (b) work with integration of lightweight encryption/decryption process, (c) energy efficient because of Frisbee model, (d) provide security from well known attacks.

4 Result and Analysis

4.1 Experimental Setup

In order to evaluate the performance, Linux operating system is selected with ns-3 platform and python language [26]. Number of nodes selected for analysis varies from 50 to 200. The parameters taken for analysis are: anomaly detection ratio (ADR), wrongly calculated anomaly ratio (WCAR), average local anomaly detection ratio (ALADR) and average local wrongly calculated anomaly ratio (ALWCAR). ADR is the ratio of anomalies detected by local-global mechanism to original number of anomalies present in the data set. WCAR is the ratio of number of normal data detected as outlier to total number of anomalies. ALADR & ALWCAR are the average values of local subgroup's ADR and WCAR respectively.

Table 1. Different detection ratios to calculate success rate

	N=50	**N=100**	**N=200**
ADR	0.860	0.770	0.700
WCAR	0.010	0.060	0.090
ALADR	0.910	0.800	0.740
ALWCAR	0.001	0.009	0.011

Table 1 shows the analysis of various ratios. It can be seen that accuracy decreases with increase in number of nodes. Second, it is important to notice that the global outlier detection ratio is having errors. It means that global outliers are not getting scores properly. In order to correct the result following correction is made to local outlier detection algorithm.

4.2 LVFA Correction

In order to reduce the error at global level, LVFA is modified. After making this modification and result analysis, it is observed that this error was because of inactiveness of mobile nodes. In order to remove the error because of inactiveness, virtual node concept is added. Instead of storing local view about anomaly score at subgroup controller, it is stored at virtual node subgroup controller. These virtual nodes are the programmable nodes without any hardware as discussed in figure 1. The correction is as follows:

Protocol: Local event or anomaly detection.
Premises: Same as local event or anomaly detection protocol.
Goal: Remove the deviation using virtual programmable nodes.
Step 1: Same as Step 1 of local event or anomaly detection protocol.
Step 2: Apply Markov chain model & found the possible Frisbee trajectory.

a. Same as Step 2a of local event or anomaly detection protocol.
b. $SG_{SC_j}^{HL_i}$ calculate anomaly score based on scheme mentioned in subsection 3.1 and update $SG_{SC_j}^{HL_{i+1}}$ with anomaly score and trajectory followed in jth subgroup at ith layer.
c. $SG_{SC_j}^{HL_i}$ will send its information to virtual node subgroup controller $VNSG_{SC_j}^{HL_i}$. Like $SG_{SC_j}^{HL_i}$, $VNSG_{SC_j}^{HL_i}$operate and exchange information about subgroup. The main advantage of these virtual nodes is that these are supposed to be active throughout lifecycle.

Step 3: Outlier node can later put a request to nearby nodes for change of their views based on new anomaly score.

a. If some $SM_{(j,k)}^{HL_i}$is found as sending data anomaly source then using previous two steps, it can be easily detected.

b. After some time interval, if same mobile node $SM_{(j,k)}^{HL_i}$ want to attend an event then it will send a request to it's subgroup controller.

c. $SG_{SC_j}^{HL_i}$will use BD protocol to prove the authenticity and shamir's threshold mechanism to recalculate it's anomaly score.

d. $SG_{SC_j}^{HL_i}$ send calculated anomaly score to $VNSG_{SC_j}^{HL_i}$. These virtual nodes further exchange information about top layer virtual subgroup controller for calculation of global anomaly score.

4.3 Results Evaluation

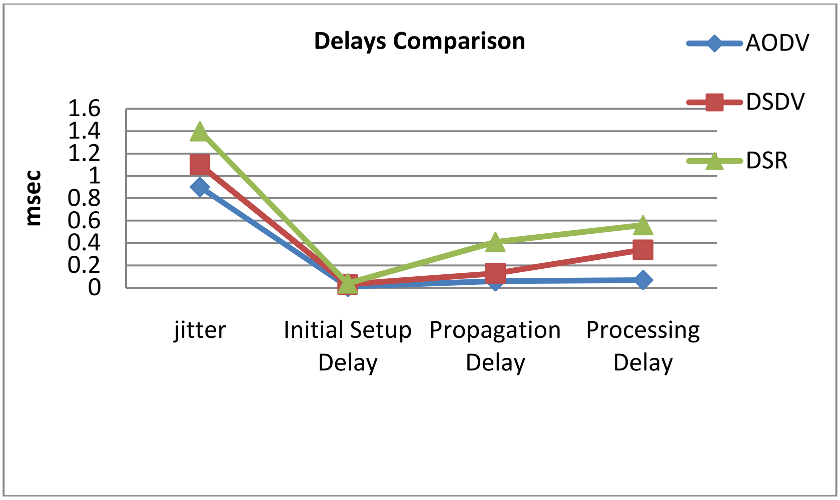


Fig. 3. Delay Comparison of proposed mechanism over MANET routing protocols

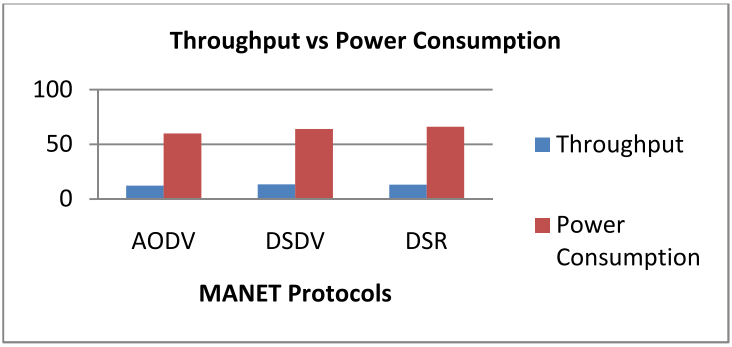


Fig. 4. Power vs Throughput comparison over MANET routing protocols

After making LVFA correction, it is found that error or deviation is negligible. Further, in order to evaluate the performance various parameters taken into the considerations are: end to end delay, throughput, jitter and power consumption. Figure 3 and Figure 4 shows that Ad On-demand Distance Vector (AODV) routing protocol is having minimum average value of end to end delay, jiiter, power consumption as compared to Dynamic Source Routing (DSR) and Destination Sequenced Distance Vector (DSDV) routing protocols. Since proposed protocol is also an on-demand protocol thus it resembles with the operations of AODV routing protocols and provide good amount of throughput. AODV and DSR are reactive routing protocols. Out of these two protocols, DSR is providing continuous increase in end to end delay than AODV because in proposed scheme single hop neighbor discovery protocol is used which is similar to the scheme used in AODV.

4.4 Verification and Validation

```
Process:
            [Process]
--- Query             [Query]
Completing………
Starting query [Query]
Goal [un] reachable: [Goal]
Abbreviations:
…………….
…………….
[Attack derivation]
…………….
RESULT  not  attacker(secret SG N_SG []) is true
RESULT  not  attacker(secret SM N_SM []) is true
RESULT  not  attacker(secret SMO N_SMO []) is true
RESULT  not  attacker(secret VNSG N_VNSG []) is true
RESULT inj –event (endHL_iparam(x_1400)) ===> inj-event (beginHL_i(x_1400)) is true
RESULT inj –event (endSM_iparam(x_1589)) ===> inj-event (beginSM_i(x_1589)) is true
RESULT inj –event (endSG_iparam(x_1623)) ===> inj-event (beginSG_i(x_1623)) is true
RESULT inj –event (endSMO_iparam(x_1801)) ===> inj-event (begin SMO_i(x_1801)) is true
RESULT inj –event (endSMO_iparam(x_1945)) ===> inj-event (begin SMO_i(x_1945)) is true
```

Fig. 5. ProvVerif results showing passing of all tests

In this subsection, automated verification tools AVISPA and ProVerif are used to verify that protocol is protected from attacks or corruption [27][28]. These tools are used to graphically test various points of protocol failure under the inspection of different probability models. AVISPA check the security of protocols using HLPSL specification language. After checking against man in the middle, replay and denial of service attacks, it is found that tests have given “no attacks found” results. This validates that local and global outlier mechanism is secure. Figure 5 shows the results of proposed mechanism using ProVerif. ProVerif is used to test the backward and forward compatibility. Here backward compatibility means previous key will not help the attacker to find new key for any node. Here, if some attacker is able to find the key then it can easily manipulate the messages. Those messages will be considered as anomalies. Similarly, forward compatibility means if new key or keys are leaked then past key should be secure. This process can again generate anomaly data by an attack. Results show that protocol is secured from both backward and forward corruption.

5 Conclusion

In this work, Traag et. al., Branch et. al., Teo and Tan and Cerpa et. al. protocols of event detection in mobile phones, local-global algorithm for anomaly view formation, group key management and Frisbee construction protocols respectively are integrated and modified for lightweight resource constraint devices. After observing the error of 3 to 5 % in anomaly detection, correction to LVFA is made and result are verified using two automated verification and validation tools: AVISPA and Proverif. Results shows that proposed mechanism work efficiently with AODV routing protocol and with no attack.

References

1. Zhou, L., Haas, Z.J.: Securing Ad Hoc Networks. IEEE Network 13(6), 24–30 (1999)
2. Heady, R., Luger, G., Maccabe, A., Servilla, M.: The architecture of a network level instrusion detection system, Computer Science Department, University of New Mexico. Tech. Rep. (1990)
3. Chen, J., Kher, S., Somani, A.: Distributed fault detection of wireless sensor networks. In: Proceedings of the 2006 Workshop on Dependability Issues in Wireless Ad Hoc Networks and Sensor Networks, pp. 65–72 (2006)
4. Luo, X., Dong, M., Huang, Y.: On distributed fault tolerant detection in wireless sensor networks. IEEE Transactions on computers 55(1), 58–70 (2006)
5. Krishnamachari, B., Iyengar, S.: Distributed Bayesian algorithms for fault tolerant event region detection in wireless sensor networks. IEEE Transactions on Computers 53(3), 241–250 (2004)
6. Martincic, F., Schwiebert, L.: Distributed event detection in sensor networks. In: Proceedings of Systems and Network Communication, pp. 43–48 (2006)
7. Ding, M., Chen, D., Xing, K., Cheng, X.: Localized fault tolerant event boundary detection in sensor networks. In: Proceesings of IEEE Conference of Computer and Communications Socities, pp. 902–913 (March 2005)

8. Silva, A.P.R., Martins, M.H.T., Rocha, B.P.S., Loureiro, A.A.F.: Decentralized intrusion detection in wireless sensor networks. In: Proceedings of the 1st ACM international Workshop on Quality of Service and Security in Wireless and Mobile Networks, pp. 16–23 (2005)
9. Bhuse, V., Gupta, A.: Anomaly intrusion detection in wireless sensor networks. Journal of High Speed Networks 15(1), 33–51 (2006)
10. Jurdak, R., Wang, X.R., Obst, O., Valencia, P.: Wireless Sensor Network Anomalies: Diagnosis and Detection Strategies. In: Tolk, A., Jain, L.C. (eds.) Intelligence-Based Systems Engineering. ISRL, vol. 10, pp. 309–325. Springer, Heidelberg (2011)
11. Buxton, H.: Learning and understanding dynamic scene activity: A review. Image and Vision Computing 21, 125–136 (2003)
12. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern., Appl. Rev. 34(3), 334–352 (2004)
13. Chandola, V., Banerjee, A., Kumar, V.: Outlier Detection: A Survey. ACM Computing Surveys, 1–72 (2009)
14. Zhang, Y., Meratnia, N., Havinga, P.: Outlier Detection Techniques for Wireless Sensor Networks: A Survey. IEEE Communication Surveys & Tutorials 12(2) (2010)
15. Gogoi, P., Borah, B., Bhattacharyya, D.K.: Anomaly Detection Analysis of Intrusion Data using Supervised and Unsupervised Approach. Journal of Convergence Information Technology 5(1) (February 2010)
16. Gogoi, P., Bhattacharyya, D.K., Borah, B., Kalita, J.K.: A Survey of Outlier Detection Methods in Network Anomaly Identification. The Computer Journal 54(4), 570–588 (2011)
17. Hawkins, D.M.: Identification of outliers. Chapman and Hall, London (1980)
18. Knorr, E.M., Ng, R.T.: Algorithm for mining distance based outliers in large datasets. In: Proceedings of the 24th International Conference on Very Large Databases, New York, USA, pp. 392–403. Morgan Kaufmann (1998)
19. Karl, H., Williz, A.: Protocols and Architectures for Wireless Sensor Networks. John Wiley & Sons (2007)
20. Branch, J.W., Giannelia, C., Szymanski, B., Wolff, R., Kargupta, H.: In-Network Outlier Detection in Wireless Sensor Networks. Knowledge and Information Systems 31 (2012)
21. Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979)
22. Teo, J.C.M., Tan, C.H.: Energy-Efficient and Scalable Group Key Agreement for Large Ad Hoc Networks. In: PE-WASUN's 2005, October 10-13, pp. 114–121 (2005)
23. Kumar, A., Aggarwal, A.: Efficient Hierarchical Threshold Symmetric Group Key Management Protocol for Mobile Ad Hoc Networks. In: IC3, pp. 335–346 (2012)
24. Cerpa, A., Elson, J., Estrin, D., Girod, L., Hamilton, M., Zhao, J.: Habitat Monitoring: Application Driver for Wireless Communication Technology. In: Proceedings of the ACM SIGCOMM Workshop on Data Communication in Latin America and the Caribbean, San Jose, Costa Rica (2001)
25. Traag, V.A., Browet, A., Calabrese, F., Morlot, F.: Social Event Detection in Massive Mobile Phone Data Using Probabilistic Location Inference. In: Traag, V.A., Browet, A., Calabrese, F., Morlot, F. (eds.) SocialCom/PASSAT, October 9-11, pp. 625–628 (2011)
26. NS3 Simulator, `http://www.nsnam.org`
27. AVISPA toolkit, `http://www.avispa-project.org`
28. ProVerif protocol verifier toolkit, `http://www.proverif.ens.fr`
29. Burmester, M., Desmedt, Y.G.: A secure and efficient conference key distribution system. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 275–286. Springer, Heidelberg (1995)

30. Yang, J., Wang, Y.: A New Outlier Detection Algorithms based on Markov chain. Advanced Materials Research 366, 456–459 (2012)
31. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying Density Based Local Outliers. In: Proceedings of the ACM SIGMOD Conference, Dallas, TX (May 2000)
32. Wang, B., Perrizo, W.: RDF: a density-based outlier detection method using vertical data representation. In: IEEE Int. Conference on Data Mining, pp. 503–506 (2004)
33. Rajagopalan, S., Karwoski, R., Bartholmai, B., Robb, R.: Quantitative image analytics for strtified pulmonary medicine. In: IEEE Int. Symposium on Biomedical Imaging (ISBI), pp. 1779–1782 (2012)
34. Cheminod, M., Bertolotti, I.C., Durante, L., Sisto, R., Valenzano, A.: Tools for cryptograhic protocols analysis: A technical and experimental comparison. Journal on Computer Standards & Interfaces 31(5), 954–961 (2009)

Image Secret Sharing in Stego-Images with Authentication

Amitava Nag[1], Sushanta Biswas[2], Debasree Sarkar[2], and Partha Pratim Sarkar[2]

[1] Academy of Technology
West Bengal University of Technology
Hoogly 721212 - India
amitavanag.09@gmail.com
[2] Départment of Engineering and Technological Studies
University of Kalyani
Kalyani 741 235 – India
ppsarkar@klyuniv.ac.in

Abstract. Recently, a polynomial-based (t,n) image sharing and hiding schemes with authentication was proposed to hid n shares of a secret image into n ordinary cover images and form n stego-images that can be transmitted securely. But each stego-image of all existing method should be expanded to 4 times of the secret image. In this paper we propose an enhanced scheme, where the size of the stego-image is reduced to $\frac{4(2n-t)}{n^2}$ times of the secret image. In addition our proposed scheme provides better authentication using hash function.

Keywords: Secret sharing, stego-image, authentication, hash function.

1 Introduction

With the rapid growth of Internet technologies as communication channel, digital media can be transmitted conveniently. But Internet is an open system; therefore, how to protect secret messages during transmission becomes an important issue. Two methods, cryptography and steganography have been used to protect secure data from malicious users on the internet. Cryptography transforms a secret data into disordered and meaningless form, which make suspicious enough to attract malicious users during transmission on the internet. The other method steganography is used to provide secure transmission by hiding a secret data into a cover media to generate gtego-media. Thus by steganography the observation of the existence of the embedded secret can be avoided. However, the weak point of steganography is that a secret message is protected in a single media carrier. Thus both cryptography and steganography are Single Point of Failure (SPOF) type as they use single storage mechanism. To overcome the weakness of Single Point of Failure (SPOF), several secret sharing techniques [1],[2],[3],[4],[5] have been proposed. It is a technique of protecting secret data like images by dividing secret data into n pieces (each piece is known as shadow share) and distributes the shares among a group of participants.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 764–770, 2013.

A secret sharing scheme is called a (t, n) threshold secret sharing scheme for t ≤ n if the following two conditions are satisfied: i) knowledge of any t or more shares can reconstruct the original secret; ii) knowledge of any t - 1 or less shares cannot recover the original secret. However, the shadow images produced by secret image sharing are noise-like, which may cause attacker's attention. From the view point of secure transmission, it is better if a (t,n) secret sharing and steganography are combined, where shadow images are embedded into cover images to form the stego-images. Moreover if authentication technique is added to detect integrity of shadow images, this scheme is called steganography and authentication based secret sharing. Some steganography and authentication based (t,n) secret sharing techniques were proposed in [6-9].

In [6], Lin and Tsai proposed a (t,n) secret sharing techniques with steganography and authentication to prevent fake stego-images. However, the authentication technique is too weak. Yang et al. [7] proposed an improved scheme that avoided Lin-Tsai's authentication weakness by hash function with secret key. But Yang et al's evaluate hash value for each pixel separately which lead a high computational cost of the authentication process. In [8], Chang et al. proposed sharing techniques with steganography and authentication scheme, where the concept of Chinese remainder theorem (CRT) is used to improve authentication ability. Eslami et al. in [9] proposed another method of secret sharing using cellular automata, where the author used double authentication to reduce the number of authentication bits. Main drawback of this method is that if all bits in stego-blocks are changed and the same eight shared bits are maintained, then tampere stego-blocks can not be located at receiver side. All these methods [6],[7],[8],[9] suffer from the problem that the size of the stego image is four times of the secret image. In this paper the size of cover images are reduced to $\frac{4(2n-t)}{n^2}$.

2 Related Work

We have already highlighted several (t,n) threshold-based Secret Sharing(SS) and (t, n)-based secret sharing in stego-images with authentication schemes. In this section we briefly describe one Secret Sharing and one secret sharing in stego-images with authentication schemes.

2.1 Lin and Wang's (k,n) Secret Image Sharing

In 2010, Lin and Wnag proposed a scalable (k,n) 2 ≤ t ≤ n secret image sharing scheme[5]. Their share generation process involves three steps:

Step 1. The secret image G is partitioned into n disjoint set of image partitions $\{P_1,P_2,\ldots,P_n\}$, such that

$$\bigcup_i P_i = G, \ \ for\, 1 \le i \le n$$

$$P_i \bigcap P_j = \emptyset, for\ 1 \le i \ne j \le n$$
$$|P_i| = \frac{1}{n}|G|, for\ 1 \le i \le n$$

where | . | denotes the size. The authors applied the three sharing modes 1) multisecret, 2) priority, and 3) progressive mode, in their sharing scheme to execute different reconstructing effects.

Step 2. Each image partition P$_i$ $(1 \le i \le n)$ is further divided into (2n – t) share images {L$_1$,L$_2$,….,L$_n$}, using Thien-Lin (n,2n – t) secret image sharing technique[3].

Step 3. The n share images, referred to as S$_i$, i $(1 \le i \le n)$ are generated as:

$$S_i = \bigcup_j L_{jk},\ (1 \le j \le n)\quad and$$

(a) If j = i, then j$\le k \le j + n - k$
(b) If j > i, then k = i
(c) If j < i, then k = i + n – k

In this (t,n) sharing method, the size of each generated shadow image is $\frac{4(2n-t)}{n^2}$ times of that the original image.

2.2 Review of Chang et al's Sharing Secrets in Stego-Images with Authentication Scheme

The main shortcoming of the steganography and authentication based secret sharing [6][7] is that the weak authentication cannot well protect the integrity of the stego-images and thus complete recovery of secret image is not possible. To overcome this drawback, in [8] Chang et al. propose a (t,n) secret sharing technique by combining LSB-based steganography and Chinese remainder theorem (CRT) based authentication together. In [8], the authors first computes four authentication bits using CRT method. Then they combine these four bits with watermark bits and produce four parity bits (p$_1$,p$_2$,p$_3$,p$_4$). Then stego-block is produced with modified pixels $\overline{W_k}$, $\overline{X_k}$, $\overline{Y_k}$ and $\overline{Z_k}$ as follows :

$$\begin{cases} \overline{W_k} = \left(w_7w_6w_5w_4w_3\boxed{\overline{w}_2\overline{w}_1\overline{w}_0}\right) = \left(w_7w_6w_5w_4w_3\boxed{s_1s_2p_1}\right) \\ \overline{X_k} = \left(x_7x_6x_5x_4x_3\boxed{\overline{x}_2\overline{x}_1\overline{x}_0}\right) = \left(x_7x_6x_5x_4x_3\boxed{s_3s_4p_2}\right) \\ \overline{Y_k} = \left(y_7y_6y_5y_4y_3\boxed{\overline{y}_2\overline{y}_1\overline{y}_0}\right) = \left(y_7y_6y_5y_4y_3\boxed{s_5s_6p_3}\right) \\ \overline{Z_k} = \left(z_7z_6z_5z_4z_3\boxed{\overline{z}_2\overline{z}_1\overline{z}_0}\right) = \left(z_7z_6z_5z_4z_3\boxed{s_7s_8p_4}\right) \end{cases}$$

3 The Proposed Scheme

The proposed scheme consists of two main phases: the first phase is sharing and embedding and the second phase is authentication and recovery.

A. The sharing and embedding phase

In sharing phase, the secret image is shared into n shadow images in a (t,n), $2 \leq t \leq n$, scalable image sharing manner in progressive mode by Lin and Wang's technique[5]. Now generated shadow images are embedded into cover image. Before embedding the cover image I of size $M \times N$ is divided into several sections of size 10×16 pixels. Each section is then further subdivided into 40 cover blocks $B_1,B_2,\dots\dots..,B_{40}$ of size 2×2 pixels as $B_k=\{W_k,X_k,Y_k,Z_k\}$, where $W_k=(w_7,w_6,\dots\dots.w_0)$, $X_k=(x_7,x_6\dots\dots x_0)$, $Y_k=(y_7,y_6,\dots\dots.y_0)$ and $Z_k=(z_7,z_6,\dots.z_0)$. Let S be a shadow pixel to be embedded in B_k block, whose binary representation is $(s_7,s_6,\dots\dots\dots s_0)$, then this secret bits are inserted into cover pixels $B_k=\{W_k,X_k,Y_k,Z_k\}$ and producing stego block $\overline{B_k}$ with pixels $\overline{W_k}$, $\overline{X_k}$, $\overline{Y_k}$ and $\overline{Z_k}$ as follows :

$$\begin{cases} \overline{W_k} = \left(w_7w_6w_5w_4w_3w_2\boxed{\overline{w}_1\overline{w}_0}\right) = \left(w_7w_6w_5w_4w_3w_2\boxed{s_7s_6}\right) \\ \overline{X_k} = \left(x_7x_6x_5x_4x_3x_2\boxed{\overline{x}_1\overline{x}_0}\right) = \left(x_7x_6x_5x_4x_3x_2\boxed{s_5s_4}\right) \\ \overline{Y_k} = \left(y_7y_6y_5y_4y_3y_2\boxed{\overline{y}_1\overline{y}_0}\right) = \left(y_7y_6y_5y_4y_3y_2\boxed{s_3s_2}\right) \\ \overline{Z_k} = \left(z_7z_6z_5z_4z_3z_2\boxed{\overline{z}_1\overline{z}_0}\right) = \left(z_7z_6z_5z_4z_3z_2\boxed{s_1s_0}\right) \end{cases}$$

To prevent the manipulation of shadow images from malicious users, a check bits stream is needed. In our proposed scheme, SHA 1 hash function is used to generate the authentication bits. The MSB of all pixels (160) of all blocks in all section are used as watermark bits (160 watermark bits). Now according to the 160 SHA-1 based authentication bits and 160 current watermark bits, 160 check bits are calculated i.e. 4 check bits per block are generated. The authentication bits of each section are evaluated as follows:

$$\begin{aligned} &(a_{159}a_{158}\dots\dots\ a_1a_0) \\ &\quad = SHA1\Big(\left(\overline{W}_{39} - c_{159}\right)\Big|\Big|\left(\overline{X}_{39} - c_{158}\right)\Big|\Big|\left(\overline{Y}_{39} - c_{157}\right)||\Big(\overline{Z}_{39} \\ &\quad - c_{156}\Big)||\ \dots||\left(\overline{W}_0 - c_3\right)||\left(\overline{X}_0 - c_2\right)||\left(\overline{Y}_0 - c_1\right)||\Big(\overline{Z}_0 \\ &\quad - c_0\Big)\Big)\ \dots\dots\dots\dots(1) \end{aligned}$$

Where $\left(\overline{W}_k - c_i\right)$ represents 7 bits exclusive the check bit c_i and "||" represent the concatenation operation. Now check bits $c_{159},\dots\dots..c_1,c_0$ are computed by

$$(c_{159}c_{158} \dots\dots c_1c_0) = \left(MSB(\overline{W}_{39})\ MSB(\overline{X}_{39})MSB(\overline{Y}_{39})MSB(\overline{Z}_{39}) \dots MSB(\overline{W}_0)\ MSB(\overline{X}_0)MSB(\overline{Y}_0)MSB(\overline{Z}_0)\right) \text{ XOR } (a_{159}a_{158} \dots\dots a_1a_0) \dots\dots\dots\dots\dots (2)$$

Finally, the proposed scheme replaces the 3rd LSB of cover pixels for example w_2,x_2,y_2,z_2 with the computed check bits c_3,c_2,c_1,c_0.

B. Authentication and Recovery Phase

In order to generate the secret image, any t or more number of stego images are gathered together. After that, each stego-image is divided into several sections of size 10×16 pixels and each section once again subdivided into 40 blocks of size 2×2 pixels. For each section, the hash value is evaluated using (1) and check bits $(\overline{c}_{159}, \dots, \overline{c}_1, \overline{c}_0)$ of the current section are generated using (2). Now extract the 160 3rd LSB (c_{159},……..c_1,c_0) from each stegopixels of the current section. If the computed check bits $(\overline{c}_{159}, \dots, \overline{c}_1, \overline{c}_0)$ are matched to those extracted bits (c_{159},……..c_1,c_0), the current section is verified successfully. Now one shadow pixel is extracted from the 2 LSB of the stego-pixels of each 2×2 stego-block. In this way, 40 shadow pixels are correctly extracted from the current verified section. Otherwise, the stego-image has been manipulated by malicious users. Thus when any t or more shadows are extracted from the stego-images, the original secret image is recovered using Lin and Wang's[5] reconstructed algorithm.

4 Experimental Results

This section presents the experimental results of the proposed scheme. The stego-images visual quality is evaluated by the Peak Signal to Noise Ratio (PSNR). The definition of PSNR is given below

$$\mathrm{PSNR(dB)} = 20\log_{10}\frac{255}{\sqrt{\mathrm{MSE}}} \tag{12}$$

MSE is the mean squared error between the original image and the modified image which is defined as

$$\mathrm{MSE} = \frac{1}{\mathrm{M{\times}N}}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(I(x,y) - I'(x,y)\right)^2 \tag{13}$$

where M and N denotes the width and height of the cover and stego image respectively. In our experiment we used (2,3) secret sharing. The image Airplane shown in figure 1 of size 256×333 pixels is chosen as secret image and three cover image Lena, pepper and baboon with 512×300 pixels are chosen as cover image as shown in figure 2. Table 1 shows the PSNR values of the stego-images.

Fig. 1. The secret image - Airplane

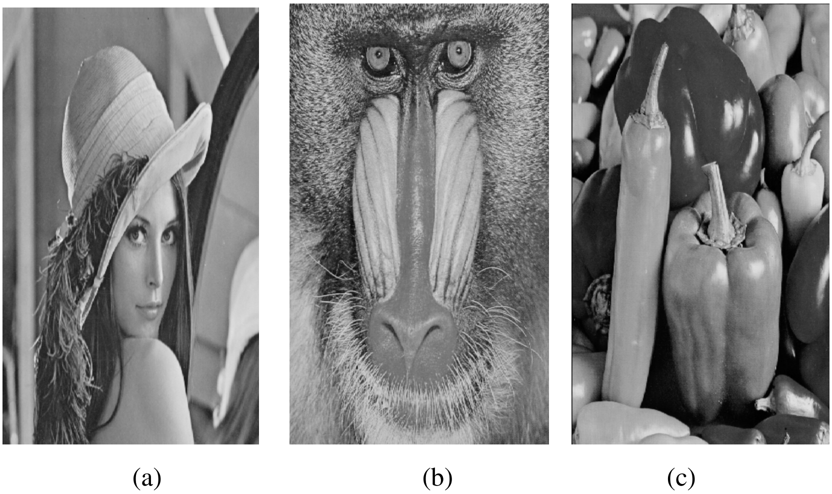


Fig. 2. Cover images (a) lena, (b) pepper, (c) baboon

Table 1. The experimental results for four (2,3) secret image sharing schemes

Images	Lin et al.'s schemes	Yang et al.'s schemes	Chang et al.'s schemes	The proposed scheme
Lena	39.21	36.20	40.97	40.9718
Pepper	39.17	36.17	40.96	40.9701
Baboon	39.18	36.19	40.93	40.9521

In addition we also used detection ratio(DR) for integrity verification. The detection ratio(DR) is defined as follows:

$$\text{detection ratio(DR)} = \frac{\text{Number of the tampered pixels (NTP)}}{\text{Number of the tampered pixels that are detected(NTPD)}}$$

The detection ration (DR) in Lin et al.'s, Yang et al.'s, Chang et al.'s , Eslami et al's and the proposed methods are 0, 0.51,0.97,0 and 0.97 respectively as given in table 2. From table 2 it is also clear that our proposed method reduces the size of expansion of stego-image and provides better authentication.

Table 2. Comparison of (k,n) secret image sharing schemes

	Lin et al.'s schemes	Yang et al.'s schemes	Chang et al.'s schemes	Eslami et al's scheme	The proposed scheme
Expansion of stego-image	4	4	4	4	$\frac{4(2n-t)}{n^2}$
Detection Ratio (DR)	0	0.51	0.97	0	0.97
The size of detection unit	-	Block of size 2×2 pixels	Block of size 2×2 pixels	-	Block of size 10×16 pixels
Authentication bits: pixel	1	1:4	1:1	0	1:1

5 Conclusion

In this paper we propose an steganography and authentication based secret sharing technique. Compared with other existing method, the size of each stegoimage is only $\frac{4(2n-t)}{n^2}$ times of the size of the secret image. This small size helps in both storage and transmission. Moreover the authentication is implemented by SHA1 which enhance the authentication ability.

References

1. Blakely, G.R.: Safeguarding cryptography keys. In: Proc. of AFIPS National Computer Conference, vol. 48, pp. 313–317 (1979)
2. Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979)
3. Thien, C.C., Lin, J.C.: Secret image sharing. Computer Graphics 26(5), 765–770 (2002)
4. Thien, C.C., Lin, J.C.: An image-sharing method with user-friendly shadow images. IEEE Transactions on Circuit System 13(12), 1161–1169 (2003)
5. Lin, Y.Y., Wang, R.Z.: Scalable Secret Image Sharing with Smaller Shadow Images. IEEE Signal Processing Letters 17(3), 316–319 (2010)
6. Lin, C.C., Tsai, W.H.: Secret Image sharing with steganography and authentication. Journal of Systems and Software 73, 405–414 (2004)
7. Yang, C.N., Chen, T.S., Yu, K.H., Wang, C.C.: Improvements of image sharing with steganography and authentication. Journal of Systems and Software 80, 1070–1076 (2007)
8. Chang, C.C., Hsieh, Y.P., Lin, C.H.: Sharing secrets in stego images with authentication. Pattern Recognition 41, 3130–3137 (2008)
9. Eslami, Z., Razzaghi, S.H., Ahmadabadi, J.Z.: Secret image sharing based on cellular automata and steganography. Pattern Recognition 43(1), 397–404 (2010)

On Generalized Nega–Hadamard Transform

Ankita Chaturvedi and Aditi Kar Gangopadhyay

Department of Mathematics
Indian Institute of Technology Roorkee
Roorkee–247667, India
{ankitac17,ganguli.aditi}@gmail.com

Abstract. In this paper, we consider generalized Boolean functions from $\mathbb{Z}_2^n$ to $\mathbb{Z}_q$ ($q \geq 2$, a positive integer). Here, we present some of the properties of generalized nega–Hadamard transform which are analogous to nega–Hadamard transform. Further, it is shown that if we represent a generalized Boolean function in terms of Boolean functions then there is a relation between their nega–Hadamard transforms.

Keywords: Generalized Boolean functions, nega–Hadamard transform, generalized nega–Hadamard transform.

1 Introduction

In the recent years, several generalizations of Boolean functions have been proposed and the effect of Walsh–Hadamard transform on them has been studied by various authors [3,9,14,15,11]. The nega-Hadamard transforms and negabent functions have been discussed [6,7,8,10,12,13]. Like the Boolean case, in the generalized setup, the functions which have flat spectra with respect to nega–Hadamard transform are said to be generalized negabent functions.

Let us list the notations:
$\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ denote the set of integers, real numbers and complex numbers respectively;
q and n are positive integers;
'$+'$ denotes the addition modulo q;
'$\oplus'$ denotes the addition modulo 2;
$\mathbb{Z}_2$ is the prime field of order 2;
$\mathbb{Z}_2^n$ is the n-dimensional vector space over field $\mathbb{Z}_2$;
$\mathbb{Z}_q$ is the ring of integers modulo q;
$\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is an element of $\mathbb{Z}_2^n$;
$< \mathbf{x}, \mathbf{y} >= x_1y_1 \oplus x_2y_2 \oplus \cdots \oplus x_ny_n$ is the inner product of vectors;
$\mathbf{x} * \mathbf{y} = (x_1y_1, \ldots, x_ny_n)$ is the intersection of two vectors.

The cardinality of the set S is denoted by $|S|$. If $z = a + b\imath \in \mathbb{C}$, then $|z| = \sqrt{a^2 + b^2}$ denotes the absolute value of z, and $\overline{z} = a - b\imath$ denotes the complex

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 771–777, 2013.

conjugate of z, where $\imath^2 = -1$, and $a, b \in \mathbb{R}$. The conjugate of a bit b will also be denoted by $\bar{b}$.

A function from $\mathbb{Z}_2^n$ to $\mathbb{Z}_2$ is said to be a Boolean function on n variables and the set of all such functions is denoted by $\mathcal{B}_n$. For more details on Boolean functions one may refer to [1,2,4]. A function from $\mathbb{Z}_2^n$ to $\mathbb{Z}_q$ ($q \geq 2$, a positive integer) is said to be a *generalized Boolean function* on n variables. In this paper, we are considering the generalization of Schmidt [9]. We denote the set of all such functions by $\mathcal{B}_n^q$.
Some basic definitions are given below:

Definition 1. *The (generalized)* Walsh–Hadamard transform *of $f \in \mathcal{B}_n^q$ at any point $\mathbf{u} \in \mathbb{Z}_2^n$ is the complex valued function defined by*

$$\mathcal{H}_f(\mathbf{u}) = 2^{-\frac{n}{2}} \sum_{\mathbf{x} \in \mathbb{Z}_2^n} \zeta^{f(\mathbf{x})} (-1)^{\mathbf{u} \cdot \mathbf{x}},$$

where $\zeta = e^{\frac{2\pi\imath}{q}}$ is the q-th primitive root of unity. A function $f \in \mathcal{B}_n^q$ is said to be generalized bent *if and only if $|\mathcal{H}_f(\mathbf{u})| = 1$ for all $\mathbf{u} \in \mathbb{Z}_2^n$.*

Definition 2. *The* nega–Hadamard transform *of $f \in \mathcal{B}_n$ at any vector $\mathbf{u} \in \mathbb{Z}_2^n$ is the complex valued function*

$$\mathcal{N}_f(\mathbf{u}) = 2^{-\frac{n}{2}} \sum_{\mathbf{x} \in \mathbb{Z}_2^n} (-1)^{f(\mathbf{x}) \oplus \mathbf{u} \cdot \mathbf{x}} \, \imath^{wt(\mathbf{x})}.$$

A function $f \in \mathcal{B}_n$ is said to be negabent *if and only if $|\mathcal{N}_f(\mathbf{u})| = 1$ for all $\mathbf{u} \in \mathbb{Z}_2^n$.*

We define generalized nega–Hadamard transform and generalized negabent function in the following manner:

Definition 3. *The* generalized nega–Hadamard transform *of $f \in \mathcal{B}_n^q$ at any point $\mathbf{u} \in \mathbb{Z}_2^n$ is defined by*

$$\mathcal{N}_f^q(\mathbf{u}) = 2^{-\frac{n}{2}} \sum_{\mathbf{x} \in \mathbb{Z}_2^n} \zeta^{f(\mathbf{x})} (-1)^{\mathbf{u} \cdot \mathbf{x}} \imath^{wt(\mathbf{x})}.$$

Definition 4. *A function $f \in \mathcal{B}_n^q$ is a* generalized negabent function *if $|\mathcal{N}_f^q(\mathbf{u})| = 1$ for all $\mathbf{u} \in \mathbb{Z}_2^n$.*

We recall the following result:

Lemma 1 ([10], Lemma 1). *For any $\mathbf{u} \in \mathbb{F}_2^n$, we have*

$$\sum_{\mathbf{x} \in \mathbb{F}_2^n} (-1)^{\mathbf{u} \cdot \mathbf{x}} \imath^{wt(\mathbf{x})} = 2^{\frac{n}{2}} \, \omega^n \, \imath^{-wt(\mathbf{u})}, \tag{1}$$

where $\omega = \frac{1+\imath}{\sqrt{2}}$ is a primitive 8th root of unity.

We shall also use the following well-known identities

$$wt(\mathbf{x} \oplus \mathbf{y}) = wt(\mathbf{x}) + wt(\mathbf{y}) - 2wt(\mathbf{x} * \mathbf{y}) \qquad \text{(given in [5])} \tag{2}$$

and

$$\sum_{\mathbf{x}} (-1)^{\mathbf{v}\cdot\mathbf{x}} = \begin{cases} 2^n & \text{if } \mathbf{v} = \mathbf{0} \\ 0 & \text{if } \mathbf{v} \neq \mathbf{0}, \end{cases} \qquad \text{(see [2, p.8])} \tag{3}$$

In this paper, we prove various results in Section 2 on the behavior of the generalized nega–Hadamard transform on affine functions and sums of functions.

In Section 3, we show that if we represent a generalized Boolean function in terms of Boolean functions then there is a relation between their nega–Hadamard transform.

2 Properties of Generalized Nega–Hadamard Transform

Stănică et al. [12,13] investigated various properties of nega–Hadamard transform.

The following theorem gives the generalized nega–Hadamard transform of various combinations of generalized Boolean functions.

Theorem 1. *Let f, g, h be in $\mathcal{B}_n^q$. The following statements are true:*

(a) *For any affine function $\ell_{\mathbf{a},c}(\mathbf{x}) = \left(\frac{q}{2}\right)\mathbf{a}\cdot\mathbf{x} + c$ and $f \in \mathcal{B}_n^q$, $\mathcal{N}_{f+\ell_{\mathbf{a},c}}^q(\mathbf{u}) = \zeta^c \mathcal{N}_f^q(\mathbf{a} \oplus \mathbf{u})$. Further, $\mathcal{N}_{\ell_{\mathbf{a},c}}^q(\mathbf{u}) = \zeta^c \omega^n \imath^{-wt(\mathbf{a}\oplus\mathbf{u})}$.*

(b) *If $h(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ on $\mathbb{Z}_2^n$, then for $\mathbf{u} \in \mathbb{Z}_2^n$,*

$$\mathcal{N}_h^q(\mathbf{u}) = 2^{-\frac{n}{2}} \sum_{\mathbf{v}\in\mathbb{Z}_2^n} \mathcal{N}_f^q(\mathbf{v})\mathcal{H}_g(\mathbf{u}\oplus\mathbf{v}) = 2^{-\frac{n}{2}} \sum_{\mathbf{v}\in\mathbb{Z}_2^n} \mathcal{H}_f(\mathbf{v})\mathcal{N}_g^q(\mathbf{u}\oplus\mathbf{v}).$$

(c) *If $h(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{Z}_2^n$, then $\mathcal{N}_{f+g}^q(\mathbf{u}, \mathbf{v}) = \mathcal{N}_f^q(\mathbf{u})\mathcal{N}_g^q(\mathbf{v})$.*

(d) *If $h(\mathbf{x}) = f(A\mathbf{x} \oplus \mathbf{a})$, then $\mathcal{N}_h^q(\mathbf{u}) = (-1)^{\mathbf{a}\cdot(A\mathbf{u})} \imath^{wt(\mathbf{a})} \mathcal{N}_f^q(A\mathbf{u} \oplus \mathbf{a})$, where A is an $n \times n$ orthogonal matrix over $\mathbb{Z}_2$ (and so, $A^T A = I_n$).*

Proof. (a) We obtain

$$\begin{aligned}
\mathcal{N}_{f+\ell_{\mathbf{a},c}}^q(\mathbf{u}) &= 2^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{(f+\ell_{\mathbf{a},c})(\mathbf{x})}(-1)^{\mathbf{u}\cdot\mathbf{x}} \imath^{wt(\mathbf{x})} \\
&= 2^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{f(\mathbf{x})+\left(\frac{q}{2}\right)\mathbf{a}\cdot\mathbf{x}+c} (-1)^{\mathbf{u}\cdot\mathbf{x}} \imath^{wt(\mathbf{x})} \\
&= 2^{-\frac{n}{2}} \zeta^c \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{f(\mathbf{x})}(-1)^{(\mathbf{a}\oplus\mathbf{u})\cdot\mathbf{x}} \imath^{wt(\mathbf{x})} \\
&= \zeta^c \mathcal{N}_f^q(\mathbf{a}\oplus\mathbf{u}).
\end{aligned}$$

Further,

$$\begin{aligned}
\mathcal{N}^q_{\ell_{\mathbf{a},c}}(\mathbf{u}) &= 2^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{\ell_{\mathbf{a},c}(\mathbf{x})}(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \\
&= 2^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{\frac{q}{2}(\mathbf{a}\cdot\mathbf{x})+c}\,(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \\
\mathcal{N}^q_{\ell_{\mathbf{a},c}}(\mathbf{u}) &= 2^{-\frac{n}{2}}\,\zeta^c \sum_{\mathbf{x}\in\mathbb{Z}_2^n} (-1)^{(\mathbf{a}\oplus\mathbf{u})\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \\
&= \zeta^c\,\omega^n\,\imath^{-wt(\mathbf{a}\oplus\mathbf{u})} \qquad \text{(using (1))}.
\end{aligned}$$

Here we show the first identity of (b). Since

$$\begin{aligned}
\mathcal{N}^q_f(\mathbf{v}) &= 2^{-\frac{n}{2}} \sum_{\mathbf{x}\in\mathbb{Z}_2^n} \zeta^{f(\mathbf{x})}\,(-1)^{\mathbf{v}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \\
\mathcal{H}_g(\mathbf{u}\oplus\mathbf{v}) &= 2^{-\frac{n}{2}} \sum_{\mathbf{y}\in\mathbb{Z}_2^n} \zeta^{g(\mathbf{y})}\,(-1)^{\mathbf{y}\cdot(\mathbf{u}\oplus\mathbf{v})}
\end{aligned}$$

we obtain (sums are over $\mathbb{Z}_2^n$)

$$\begin{aligned}
\sum_{\mathbf{v}} \mathcal{N}^q_f(\mathbf{v})\mathcal{H}_g(\mathbf{u}\oplus\mathbf{v}) &= 2^{-n} \sum_{\mathbf{v},\mathbf{x},\mathbf{y}} \zeta^{f(\mathbf{x})+g(\mathbf{y})}\,(-1)^{\mathbf{v}\cdot\mathbf{x}\oplus\mathbf{y}\cdot(\mathbf{u}\oplus\mathbf{v})}\,\imath^{wt(\mathbf{x})} \\
&= 2^{-n} \sum_{\mathbf{x},\mathbf{y}} \zeta^{f(\mathbf{x})+g(\mathbf{y})}\,(-1)^{\mathbf{u}\cdot\mathbf{y}}\,\imath^{wt(\mathbf{x})} \sum_{\mathbf{v}} (-1)^{\mathbf{v}\cdot(\mathbf{x}\oplus\mathbf{y})} \\
&= \sum_{\mathbf{x}} \zeta^{f(\mathbf{x})+g(\mathbf{x})}\,(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \qquad \text{(using (3))} \\
&= 2^{\frac{n}{2}} \mathcal{N}^q_{f+g}(\mathbf{u}).
\end{aligned}$$

Similarly we can prove second identity.

(c) If $h(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) + g(\mathbf{y})$, where $\mathbf{x}$, $\mathbf{y} \in \mathbb{F}_2^n$.
We obtain (sums are over $\mathbb{Z}_2^n$)

$$\begin{aligned}
\mathcal{N}^q_h(\mathbf{u},\mathbf{v}) &= 2^{-n} \sum_{\mathbf{x},\mathbf{y}} \zeta^{h(\mathbf{x},\mathbf{y})}(-1)^{\mathbf{u}\cdot\mathbf{x}\oplus\mathbf{v}\cdot\mathbf{y}}\,\imath^{wt(\mathbf{x})+wt(\mathbf{y})} \\
&= 2^{-\frac{n}{2}} \sum_{\mathbf{x}} \zeta^{f(\mathbf{x})}(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})}\, 2^{-\frac{n}{2}} \sum_{\mathbf{y}} \zeta^{g(\mathbf{y})}(-1)^{\mathbf{v}\cdot\mathbf{y}}\,\imath^{wt(\mathbf{y})} \\
&= \mathcal{N}^q_f(\mathbf{u})\,\mathcal{N}^q_g(\mathbf{v}).
\end{aligned}$$

To show (d), we compute, for $h(\mathbf{x}) = f(A\mathbf{x}\oplus\mathbf{a})$, where A is an $n\times n$ orthogonal matrix over $\mathbb{Z}_2$.
We compute (sums are over $\mathbb{Z}_2^n$)

$$\begin{aligned}
\mathcal{N}^q_h(\mathbf{u}) &= 2^{-\frac{n}{2}} \sum_{\mathbf{x}} \zeta^{h(\mathbf{x})}\,(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})} \\
&= 2^{-\frac{n}{2}} \sum_{\mathbf{x}} \zeta^{f(A\mathbf{x}\oplus\mathbf{a})}\,(-1)^{\mathbf{u}\cdot\mathbf{x}}\,\imath^{wt(\mathbf{x})}
\end{aligned}$$

$$
\begin{aligned}
&= 2^{-\frac{n}{2}} \sum_{\mathbf{y}} \zeta^{f(\mathbf{y})} \, (-1)^{\mathbf{u}\cdot A^T(\mathbf{y}\oplus\mathbf{a})} \, \imath^{wt(A^T(\mathbf{y}\oplus\mathbf{a}))} \\
&= 2^{-\frac{n}{2}} \sum_{\mathbf{y}} \zeta^{f(\mathbf{y})} \, (-1)^{A\mathbf{u}\cdot\mathbf{y}+A\mathbf{u}\cdot\mathbf{a}} \, \imath^{wt(\mathbf{y})+wt(\mathbf{a})-2wt(\mathbf{y}*\mathbf{a})} \qquad \text{(using (2))} \\
\mathcal{N}_h^q(\mathbf{u}) &= 2^{-\frac{n}{2}} (-1)^{A\mathbf{u}\cdot\mathbf{a}} \, \imath^{wt(\mathbf{a})} \sum_{\mathbf{y}} \zeta^{f(\mathbf{y})} \, (-1)^{(A\mathbf{u}\oplus\mathbf{a})\cdot\mathbf{y}} \, \imath^{wt(\mathbf{y})} \\
&= (-1)^{A\mathbf{u}\cdot\mathbf{a}} \, \imath^{wt(\mathbf{a})} \mathcal{N}_f^q(A\mathbf{u}\oplus\mathbf{a}).
\end{aligned}
$$

Since $\imath^{2wt(\mathbf{y}*\mathbf{a})} = (-1)^{\mathbf{y}\cdot\mathbf{a}}$ and $wt(A^T(\mathbf{y}\oplus\mathbf{a})) = wt(\mathbf{y}\oplus\mathbf{a})$ (here we needed the orthogonality of A, since it preserves vectors lengths).

3 Relation between Generalized Nega–Hadamard Transform and Nega–Hadamard Transform

In this section, the generalized Boolean function $f : \mathbb{Z}_2^{2n} \to \mathbb{Z}_4$ is considered.

Generalized Boolean function can be represented by combination of Boolean functions. In the following theorem, it is shown that the generalized nega–Hadamard transform can be derived from the nega–Hadamard transform of Boolean functions.

Theorem 2. *Let $f : \mathbb{Z}_2^{2n} \to \mathbb{Z}_4$ be any generalized Boolean function. Represented it as $f(\mathbf{x},\mathbf{y}) = a(\mathbf{x},\mathbf{y}) + 2b(\mathbf{x},\mathbf{y})$; for any $\mathbf{x},\mathbf{y} \in \mathbb{Z}_2^n$, where $a, b : \mathbb{Z}_2^{2n} \to \mathbb{Z}_2$ are Boolean functions. Between nega–Hadamard transforms of $f, a+b, b$ there is the relation*

$$
\mathcal{N}_f^4 = \frac{1}{2}\left[\mathcal{N}_b(\mathbf{u},\mathbf{v}) + \mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})\right] + \frac{i}{2}\left[\mathcal{N}_b(\mathbf{u},\mathbf{v}) - \mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})\right]
$$

and

$$
|\mathcal{N}_f^4(u,v)|^2 = \frac{1}{2}|\mathcal{N}_b(\mathbf{u},\mathbf{v}) - \imath\mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})|^2
$$

Proof.

$$
\begin{aligned}
\mathcal{N}_f^4(u,v) &= 2^{-n} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} \imath^{a(\mathbf{x},\mathbf{y})+2b(\mathbf{x},\mathbf{y})} \, (-1)^{\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}} \, \imath^{wt(\mathbf{x},\mathbf{y})} \\
&= 2^{-n} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} \imath^{a(\mathbf{x},\mathbf{y})} \, (-1)^{\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}+b(\mathbf{x},\mathbf{y})} \, \imath^{wt(\mathbf{x},\mathbf{y})}
\end{aligned}
$$

Applying the formula $\imath^s = \frac{1+(-1)^s}{2} + \left(\frac{1-(-1)^s}{2}\right)\imath$,
for $s = a(\mathbf{x},\mathbf{y})$, We have

$$
\mathcal{N}_f^4(u,v) = \frac{2^{-n}}{2} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} \left[1+(-1)^{a(\mathbf{x},\mathbf{y})} + \imath\left(1-(-1)^{a(\mathbf{x},\mathbf{y})}\right)\right] \, (-1)^{\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}+b(\mathbf{x},\mathbf{y})} \, \imath^{wt(\mathbf{x},\mathbf{y})}
$$

$$
\begin{aligned}
&= 2^{-(n+1)} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} (-1)^{b(\mathbf{x},\mathbf{y})+\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}}\, \imath^{wt(\mathbf{x},\mathbf{y})} \\
&\quad +2^{-(n+1)} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} (-1)^{a(\mathbf{x},\mathbf{y})+b(\mathbf{x},\mathbf{y})+\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}}\, \imath^{wt(\mathbf{x},\mathbf{y})} \\
&\quad +\imath\, 2^{-(n+1)} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} (-1)^{b(\mathbf{x},\mathbf{y})+\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}}\, \imath^{wt(\mathbf{x},\mathbf{y})} \\
&\quad -\imath\, 2^{-(n+1)} \sum_{\mathbf{x},\mathbf{y}\in\mathbb{Z}_2^n} (-1)^{a(\mathbf{x},\mathbf{y})+b(\mathbf{x},\mathbf{y})+\mathbf{u}\cdot\mathbf{x}+\mathbf{v}\cdot\mathbf{y}}\, \imath^{wt(\mathbf{x},\mathbf{y})}
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{N}_f^4(u,v) &= \frac{1}{2}\left[\mathcal{N}_b(\mathbf{u},\mathbf{v}) + \mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})\right] + \frac{\imath}{2}\left[\mathcal{N}_b(\mathbf{u},\mathbf{v}) - \mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})\right] \\
|\mathcal{N}_f^4(u,v)|^2 &= \frac{1}{2}\left(|\mathcal{N}_b(\mathbf{u},\mathbf{v})|^2 + |\mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})|^2\right) \\
&\quad + \frac{\imath}{2}\left(\mathcal{N}_b(\mathbf{u},\mathbf{v})\overline{\mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})} - \mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})\overline{\mathcal{N}_b(\mathbf{u},\mathbf{v})}\right) \\
|\mathcal{N}_f^4(u,v)|^2 &= \frac{1}{2}|\mathcal{N}_b(\mathbf{u},\mathbf{v}) - \imath\mathcal{N}_{a+b}(\mathbf{u},\mathbf{v})|^2
\end{aligned}
$$

4 Conclusion

In this paper, we have investigated some properties of generalized nega–Hadamard transfom. Moreover, it is shown that if a generalized Boolean function is represented by the combination of Boolean functions, there is the relation between their nega–Hadamard transform.

References

1. Carlet, C.: Boolean functions for cryptography and error correcting codes. In: Crama, Y., Hammer, P. (eds.) Boolean Methods and Models. Cambridge Univ. Press, Cambridge, `http://www-roc.inria.fr/secret/Claude.Carlet/pubs.html`
2. Cusick, T.W., Stănică, P.: Cryptographic Boolean functions and Applications. Elsevier–Academic Press (2009)
3. Kumar, P.V., Scholtz, R.A., Welch, L.R.: Generalized bent functions and their properties. Journal of Combinatorial Theory (Series A) 40(1), 90–107 (1985)
4. Lidl, R., Niederreiter, H.: Introduction to finite fields and their applications. Cambridge University Press (1983)
5. MacWilliams, F.J., Sloane, N.J.A.: The theory of error–correcting codes. North-Holland, Amsterdam (1977)
6. Parker, M.G., Pott, A.: On Boolean functions which are bent and negabent. In: Golomb, S.W., Gong, G., Helleseth, T., Song, H.-Y. (eds.) SSC 2007. LNCS, vol. 4893, pp. 9–23. Springer, Heidelberg (2007)
7. Riera, C., Parker, M.G.: One and two-variable interlace polynomials: A spectral interpretation. In: Ytrehus, Ø. (ed.) WCC 2005. LNCS, vol. 3969, pp. 397–411. Springer, Heidelberg (2006)
8. Riera, C., Parker, M.G.: Generalized bent criteria for Boolean functions. IEEE Trans. Inform. Theory 52(9), 4142–4159 (2006)
9. Schmidt, K.-U.: Quaternary Constant-Amplitude Codes for Multicode CDMA. In: IEEE International Symposium on Information Theory, pp. 2781–2785 (2007), `http://arxiv.org/abs/cs.IT/0611162`

10. Schmidt, K.-U., Parker, M.G., Pott, A.: Negabent functions in the maiorana–mcFarland class. In: Golomb, S.W., Parker, M.G., Pott, A., Winterhof, A. (eds.) SETA 2008. LNCS, vol. 5203, pp. 390–402. Springer, Heidelberg (2008)
11. Solé, P., Tokareva, N.: Connections between Quaternary and Binary Bent Functions, `http://eprint.iacr.org/2009/544.pdf`
12. Stănică, P., Gangopadhyay, S., Chaturvedi, A., Gangopadhyay, A.K., Maitra, S.: Nega–Hadamard transform, bent and negabent functions. In: Carlet, C., Pott, A. (eds.) SETA 2010. LNCS, vol. 6338, pp. 359–372. Springer, Heidelberg (2010)
13. Stănică, P., Gangopadhyay, S., Chaturvedi, A., Gangopadhyay, A.K., Maitra, S.: Investigations on bent and negabent functions via the nega–Hadamard transform. IEEE Transactions on Information Theory 58(6), 4064–4072 (2012)
14. Stănică, P., Gangopadhyay, S., Singh, B.K.: Some Results concerning generalized bent functions, `http://eprint.iacr.org/2011/290.pdf`
15. Stănică, P., Martinsen, T., Gangopadhyay, S., Singh, B.K.: Bent and generalized bent Boolean functions. In: Designs, Codes and Cryptography (2012), doi:10.1007/s10623-012-9622-5

Design of High Performance MIPS Cryptography Processor

Kirat Pal Singh[1,*], Shivani Parmar[2], and Dilip Kumar[3]

[1] Department of Electronics and Communication Engineering,
SSET, Surya World University, Bapror, Rajpura, Punjab, India
[2] Department of Electronics and Communication Engineering,
Sachdeva Engineering College for Girls, Gharuan, Punjab, India
[3] ACS Division, Centre for Development of Advanced Computing,
Mohali, Punjab, India
kirat_addiwal@yahoo.com,
{shivaniparmar03,dilip.k78}@gmail.com

Abstract. This paper presents the design and implementation of low power 32-bit encrypted and decrypted MIPS processor for Data Encryption Standard (DES), Triple DES, Advanced Encryption Standard (AES) based on MIPS pipeline architecture. The organization of pipeline stages has been done in such a way that pipeline can be clocked at high frequency. Encryption and Decryption blocks of DES, TDES and AES cryptography algorithms on MIPS processor and dependency among themselves are explained in detail with the help of architecture. Clock gating technique is used to reduce the power consumption in MIPS crypto processor. This approach results in processor that meets power consumption and performance specification for security applications. Proposed design Implementation concludes higher system performance and reducing gate propagation delay while reducing operating power consumption. The purpose this processor is to find the maximum clock frequency and adjusted setup and hold time based on propagation delay for circuits with combinational and sequential gates. Testing results shows that the MIPS crypto processor operates successfully at a working frequency of DES, TDES & AES crypto processor at 218MHz, 209MHz, & 210MHz and a operating bandwidth of 664Mbits/s, 636Mbits/s, and 560Mbits/s.

Keywords: Cryptography, Delay, Datapath, Throughput, MIPS.

1 Introduction

Security attacks against network are increasing significantly with time. Our communication media should also be secured are confidential. Cryptanalysis is the study used to describe the methods of code-breaking or cracking the code without using the security information, usually used by hackers. For this purpose, these three suggestions arrive in everyone's mind: (i) one can transmit the message secretly, so

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 778–793, 2013.

that it can be saved from hackers, (ii) the sender ensures that the message arrives to the desired destination, and (iii) the receiver ensure that the received message is in its original form and coming from the authenticate person. In order to achieve the same one can use the two techniques, (i) one can use invisible link for writing the message through the confidential person, and (ii) use of scientific approach called "Cryptography". Cryptography is the technique used to avoid unauthorized access of data. Data can be encrypted using a cryptographic algorithm by various keys. It will be transmitted in an encrypted state, and later decrypted by the intended party. If a third party intercepts the encrypted data, it will be difficult to decipher. The security of modern cryptosystem is not based on the secrecy of the algorithm, but on the secrecy of a relatively small amount of information, called a secret key. The fundamental and classical task of cryptography is to provide confidentiality by encryption methods. It is used in applications present in technologically advanced societies; it includes the security of ATM cards, computer passwords, and electronic commerce.

An encryption algorithm provides Confidentiality, Authentication, Integrity and Non-repudiation. Confidentiality ensures that the information is accessible to only authorized set of people. Authentication is the act of establishing that the algorithms are genuine. Integrity in the general means completeness but in encryption it is adhering to some mathematical proof. Non-repudiation in cryptology means that it can be verified that the sender and the recipient were, in fact, two parties who claimed to send or receive the message, respectively.

The MIPS is simply known as Millions of instructions per second and is one of the best RISC (Reduced Instruction Set Computer) processor ever designed. MIPS architecture is employed in a wide range of applications. The architecture remains the same for all MIPS based processors while the implementations may differ [1]. There is a 16- bit RSA cryptography MIPS cryptosystem have been previously designed [2]. Some adjustments and minor improvements in the MIPS pipelined architecture design are made using authenticating devices [3] such as Data Encryption Standard [DES], Triple-DES and Advanced Encryption Standard [AES] to protect data transmission over insecure medium. High speed MIPS processor possesses Pipeline architecture to speed up the processing as well as increase the frequency and performance. A MIPS based RISC processor was described in [4]. It consists of basic five stages of pipelining that are Instruction Fetch, Instruction Decode, Instruction Execution, Memory Access and Write Back. These five pipeline stages generate 5 clock cycles processing delay and several Hazards during the operation [2]. These pipelining Hazard are eliminates by inserting NOP (No Operation Performed) instruction which generate some delays for the proper execution of instruction [4]. The pipelining Hazards are of three types: data, structural and control hazard. These hazards are handled in the MIPS processor by the implementation of Forwarding Unit, Pre-fetching or Hazard detection unit, Branch and Jump Prediction Unit [2]. The Forwarding unit is used for preventing data hazards which detects the dependencies and forward the required data from the running instruction to the dependent instructions [5]. Stall occurs in the pipelined architecture when the consecutive instruction uses the same operand as that of the instruction and requires more clock cycles for execution. This reduces the performance. To overcome this situation, Instruction Pre-fetching Unit is used which reduces the Stalls and improves performance. The control hazard occurs when a branch prediction is mistaken or in general, when the system has no mechanism for handling the control hazards [5]. The control hazard is handled by two mechanisms: Flush mechanism and Delayed jump mechanism. The branch and jump prediction unit uses these two mechanisms for preventing control hazards. The flush mechanism runs instruction after a branch and flushes the pipe after the misprediction [5]. Frequent flushing may increase the clock cycles and reduce performance. In the delayed jump mechanism, Specific numbers of NOP's are pipelined after the Jump instruction to handle the control hazard. The branch and jump prediction unit placement in the pipelining architecture may affect the critical or the longest path. The standard method of increasing performance of the processor is to detect the longest path and design hardware that results in minimum clock period.

2 MIPS Crypto Processor Architecture

The single chip MIPS crypto processor (shown in Fig. 1) consists of various components like Datapath, Data I/O unit, Control Unit, Memory unit, Crypto Specific Unit, Dependency Resolver and Arithmetic Logic Unit. The dedicated data processing block consist of Datapath and Crypto IP core (coprocessor) that performs the128-bit AES cipher operation and a 64-bit DES/TDES cipher or decipher operation. Advanced Encryption Standard (AES) algorithm operates on 128bits block size by using cipher keys with lengths 128, 192 and 256 bits for encryption process respectively. The incoming data and key are stored in a matrix called state matrix and all the operations are performed over the state matrix [6]. Data Encryption Standard (DES) and Triple DES is a Symmetric crypto algorithm, which operates on 64-bit block size with 16 rounds. The input plaintext, cipher keys and output cipher text are of 64-bit. The main operation in DES and TDES is bit permutation and substitution in

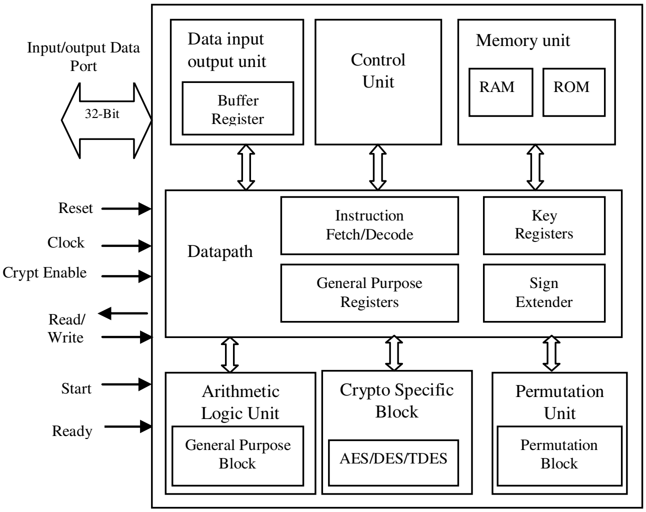


Fig. 1. MIPS crypto processor architecture

one round which is performed by the permutation unit. Datapath processing unit performs the 5 stages pipelining process inside the processor. It consists of Program Counter, 32-bit General Purpose Registers, Key Register and Sign Extender Unit. The program counter unit updates the values available at its input bus at every positive edge clock cycle and also fetches the next instruction from the instruction ROM memory. The registers are read from the General purpose register and the opcode is passed to the control unit which asserts the required control signals. Sign extension is used for calculating the effective address. The data and instruction memory have capability of storing 256 bytes and each byte is referred by the address in between 0 to 256. The address is represented by 8-bits.

The MIPS controller is the main core of the architecture which consists of control unit and ALU control signal unit. The function of controller is to controls the dedicated crypto block and performs the interface and specific operation with the external devices such as Memory, I/O bus interface controller. Single control unit controls the activities of other modules according to the instruction stored inside memory. The crypto specific block executes various other private and public key algorithms such as RSA, DSA, elliptic curve and IDEA with other application programs such as user authentication programs.

The arithmetic logic unit (ALU) performs the NOP (no operation), addition, subtraction, OR, NOR, set less than, shift left logic operation. The data and address calculations for load and store instruction are performed by ALU. The Load and Store instructions write to and read from the RAM memory in the memory unit while the ALU results and the data read from RAM are written in to the register file by the register type and Load instruction respectively. Data I/O has two different external interfaces which stored data initially at buffer registers or move data to output. The bit permutation operation has a big process part in DES and TDES algorithms as it improves diffusion properties. The incoming data is subjected to some bit position according to the permutation type. The dependency resolver block has a function to avoid stall by rearranging the instruction sequence and checking the successive instruction for their stall possibility by comparing their operands. This module handles both stalling as well as data forwarding of previous stage. In case of data dependency between two consecutive instructions the receiving instruction waits for one clock cycle. Thus dependency resolver controls the data forwarding in pipeline stages.

2.1 Data Encryption Standard

Data Encryption Standard (DES) algorithm uses the complicated logical function such as non-linear permutation and substitution. In this algorithm, there are 16 rounds of identical operation and in each round, 48-bit sub keys are generated, and substitution using S-box, bitwise shift, and XOR (exclusive –OR) operation are performed. The algorithm is designed to encrypt and decrypt blocks of data consisting of 64-bit using 56-bit key. Sometimes the key is considered as 64-bits in length for computational purpose (but only 56bits are used for conversion purpose and rest bits are used for parity checking). DES acts on 64-bit block of the plaintext, involving 16 rounds of permutations, swap, and substitutes as shown in Fig. 2. The standard includes labels describing all of the selection, permutation and expansion operations mentioned below; these aspects of the algorithm are not secrets. The basic DES steps are:

(1) The 64-bit block to be encrypted undergoes an initial permutation (IP), where each bit is moved to a new bit position; e.g., the 1st, 2nd and 3rd bits are moved to the 58th, 50th and 42nd position, respectively.
(2) The 64-bit permuted input is divided into two 32-bit blocks, called left and right, respectively. The initial values of the left and right blocks are denoted L0 and R0.
(3) These are then 16 rounds of operation on L and R blocking. During each iteration (where n ranges from 1 to 16).

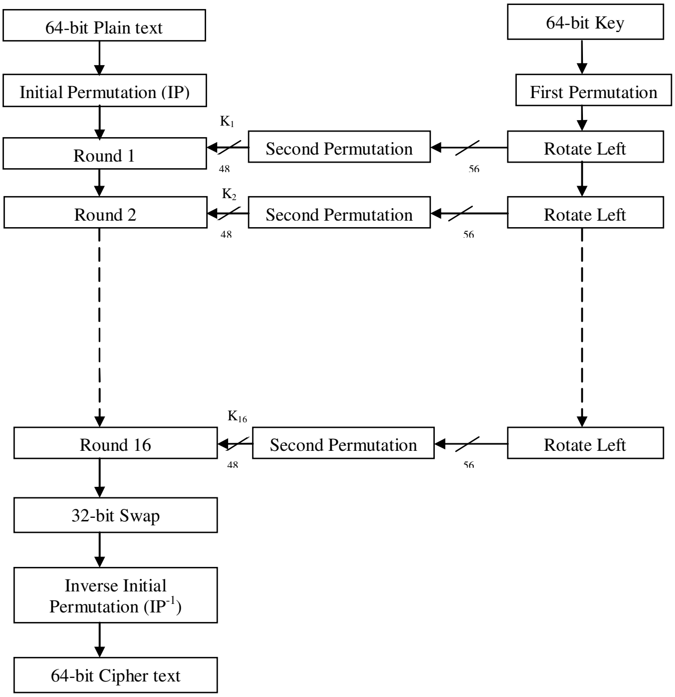


Fig. 2. DES Algorithm

At any given step in the process, the new L block value is merely taken from the prior R block value. The new R block is calculated by taking the bit-by-bit exclusive-OR (XOR) of the prior L block with the results of applying the DES cipher function f, to the prior L block and K_n. (K_n is a 48 bit value derived from the 64 bits DES key. Each round uses a different 48 bits according to the standards key schedule algorithm).

The cipher function, f, combines the 32-bit R block value and the 48-bit sub key in the following way. First, the 32-bits in the R-block and expanded to 48 bits by an expansion function (E); the extra 16 bits are found by repeating the bits in 16 predefined positions. The 48-bit expanded R block is then XORed with the 48-bit value that is then divided into eight 6-bit blocks.

There are fed as input into 8 sections (S) boxes, denoted S1,…, S8. Each 6bit input yields a 4-bit output using a lookup table (LUT) based on the 64 possible inputs; this results in a 32-bit output from the S-box. The 32-bits are then arranged by a permutation function (P), producing the results from the cipher function.

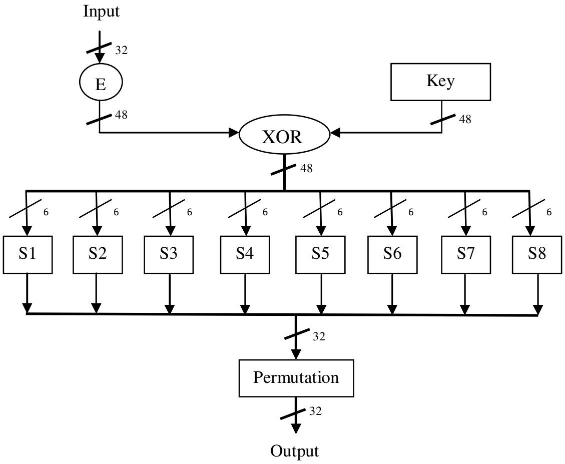


Fig. 3. DES Round (F[R,K] box) Detail

(4) The results from the final DES round- i.e., L16 and R16 are recombined into a 64-bit value and fed into an inverse initial permutation (IP^{-1}). At this step, the bits are rearranged into their original positions, so that 58th, 50th, and 42nd bits, for example, one moved back into the 1st, 2nd and 3rd positions, respectively. The output from IP^{-1} is the 64 bit cipher text block.

2.2 Triple Data Encryption Standard (TDES)

A DES algorithm is no longer considered to be a secure algorithm for many applications by the NIST (National Institute of Standard and Technology). A more secure algorithm based on DES is called Triple Data Encryption Algorithm (triple DES, 3DES, or TDEA) which is still supported by NIST. Fig. 4 shows the Triple Data Encryption Algorithm. This involves applying DES, then DES^{-1}, followed by a final DES to the plain text using three different options [7]. This results in a cipher text that is much harder to break. TDEA uses the same set of operations as DES.

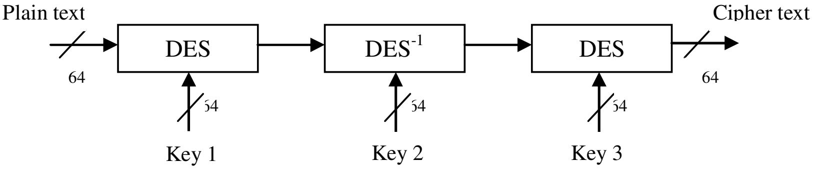


Fig. 4. TDES Block Representation

2.3 Advanced Encryption Standard (AES)

There are numerous encryption algorithms that are now commonly used in computation, but U.S government has adopted the Advanced Encryption Standard (AES) to be used by Federal departments, and agencies for protecting sensitive information. The AES algorithm is a symmetric cipher and used a single secret key for both the encryption and decryption. In addition, the AES algorithm is a block cipher as it operates on fixed-length groups of bits (blocks), whereas in stream ciphers, the plaintext bits are encrypted one at a time, and the set of transformation applied to successive bits may very during the encryption process. The AES algorithm operates on block length [N_b] of 128-bits, by using cipher keys with key length [N_k] of 128, 192 or 256 bits or the encryption process.

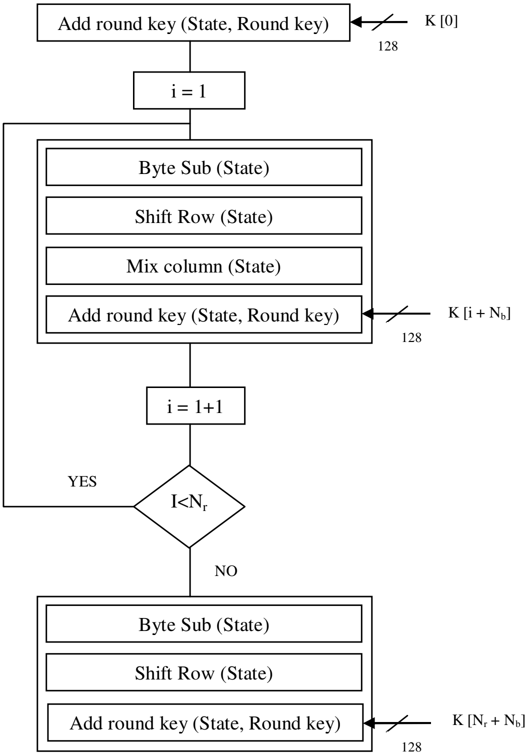


Fig. 5. AES Block Diagram for Key Length of 128bits and the number of Iterations required are 10(Nr = 10)

The encryption and decryption process of AES block consists of number of different transformation applied consecutively over the data block bits, considered as a 4x4 array of 8 bit bytes (also called "state" in the algorithm). The state undergoes four different transformations in each round having fixed number of iterations. These transformations are *"Sub Byte", "Shift Row", "Mix Column",* and *"Add Round Key"* transformations. *"Sub Byte"* can be implemented by non–linear substitution of bytes that operates independently on each byte of the state using a substitution LUT

(S-box). In this S-box; each byte in the state matrix is an element of a Galois Field GF (2^8), with irreducible polynomial m(x) = $x^8 + x^4 + x^3 + x+1$. In simple terms, GF (2^n) is a set of 2^n elements each represented by an n-bit string of 0's and 1's and affine transformation is applied (over GF (2)). The *"Shift Row"* can be implemented using a cyclically shift the rows of the state over different offsets. *"Mix Column"* are considered as most complicated operation in the algorithm and need GF (2^8) fields and multiply by modulo x^4+1 with a fixed polynomial a(x)={03}x^3 + {01}x^2 + {02}x. *"Add Round Key"* is added to the state by a logical XOR operation. Each round key consists of N_b words from the key expansion. These N_b words are added into the state columns. Each round key is a 4-word (128bit) array generated as a product of previous round key, and a sense of substitution LUT for each 32-bit word of the key. The key expansion generated a total of $N_b(N_r+1)$.

3 Design and Implementation Methodology

Current applications demand high speed processor for large amount of data transmission in real time. As compared to software alternatives, hardware implementation provides highly secure algorithms and fast solutions approaches for high performance applications. Software approaches could be a good choice but it has some limitations like low performance and speed. Main advantages of software are low cost and short time to market. But they are unacceptable in terms of high speed and performance specification. So that, Hardware alternatives could be selected for implementing MIPS crypto processor architecture.

Table 1. Hardware v/s software alternatives for crypto processor

Parameters	Software	Hardware	
		FPGA	ASIC
Performance	Low	Medium-High	Very High
Power consumption	Depends	Very high	Low
Logic integration	Low	Low	High
Tool cost	Low	Low	Low
Test development complexity	Very low	Very low	High
Density	High	Very low	High
Design efforts	Low-medium	Low-medium	High
Time consumed	Short	Short	High
Size	Small-medium	Small	Large
Memory	Fine	Fine	Fine
Flexibility	High	High	-
Time to market	Short	Short	High
Run time configuration	-	high	-

Hardware implementation supports both Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASIC) at high data rates. Such design has high performance but more time consuming and expensive as compared to software alternatives. The detailed comparison of hardware v/s software solutions for implementing the MIPS crypto processor architecture is shown in Table 1. Based on the comparison, hardware solution is a better choice in most of the cases because they have high performance. The main advantage of FPGA in hardware alternative, FPGA are low density and low area consumption. Logic integration, size and density are the major drawbacks in ASIC but have higher performance than FPGA.

3.1 Hardware Implementation of Cryptographic Engine

The global architecture of encrypted and decrypted MIPS pipeline processor is modified in a way that it executes encrypted instruction. Fig. 6 shows the block diagram of encrypted MIPS processor. To modify MIPS processor for encryption, we insert the cryptography module such as Data Encryption Standard (DES), Triple Data Encryption Standard (T-DES), Advanced Encryption Standard (AES) etc. to the pipeline stage. Only single cryptographic module is used in same hardware implementation. The instruction fetch unit of encrypted MIPS contains Program Counter (PC), Instruction Memory, Decryption core and MUX. The Instruction memory reads address from the PC and stores instruction value at the particular address that is pointed by the PC. Instruction Memory sends encrypted instruction to MUX and decryption core. The decryption core gives decrypted instructions which are further sent to the MUX. The output of MUX is fed to the IF register. The MUX control signal comes from control unit. The instruction decode unit contains Register file and Key register. Key register stores the key data of encryption/decryption core. Key address and Key data comes from write back stage. The key data to be stored into the register file and remains same for all program instruction execution. The control unit provides various control signals to other stages. This acts as select line for two multiplexers. When the control unit detects a store/branch/jump it asserts the control signal high and keep it asserted till a load instruction is detected. During that period,

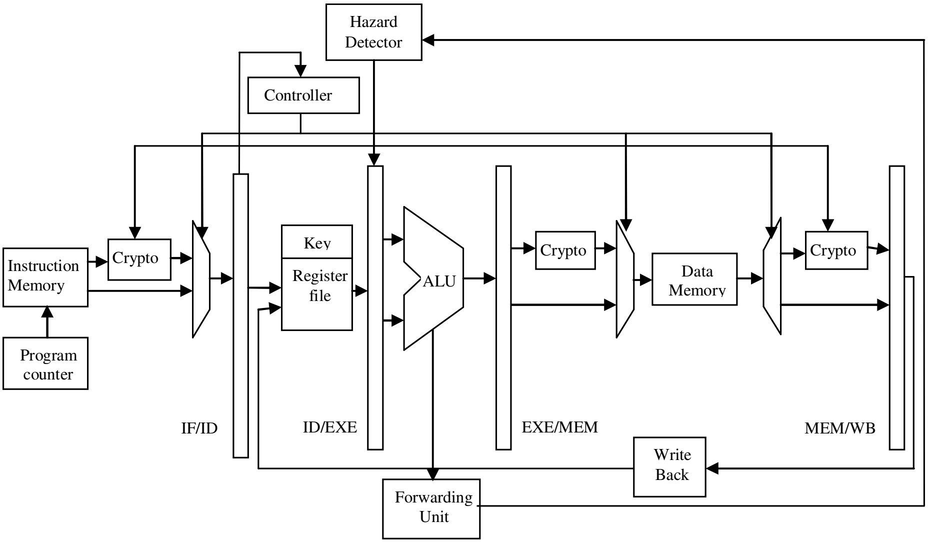


Fig. 6. Detailed MIPS crypto processor architecture

the write back stage gets the forwarded data and the memory stage gets a constant zero value thus preventing only further transitions. When the control signal is de-asserted, then the data pass through the standard pipeline structure. The execute unit executes the register file output data and performs the particular operation determined by the ALU. The ALU output data is sent to EXE register which temporarily store address value. The Memory Access Unit contains Encryption core, Decryption core, Data Memory, MUX and DEMUX. The second register data from register file is fed to the encryption core and the MUX. Here the crypt signal enable/disable encryption operation when occurs. The read/write signal of data memory describes whether reading/writing operation is done. Output of data memory pass through DEMUX whose one output goes to decryption core and other to MEM register. Here the unencrypted memory data and decrypted data are temporarily stored to the MEM register. The MEM output is fed to the write back data MUX and according to the control signal, the output of MUX goes to register file.

3.2 Initialization

The operational mode of the MIPS crypto processor is controlled by a RESET signal. When the RESET signal is at logic “0”, the crypto processor is in the reset mode and the processing unit writes the memory and register contents using the 32-bit bidirectional data bus, 10-bit address bus, and four control signals. When the reset signal is at logic “1”, the crypto processor is in the running mode and acts as an FPGA, implementing one of the three algorithms based on the preloaded contents of the memory blocks. The keys are kept in the key registers of the register file of crypto processor that are available to other stages of processor.

3.3 Microinstruction Set

The MIPS instruction set is straightforward like any other RISC designs. MIPS are a load/store architecture, which means that only load and store instructions access memory. Other instructions can only operate on values in the registers [8]. Generally, the MIPS instructions can be broken into three classes: the memory-reference instructions, the arithmetic- logical instructions, and the branch instructions. Also, there are three different instructions formats (as shown in Fig.7) in MIPS architecture: R-Type instructions, I-Type instructions, and J-Type instructions. A subset of the instruction has been implemented in our design, the list of which is given in Table 2.

Instruction Type	Instruction
R-Type	AND, OR, NOR, ADD, SUB, SLT
I-Type	ADDI, SUBI, NORI, ANDI, SLTI,SLL, SRL
	LW,SW, LKLW, LKUW
	BEQ, BNE
J-Type	J, JR, JAL, CRYPT

Fig. 7. Implemented MIPS Instruction Types

Table 2. MIPS Instruction Format

R-Type	Op	RS	RT	RD	Shamt	Funct

Arithmetic instruction format

I-Type	Op	RS	RT	Address/immediate

Transfer, branch, immediate

J-Type	Op	Target address

Jump instruction

Field	**Description**
Op[31-26]	6-bit operation code
RS[25-21]	5-bit source register
RT[20-16]	5-bit target register
Immediate[15-	16-bit immediate address
Target[25-0]	26-bit jump target
RD[15-11]	5-bit destination register
Shamt[10-6]	5-bit shift amount
Funct[5-0]	6-bit function field

The MIPS instruction Format is used to minimize the number of bits in each instruction, note that the 6-bit operation code field in the instruction format is used to have the word length of the memory as 32-bit and used standard memory blocks for the program memory. Only 32-bit instruction set is required for the current implementation as shown in Table 3. There are three more new instructions that support encrypted and decrypted operation. These instructions are load key upper word (LKUW), load key lower word (LKLW) and encryption mode (CRYPT). These instructions randomly use opcodes in the hardware implementation. LKLW and LKUW come under I-type instruction and variant of load word (LW). These two instructions do not need to specify a destination address in the assembly code. CRYPT instruction comes under J-type instruction and instead of address, only single argument i.e., Boolean value is assigned. This indicates enable/disable encryption and decryption process. Any non zero value enables the encryption/decryption process and zero value disables the encryption process.

4 Implementation Results

The complete pipeline processor stages are modeled in VHDL. The syntax of the RTL design is checked using Xilinx tool. For functional verification of the design the MIPS processor is modeled in Hardware Descriptive Language. The design is verified both at the block level and top level. The complete design along with all timing constraints, area utilization and optimization options are described using Synthesis Report. The design has been synthesized targeting 40nm triple oxide process technology using Xilinx FPGA Virtex-6 (xc6vlx240t-3ff1156) device. The Virtex family is the latest and fastest FPGA which aims to provide up to 15% lower dynamic and static power and 15% improved performance than the previous generation [18]. It is obvious that there is a trade-off between maximum clock frequency and area utilization (number of slices LUT's) because the basic programmable part of FPGA is the slice that contains four LUTs (look up table) and eight Flip flops. Some of the slice can use their LUT's as distributed RAM.

Table 3. ISA overview of Implemented MIPS Crypto Processor

Instruction	Operation	Syntax	Opcode	Clock cycle
ADD	Arithmetic Addition operation	*Add $rd, $rs, $rt*	0(0x000000)	4
SUB	Arithmetic Subtraction operation	*Sub $rd, $rs, $rt*	0(0x000000)	4
AND	Logical AND operation	*AND $rd, $rs, $rt*	0(0x000000)	4
OR	Logical OR operation	*OR $rd, $rs, $rt*	0(0x000000)	4
NOR	Logical NOR operation	*NOR $rd, $rs, $rt*	0(0x000000)	4
SLT	Set less than manipulation operation	*SLT $rd, $rs, $rt*	0(0x000000)	4
ADDI	Immediate arithmetic addition operation	*ADDI $rd, $rs, constant*	8(0x001000)	4
SUBI	Immediate Arithmetic Subtraction operation	*SUBI $rd, $rs, constant*	8(0x001000)	4
SLTI	Immediate Set less than manipulation operation	*SLTI $rd, $rs, constant*	8(0x001000)	4
ORI	Immediate Logical OR operation	*ORI $rd, $rs, constant*	8(0x001000)	4
ANDI	Immediate Logical AND operation	*ANDI $rd, $rs, constant*	8(0x001000)	4
NORI	Immediate Logical NOR operation	*NORI $rd, $rs, constant*	8(0x001000)	4
SLL	Shift left logic operation	*SLL $rd, $rs, shamt*	0(0x000000)	4
SRL	Shift right logic operation	*SRL $rd, $rs, shamt*	0(0x000000)	4
BEQ	Branch equal operation	*Beq $rd, $rs, label*	4(0x000100)	3
BNE	Branch not equal operation	*Bne $rd, $rs, label*	4(0x000100)	3
JR	Conditionally jump to register	*Jr $rd*	2(0x000010)	3
JAL	Unconditionally jump to program	*Jal $rd*	2(0x000010)	3
J	Conditionally jump to program	*J $rd*	2(0x000010)	3
CRYPT	Encryption/decryption enable	*Crypt $rd*	65(0x111111)	3
LW	Load data word to CPU	*Lw $rd, offset($rs)*	35(0x100011)	5
SW	Store data to memory	*sw $rd, offset($rs)*	43(0x101011)	4
LKUW	Load key upper word to target register	*LKUW $rd, offset($rs)*	64(0x111110)	5
LKLW	Load key load word to target register	*LKLW $rd, offset($rs)*	60(0x111100)	5

4.1 Design Performance, Area and Power Requirement

The performance of MIPS crypto processor based on three different crypto modules such as DES, TDES, and AES algorithms. For DES and TDES, 16 clock cycles are used for DES/TDES crypto specific block to execute data, 20 clock cycles are needed to execute the R-type instruction, 21 clock cycle are needed for I-type instruction and 19 clock cycle for J-type instruction data.

The power consumption is estimated by the Xilinx XPOWER Analyser tool, using the post layout netlist of the crypto processor along with the node activity data for each algorithm. The power consumption can be further reduced by running the processor at lower voltages than the normal voltage of 1.5v (as long as the speed and throughput requirements are satisfied). Power analysis was done for the portion between the EXE/MEM and MEM/WB stage. This is performed for both the encryption and decryption process. Clock gating technique is used to minimize energy reduction during pipeline stall stages. This technique identifies low processing requirement periods and reduces operating voltage with clock frequency (voltage-frequency scaling), resulting in reduced average operating power consumption. This may or may not occur frequently depending upon compiler efficiency. The power analysis result is carried out on the same clock frequency. in our design, a symbol is processed every clock cycle, the throughput is calculated on the basis of number of instruction execution per second. The formula for calculating throughput is:

Throughput = f * symbol width/total clock frequency

Where f is the operation frequency and symbol width is one of our parameterized values. Table 4 and Table 5 show the performance throughput, area and the estimated power consumption of DES and TDES MIPS crypto processor. Maximum throughput of MIPS DES based crypto processor is of 664Mbits/s at 4.58ns and for TDES based crypto processor is 636Mbits/s at 4.78ns.

Table 4. Throughput estimates for the MIPS crypto processor based on DES

Features	Processor
Crypto processor	DES
Data length	64-bits
Speed	218MHz (clock rate)
Throughput	664 Mbits/s (Data Bandwidth)
Area	66072 Slice LUT's(look up tables)
Latency	21 clock cycles(both for encryption and decryption)
Power consumption	1.746W(quiescent-1.303 and dynamic-0.444)

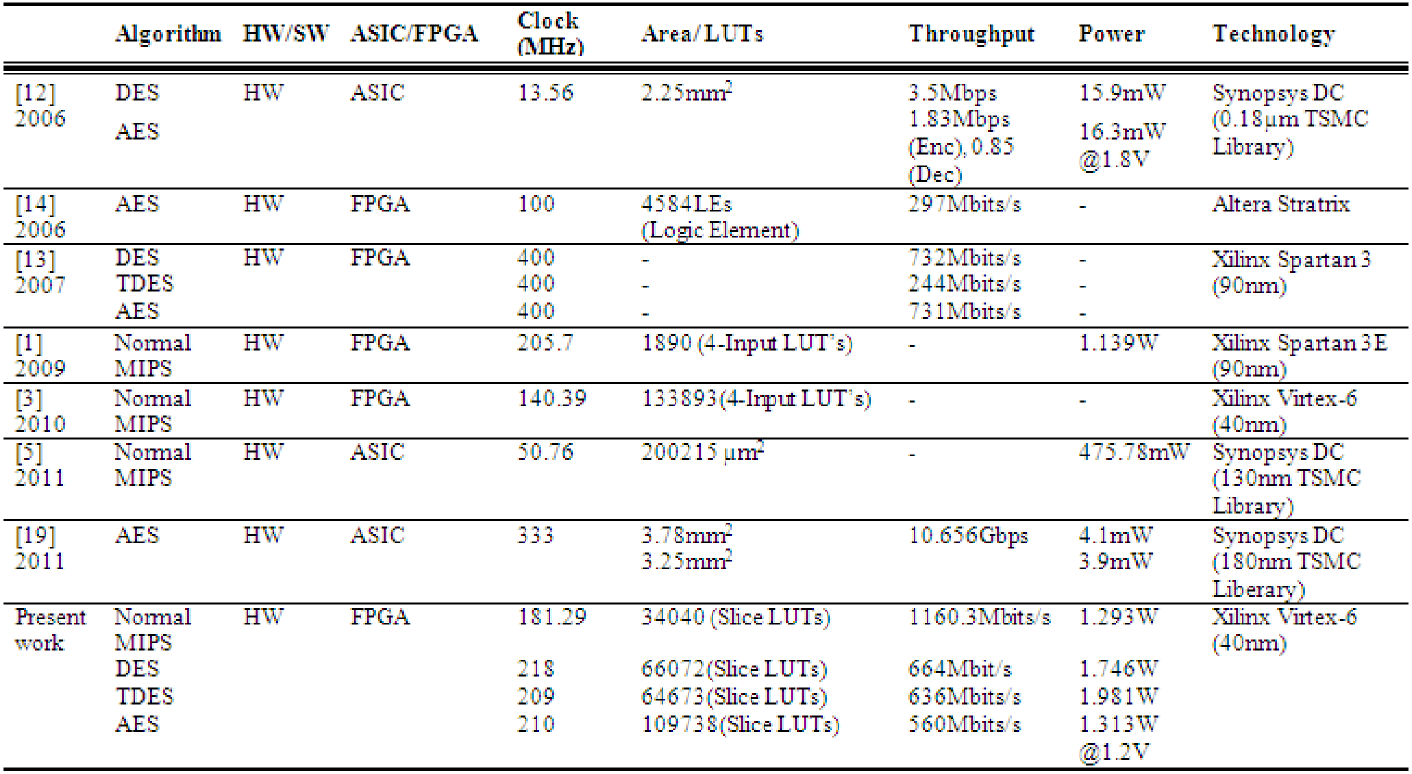
Table 5. Throughput estimates for the MIPS crypto processor based on TDES

Features	Processor
Crypto processor	TDES
Data length	64-bits
Speed	209MHz(clock rate)
Throughput	636 Mbits/s (Data Bandwidth)
Area	64673 Slice LUT's(look up tables)
Latency	21 clock cycles(both for encryption and decryption)
Power consumption	1.981W(quiescent-1.131 and dynamic-0.851)

Table 6. Throughput estimates for the MIPS crypto processor based on AES

Features	Processor
Crypto processor	AES
Data length	128-bits
Speed	210MHz (clock rate)
Throughput	560Mbits/s (Data Bandwidth)
Area	109738 Slice LUT's(look up tables)
Latency	48 clock cycles(for encryption)
Power consumption	1.313W(quiescent-1.008 and dynamic-0.396)

Table 7. Some Recent Cryptography Algorithm Implementation Specification



	Algorithm	HW/SW	ASIC/FPGA	Clock (MHz)	Area/ LUTs	Throughput	Power	Technology
[12] 2006	DES AES	HW	ASIC	13.56	2.25mm^2	3.5Mbps 1.83Mbps (Enc), 0.85 (Dec)	15.9mW 16.3mW @1.8V	Synopsys DC (0.18μm TSMC Library)
[14] 2006	AES	HW	FPGA	100	4584LEs (Logic Element)	297Mbits/s	-	Altera Stratrix
[13] 2007	DES TDES AES	HW	FPGA	400 400 400	- - -	732Mbits/s 244Mbits/s 731Mbits/s	- - -	Xilinx Spartan 3 (90nm)
[1] 2009	Normal MIPS	HW	FPGA	205.7	1890 (4-Input LUT's)	-	1.139W	Xilinx Spartan 3E (90nm)
[3] 2010	Normal MIPS	HW	FPGA	140.39	133893(4-Input LUT's)	-	-	Xilinx Virtex-6 (40nm)
[5] 2011	Normal MIPS	HW	ASIC	50.76	200215 μm^2	-	475.78mW	Synopsys DC (130nm TSMC Library)
[19] 2011	AES	HW	ASIC	333	3.78mm^2 3.25mm^2	10.656Gbps	4.1mW 3.9mW	Synopsys DC (180nm TSMC Liberary)

Present work	Normal MIPS
	DES
	TDES
	AES
		HW	FPGA	181.29
				218
				209
				210
					34040 (Slice LUTs)
					66072(Slice LUTs)
					64673(Slice LUTs)
					109738(Slice LUTs)
						1160.3Mbits/s
						664Mbit/s
						636Mbits/s
						560Mbits/s
							1.293W
							1.746W
							1.981W
							1.313W @1.2V
								Xilinx Virtex-6 (40nm)

In case of AES crypto processor, 43 clock cycles are used for crypto specific block to execute data, 47 clock cycles are needed to execute the R-type instruction, 48 clock cycles are needed for I-type instruction and 46 clock cycles for J-type instruction data. Table 6 shows the performance throughput; area and the estimated power consumption of AES based MIPS crypto processor. Maximum throughput of AES based MIPS Crypto processor is 560Mbits/s at 4.76ns. Moreover, it is possible to trade performance with area and power in the implementation. For example, higher performance can be obtained by running processor at higher frequency up to 300MHz for the current design (increasing power consumption) and/or using pipeline (increasing area) for more performance demanding applications. Some recent cryptography algorithms specification are shown in table 7.

5 Conclusion

In this paper, we have presented the design of high performance 32-bit cryptography MIPS processor that executes encrypted/decrypted instructions. Initially it read encrypted data from instruction memory and decrypts the same data and sent it to the next pipeline stages. The processor uses the symmetric block viz., DES, TDES and AES plain/cipher that can process data length of 64 bits, 64bits and 128bits respectively. The crypto algorithm block inside the MIPS processor performs data encryption and decryption algorithm. The design has been modeled in VHDL and synthesize using Xilinx tool and functional verification and optimization policies are adopted for it. Optimization and synthesis of design is carried out at latest and fastest FPGA Viretx-6 device that improves performance. Each program instructions are tested with some of vectors provided by MIPS. We conclude that the performance of MIPS crypto processor using DES and AES is High 664Mbits/s and 560Mbits/s respectively.

References

[1] Gautham, P., Parthasarathy, R., Balasubramanian, K.: Low-power pipelined MIPS processor design. In: International Symposium on Integrated Circuit (ISIC 2009), pp. 462–465 (2009)

[2] Zulkifli, Yudhanto, Soetharyo, Adinono: Reduced Stall MIPS architecture using Pre-fetching accelerator. In: IEEE International Conference on Electrical Engineering and Informatics, pp. 611–616. IEEE (August 2009) ISBN: 978-1-4244-4913-2

[3] Ghewari, P.B., Patil, J.K., Chougule, A.B.: Efficient hardware design and implementation of AES cryptosystem. International Journal of Engineering Science and Technology 2(3), 213–219 (2010)

[4] Patterson, D.A., Hennessy, J.L.: Computer Organization and Design, The hardware/Software Interface. Morgan Kaufmann (2005)

[5] Lotfi, P., Salehpour, A.-A., Rahmani, A.-M., Afzali-kusha, A., Navabi, Z.: Dynamic power reduction of stalls in pipelined architecture processors. International Journal of Design, Analysis and Tools for Circuits and Systems 1(1), 9–15 (2011)

[6] Sever, R., Neslin Ismailoglu, A., Tekmen, Y.C., Askar, M.: A high speed ASIC Implementation of the rijndael Algorithm. In: IEEE International Symposium on Circuits and Systems (2004)

[7] Advanced Encryption Standard (AES). Fed. Inf. Process. Syandards Pub. (November 2001)

[8] Balpande, R.S., Keote, R.S.: Design of FPGA based Instruction fetch & decode Module of 32-bit RISC (MIPS) processor. In: International Conference on communication Systems and Network Technologies, pp. 409–413. IEEE (2011) ISBN: 978-0-7695-4437-3

[9] Taherkhani, S., Ever, E., Gemikonakli, O.: Implementation of Non-pipelined and pipelined data encryption standard (DES) using Xilinx Virtex-6 technology. In: 10th IEEE International Conference on Computer and Information Technology (CIT 2010), pp. 1257–1262 (2010)

[10] Navabi, Z.: VHDL: Modular design and synthesis of cores and systems, pp. 283–291. McGrew-Hills (2007) ISBN: 978-0-07-147545-7

[11] Floyd, L.: Digital Fundamental with VHDL, pp. 362–368. Pearson Education (2003) ISBN: 0-13-099527-4

[12] Eslami, Y., Sheikholeslami, A., Glenn Gulak, P., Masui, S., Mukaida, K.: An Area-Efficient Universal Cryptography Processor for Smart Cards. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 14(1), 43–56 (2006), doi:10.1109/TVLSI. 2005.863188, ISBN: 1063-8210

[13] Askar, M., Egemen, T.: Design and SystemC Implementation of a Crypto Processor for AES and DES Algorithms. In: Information Security & Cryptology Conference with International Participation (ISC Turkey). Bildiriler kitabi Proceedings, pp. 145–149 (December 2007)

[14] Hani, M.K., Wen, H.Y., Paniandi, A.: Design and Implementation of a Private and Public key Crypto Processor for Next-Generation IT Security Applications. Malaysian Journal of Computer Science 19(1), 29–45 (2006)

[15] Sklavos, N.: On the Hardware Implementation Cost of Crypto-Processors Architectures. Information Security Journal: A Global Perspective, 53–60 (June 2010), doi:10.1080/19393551003649016, ISSN: 1939-3555 print/ 1939-3547

[16] Xilinx, ISE Simulator, `http://www.xilinx.com/tools/isim.htm`

[17] Xilinx, XST Synthesis, `http://www.xilinx.com/tools/xst.htm`

[18] Xilinx, ISE In-Depth tutorial, pp. 95–120 (June 2009), `http://www.xilinx.com/support/documentation/sw_manuals/xilinx11/ise11tut.pdf`

[19] Saravanan, P., RenukaDevi, N., Swathi, G.: A High-Throughput ASIC implementation of Configurable Advanced Encryption Standard (AES) Processor. International Journal of Computer Applications Special Issue on "Network Security and Cryptography" 3, 1–6 (2011)

Design of an Edge Detection Based Image Steganography with High Embedding Capacity

Arup Kumar Pal[1] and Tarok Pramanik[2]

[1] Department of C.S.E, ISM Dhanbad,Jharkhand, India
[2] Department of C.S.E, R.V.S College of Engineering and Technology, Jamshedpur, India
{arupkrpal,tarok.kgec}@gmail.com

Abstract. In this paper, the authors have proposed an image steganography method for improving the embedding capacity of the gray-scale cover image. In general, the embedding of the secret message into the sharp areas i.e. edge region rather than in the smooth areas i.e. non edge region of the cover image attains relatively better quality stego-image. So in the proposed work, we have also exploited the presence of edges in the cover image to embed a large amount of secret message into the cover image. The proposed method carried out into two phases: in the first phase the cover image is classified into edge region and non-edge region. Subsequently in the second phase, the binary secret message bits are embedded by replacing some least significant bits (LSBs) of each pixel. In the proposed work, x LSBs replacement are preferred for the pixels belongs to edge region and y LSBs replacement are considered for non-edge region pixels where x>y. The proposed method increases the embedding capacity of the cover image compare to some existing standard steganographic methods. The scheme has been tested on several standard gray-scale test images and the obtained simulation results depict the feasibility of the proposed scheme.

Keywords: Edge Detection, Image Steganography, Information Security, LSB Substitute method.

1 Introduction

The widespread popularity and usage of Internet make the information interchange or communication easier and faster. In such type of communication, the data or information travel through the public channel. So sometime, it becomes essential to prevent the unauthorized accessibility of these data from the illegitimate users. To keep the data secure, various types of cryptography [1] and steganography [2] techniques have been devised. In cryptography, the encryption technique [1] transferred the secret data into corresponding cipher text using a secret key. This cipher text is unintelligible to the illegitimate user. Only the decryption technique [1], reverse process of the encryption, extracts the original data from the cipher text using the secret key. Although after encryption process, the produced cipher text is remained secure but still it flags the importance of the cipher text and attracts

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 794–800, 2013.

illegitimate users or opponent's attention to employ different possible cryptography attacks on the cipher text to extract its original content. To sidestep this kind of drawback of the cryptography, steganography [2] is another widely accepted approach for secure communication of the secret message. In this method, the actual data or secret message are embedded into a digital cover media like digital image, audio or video in such a way that the secret message will be visually imperceptible after embedding into the cover media. In steganography, after embedding secret message into the cover media, the modified cover media, which is known as a stego-media should look visually almost similar to the cover media. So in this approach, the secret message becomes imperceptible and also such approach is capable to divert opponents' attention so that the chance of attacks will be reduced. Therefore to protect the secret message from being illegally accessed, steganography is a better choice than the cryptograpy.

In general, steganography is classified as video steganography [3], audio steganography [4], image steganography [5] and text steganography [6] based on the used carrier or cover media like video, audio, image and text respectively to embed the secret data into them. In image steganography, the used cover media is a digital image which is termed as cover image or host image. The cover image with the embedded secret data forms the stego-image. During designing of an image steganography, it is found a trade-off between the hiding capacity of the cover-image and the visual quality of the stego-image. So the aim of an effective image steganography is to preserve high quality of stego-image with high embedded secret message . To achieve this goal, in literature several well accepted image stegnography[5,7-10] have been proposed. Among them simple least significant bit (LSB) substitution [2] is one of the widely used methods due to its low computational complexity and high hiding capacity. In this approach, the secret data are initially converted into binary data and the corresponding binary bits sequence are concealed into cover image pixels by replacing a number of the least significant bits (LSB) of each pixel. In digital image most of the significant information is carried out by the most significant bits (MSB) of each pixel so changing the MSBs of the cover image pixel will cause serious degradation of the visual quality of the stego-image [5]. Thus, the LSB substitution method fixes on to embed secret data into the parts of LSB of the cover image pixel. In general up to first *3* LSB bits of the cover image pixels are used to conceal the secret message bits otherwise the quality of the stego-image is found to degrade extremely [5]. In literature, some modified schemes based on LSB substitution [5, 7] were proposed but all those schemes were only capable to improve the visual quality of the stego-image where no improvement of the hiding capacity is found. So in this paper, our aim is to improve the embedding capacity without compromising the visual quality of the stego-image. It has been found in many experiments that the insertion of the secret data into the sharp or edge region of a cover image is more effective than embedding into the smooth region [9], as it causes less distortion on the cover-image. So in the proposed work, we have taken advantage of sharp/edge region of the cover image pixels to hide large amount of secret data by replacing 4 LSB bits where in non-edge or smooth region pixels are replaced up to maximum 3 LSB bits. In the proposed method, hiding capacity has been increased as well as the quality of the stego-image is well preserved. The rest of the paper is organized as follows. The details of the proposed image steganography method are presented in section 2. The experimental results are described in sections 3 to show the effectiveness of the proposed method and finally, section 4 provides some conclusions.

2 Proposed Image Steganography

The intention of the proposed image steganography method is to improve/increase the hiding capacity of the cover image. The proposed method takes advantage of the presence of edge region in the cover image to embed more secret information than embedding into non-edge region. The embedding process of the proposed work is implemented broadly in two phases. In the first phase, we have used edge detection mask like sobel [11] with a suitable threshold value on the cover image to find the edge detected image. In the subsequent stage, this edge detected image helps to locate the edge and non-edge region of a cover image. In the last phase, the embedding process has been carried out where we have used simple LSB substitution methods for embedding binary bit sequence of the secret message into the cover image using varying LSB substitute values. In the embedding process, the pixels belongs to edge region are replaced by higher LSB substitute values than the pixels belongs to non-edge region. However in edge detection based steganography method, the main challenge is to identify the same edge and non-edge region from the resulted stego-image during extraction process. So in this work, we have embedded some additional information during embedding process so that the receiver can easily found out the edge and non-edge region from the stego-image. The embedded information may consider as indicator. The algorithmic steps of the proposed embedding procedure are presented as follows.

Algorithm 2.1: Embedding Procedure
Input: A secret image and a cover image
Output: A stego-image.
Step 1: Employ any edge detection mask/operator with a suitable threshold value, *Th* on the cover image to obtain the edge detected cover image. Convert the secret image into secret binary data *(SB).*
Step 2: Decompose both the edge detected image and the cover image into non-overlapping blocks of size $n \times n$. Find the presence of the prominent edge line in each block from the cover image based on the occurrence of the edge line in corresponding block of the edge detected image.
Step 3: Use first pixel of every block as an indicator pixel. Replace *3* LSBs of indicator pixel for each block by one of the following *3* bits information *000, 001, 010, 011* and *100* for indicating the presence of smooth region, horizontal edge, vertical edge, 45^0 diagonal edge and 135^0 diagonal edge respectively.
Step 4: Process each block as raster scan order (i.e. from left to right followed by top to down) and embed secret message bits from *SB* into each block where during embedding process for each block if the pixels belong to the edge line then choose ‘x’ bits secret message from *SB* for embedding into each pixel LSBs position.

Otherwise if the pixels belong to the non- edge region then consider ‘y’ bits secret message from *SB* for insertion into the corresponding pixels LSBs position. Here x>y. Now after completion of embedding all binary bits sequence of *SB* into cover image, the stego-image is formed.

The schematic diagram of the proposed embedding procedure is shown in Fig.1. The secret message can be extracted from the stego-image in comparatively less number of steps than embedding procedure. The algorithmic steps are presented as follows.

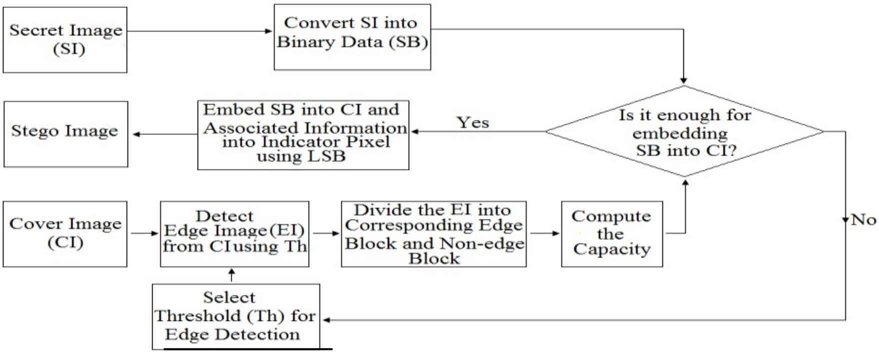

Secret Image (SI)
Convert SI into Binary Data (SB)
Stego Image
Embed SB into CI and Associated Information into Indicator Pixel using LSB
Yes
Is it enough for embedding SB into CI?
No
Cover Image (CI)
Detect Edge Image(EI) from CIusing Th
Divide the EI into Corresponding Edge Block and Non-edge Block
Compute the Capacity
Select Threshold (Th) for Edge Detection

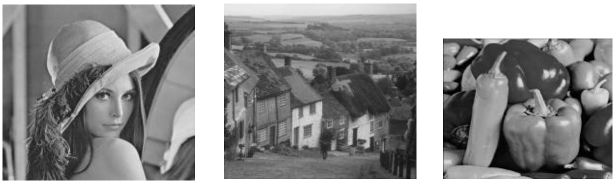
Fig. 1. The Proposed Embedding Procedure

Algorithm 2.2: *Extraction Procedure*
Input: The stego-image
Output: The secret image (SI)
Step 1: Decompose the stego-image into non-overlapping blocks of size $n \times n$.
Step 2: Process each block as raster scan order and extract 3 bits information from first pixel 3LSBs position of every block. Extracted 3 bits information is the indicator value to locate the presence of prominent edge direction in the corresponding block.
Step 3: Based on indicator value find the edge pixels and non-edge pixels from each block. Extract from each pixel (except indicator pixel) LSB position either ‘x’ bits data for edge pixel or ‘y’ bits data for non-edge pixel to form secret binary data (SB). Convert the secret binary data (SB) into corresponding image i.e. the secret image.

3 Experimental Results and Discussion

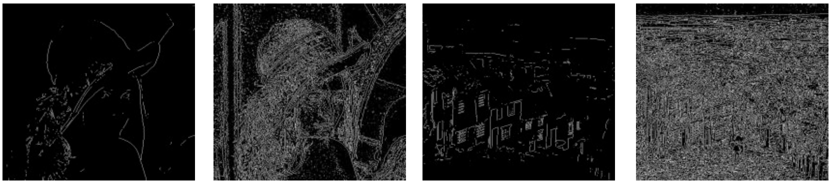
In he proposed scheme has been tested on number of standard gray-scale images but in this paper, we have presented the simulation results for two different types of cover images. For making comparisons convenience, we have chosen same secret message i.e. a standard gray-scale image for embedding into different cover images. The used cover images, each of size *510*×*510* and the secret image of size *289*×*324* are shown in Fig. 2.

In our experiment, the detection of an edge image (EI) from the cover image (CI), using an appropriate threshold value (Th), has been carried out by a MATLAB command *EI=edge (CI,'sobel',Th).* Here, we have considered *Th=0.1* and *0.01* for comparative study. Fig. 3 shows the presence of edge pixels in the corresponding cover image. In our experiment, the cover image are decomposed into non-overlapping blocks of size *5×5* and afterwards the secret message bits and the corresponding indicator value are embedded into each blocks using LSB substitute method. The pixels those belongs to edge line, contains 4-bits secret message in LSB position whereas the non-edge pixels contain 3-bit secret message in their corresponding LSB position. The stego-images thus formed are shown in Fig.4. Based on the human visual perception, it can be observed that the stego-images shown in Fig.4 are quite good and almost similar as their corresponding original cover image. We also compared our proposed work with LSB substitute method and OPAP method [5] in terms of the peak signal-to-noise ratio (PSNR) value and by computing embedding capacity respectively. Table 1 shows the obtained PSNR values from different stego-images, resulting after embedding the secret pappers image, using different method like LSB substitute method, OPAP [5] and the proposed method with threshold value 0.01 and 0.1 respectively. According to the results shown in Table 1, it is evident that the PSNR values of the stego-images are slightly inferior than the other compared methods. However in our proposed method, the obtained PSNR values are reasonably good since it has been observed that an image with a PSNR value greater than 30 dB is acceptable by human visual perception [10] . The LSB substitute and OPAP methods provide same embedding capacity and it always fixed for any cover image. The embedding capacity for a cover image of size 510×510 is 95.25 KB when 3 bits LSB replacement is considered. But in the proposed scheme, the embedding capacity varies for image to image since every image has different edge patterns. Table 2 shows the embedding capacity of the proposed method for various cover images. According to the results shown in Table 2, it is clear that the embedding capacity of the proposed method with Th=0.1 is not better than LSB substitute and OPAP methods. However embedding capacity of the proposed method with Th=0.01 is superior to the other mentioned methods. So the prposed method with Th=0.01 is capable of providing high embedding capacities while retaining image quality higher than 30dB.



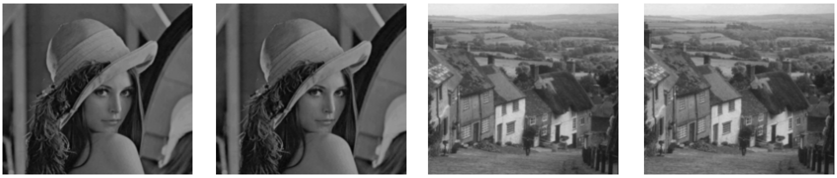
(a) (b) (c)

Fig. 2. The original Cover image (a) Lena Image and (b) Goldhill; the secret message (c) Peppers image



(a) (b) (c) (d)

Fig. 3. Edge detected image of (a) Lena using threshold value 0.1;(b) Lena using threshold value 0.01;(c) Goldhill using threshold value 0.1;(d) Goldhill using threshold value 0.01



(a) (b) (c) (d)

Fig. 4. Stego-image of (a) Lena using threshold value 0.1; (b) Lena using threshold value 0.01; (c) Goldhill using threshold value 0.1;(d) Goldhill using threshold value 0.01

Table 1. Comparative Study in Terms of PSNR (in dB) of the Stego-images Created by Various Embedding Algorithms

Cover Image	**Embedding Method**			
	LSB Substitute	**OPAP [5]**	**Proposed (Th=0.1)**	**Proposed (Th=0.01)**
Lena	38.1826	39.0843	**37.6974**	**36.6174**
Goldhill	38.1741	39.0747	**37.5753**	**36.3499**

Table 2. The Embedding Capacity of the Proposed Method

Embedding Method	**Cover Image**	**Number of Non-Edges Blocks**	**Number of Edges Blocks**	**Embedding Capacity (KB)**
Proposed Method (Th=0.1)	Lena	9526	878	**91.98**
	Goldhill	8811	1593	**92.41**
Proposed Method (Th=0.01)	Lena	2860	7544	**96.05**
	Goldhill	698	9706	**97.37**

4 Conclusion

In image steganography, a trade off always exists between hiding capacity of the cover image and the visual quality of the stego-image. In this paper, the proposed method is proficient to enlarge the hiding capacity with retaining high quality stego- image. In steganography, it has been found that up to 3 bits secret message is possible to embed in the pixels LSB position where in contrast more than 3 bits secret message can be embedded into the pixels belong to edge region. Therefore in this paper for improving the hiding capacity, we have considered 4 LSBs replacement for embedding secret message bits into the pixels belongs to edge region whereas 3 LSBs replacement are preferred for pixel belong to non-edge region. The experimental results shows that the proposed method has attain better hiding capacity than the LSB substitute and OPAP methods with maintaining acceptable PSNR value for the stego- image.

References

[1] Stalling, W.: Cryptography and Network Security: Principles and Practices, 4th edn. Pearson Education, India (2007)

[2] Lu, C.-S.: Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property. Idea Group Publishing (2005)

[3] Hanafy, A.A., Salama, G.I., Mohasseb, Y.Z.: A secure covert communication model based on video steganography. In: Military Communications Conference, MILCOM 2008, vol. 1-6, pp. 16–19. IEEE (November 2008)

[4] Cvejic, N., Seppanen, T.: Increasing the capacity of LSB-based audio steganography. In: 2002 IEEE Workshop on Multimedia Signal Processing, December 9-11, pp. 336–338 (2002)

[5] Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recognition 37(3), 469–474 (2004)

[6] Satira, E., Isikb, H.: A compression-based text steganography method. The Journal of Systems and Software 85, 2385–2394 (2012)

[7] Wang, R.Z., Lin, C.P., Lin, J.C.: Hididng data in images by optimal moderately-significant-bit replacement. IEE Electronics Letters 36(2), 2069–2070 (2000)

[8] Wang, R.Z., Lin, C.F., Lin, J.C.: Image hiding by optimal LSB substitution and genetic algorithm. Pattern Recognition 34(3), 671–683 (2001)

[9] Chen, W.J., Chang, C.C., Le, T.H.: High payload steganography mechanism using hybrid edge detector. Expert Systems with Applications 37, 3292–3301 (2010)

[10] Yu, Y.H., Chang, C.C., Lin, I.C.: A new steganographic method for color and grayscale image hiding. Computer Vision and Image Understanding 107, 183–194 (2007)

[11] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Pearson Education, India (2008)

Trust Management Method for Vehicular Ad Hoc Networks

Riaz Ahmed Shaikh and Ahmed Saeed Alzahrani

King Abdulaziz University, Jeddah, Saudi Arabia
{rashaikh,asalzahrani}@kau.edu.sa

Abstract. In vehicular ad hoc networks, evaluating trustworthiness of data is utmost necessary for the receiver to make reliable decisions that are very crucial in safety and traffic-efficiency related applications. Existing trust management schemes that have been proposed so far for the vehicular networks has suffered from various limitations. For example, some schemes build trust based on the history of interactions. However, vehicular networks are ephemeral in nature, which makes that approach infeasible. Furthermore, in most of the existing approaches, unique identities of each vehicle must be known. This violates user privacy. In order to overcome these limitations, we have proposed a novel trust management scheme for the vehicular networks. The proposed method is simple and completely decentralized, which makes it easy to implement in the vehicular networks. We have analytically proved its robustness with respect to various security threats. Furthermore, it introduces linear time complexity, which makes it suitable to use in real-time.

Keywords: Ad-Hoc networks, Privacy, Trust model, Vehicular networks.

1 Introduction

In the last decade, we have witnessed a large increase in research and development in the domain of a vehicular ad hoc networks (VANETs). In the USA, the Federal Communications Commission (FCC) has already allocated 75 MHz of a Dedicated Short Range Communications (DSRC) spectrum at 5.9 GHz band to support vehicular networking [1]. Also, in August 2008, the European Telecommunications Standards Institute (ETSI) has allocated 30 MHz of spectrum in the 5.9 GHz band for vehicular networking [2]. Allocation of a wide DSRC spectrum enables a great number of potential applications including safety applications, real-time traffic management, on-board entertainment and mobile Internet access [3]. Many applications are proposed so far, e.g., General Motors (GM)'s collision warning system [4], Inter-vehicle hazard warning system [5], and Traffic view system [6]. However, most of the focus has been placed on reliable delivery of messages among vehicles, and less focus has been placed on evaluating the reliability of the data sent by the peers [7,8]. This motivates us to work in this direction. We firmly believe that the evaluating quality and reliability of the data is utmost necessary for the receiver to make reliable decisions, which are

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 801–815, 2013.

very crucial in safety and traffic efficiency related applications. For example, a malicious peer wants to create congestion on the road to achieve some criminal goal. For this, he reports the roads on his path as slippery. In the absence of data reliability mechanism, other peers would slow down, thus creating congestion.

Cryptographic-based security solutions do not provide any guarantee or assurance of the quality or reliability of the data itself [9,10]. That can be evaluated by measuring a trust on the sender, where trust is defined as "confidence in or reliance on some quality or attribute of a person or thing, or the truth in a statement" (Oxford English Dictionary, p. 432). The trust level will help to distinguish trusted, malicious, faulty and selfish senders. Only a few trust models have been proposed for the VANETs, e.g., [11,12,13,14,8,15], which has suffered from various limitations that are discussed in the related work section.

Trust management in the VANETs is more complex than the trust management in other networks like sensor networks, MANETs etc., because of the following reasons:

- Nodes move at very high speed, in which time to react to an imminent situation is very critical [7]. Therefore, nodes in the VANETs should be able to evaluate trust in real-time.
- Nodes in the VANETs remains in contact with each other for a short period of time, which may not be enough to establish trust based on reputation or history of interactions [16]. Therefore, trust management schemes in the VANETs should be able to cope with this time scarcity problem.

In this work, we have proposed a novel trust management scheme for the vehicular networks that efficiently deals with the above-mentioned challenges. In addition to that our proposed method has the following properties:

- *Privacy assurance*: Proposed method operates in identity anonymous environment, which ensures identity and location privacy of the user.
- *Distributed trust establishement*: Proposed trust management scheme is completely decentralized, which makes it easy to implement in the VANETs.
- *Robustness*: Proposed method is resilient against attacks on the trust model itself.

Our proposed method works in three phases. In the first phase, receiver nodes will calculate their confidence value on each message that comes from unique senders about a particular event. It is calculated based on three parameters: 1) location closeness, 2) time closeness, and 3) location verification. In the second phase, method will calculate the trust value for each unique message related to the same event. In the last phase, receiver will take the decision of an acceptance of the message, which has the highest trust value.

The rest of this paper is organized as follows: Section 2 contains related work. Section 3 describes the proposed trust management method. Section 4 consists of analysis and evaluation of the proposed method from the perspective of security resiliency and time complexity. Finally, Section 5 concludes the paper and highlights some future work.

2 Related Work

F. G. Mármol and G. M. Pérez [15] have proposed a trust and reputation infrastructure-base proposal (TRIP) for the vehicular ad hoc networks. In that work, the reputation of a node is first calculated based on the three factors: 1) direct previous experiences with the target node, 2) recommendations from other surrounding vehicles, and 3) recommendation from central authority through RSU. After that, system will map the reputation score with one of the three trust levels (1. Trust, 2. Not Trust, and 3. +/- Trust), which are represented as fuzzy sets. The proposed scheme is based on one very strong assumption that is; a vehicle usually circulate over the same road, and at the same time of the day. This will help to built history. We argue that this assumption is not realistic. Furthermore, in order to build a history and reputation, actual identities of vehicles must be known. However, in order to ensure privacy in vehicular-to-vehicular (V2V) communication, the use of temporal pseudo-identities is recommended [7,17].

D. Huang *et al.* [12] have proposed a Situation-Aware Trust (SAT) architecture for the vehicular networks. The SAT includes three main components:

- An attribute-based policy control module, which is used to address a number of trust situations and application scenarios on road,
- Proactive trust module, which is used to build inter-vehicle trust in a timely fashion, and
- An email-based social network trust module, which is used to enhance trust and to allow the set up of a decentralized trust framework.

The SAT requires deployment of both global and local trust agents that makes it hybrid architecture. Authors have suggested various parameters and high level mechanisms that can be used to compute trust. However, they did not provide mathematical model that could show how to combine the various parameters together. Furthermore, authors have suggested the use of email addresses and social networks to compute trust that violates the identity and location privacy of a user.

M. Raya *et al.* [11] have proposed a data-centric trust establishment method for the ephemeral ad hoc networks. In their model, they evaluate trustworthiness of the data reports instead of the trustworthiness of the sender entities themselves. They define various trust matrices, such as, a priori trust relationship (default trustworthiness), event or task-specific trustworthiness, and time and location closeness. They evaluate data reports with corresponding trust metrics using several decision logics, such as weighted voting, Bayesian inference, and Dempster-Shafer Theory. This scheme is suitable only in a scenario, when enough evidence (either in support or against a specific event) is available [7].

U. F. Minhas *et al.* [13] have proposed an expanded trust model for agents in the VANETs. In their model, they have incorporated role-based trust and experience-based trust, that are both combined into the priority-based model which can be used to choose proper advisers. After that, they use majority-opinion approach to aggregate feedback from selected advisers. During feedback aggregation, they also consider time and location closeness factors. In their model, they assume that roles are pre-defined by the authorities, and are expected to behave in a certain way. Furthermore, robustness has not been extensively addressed [7].

A. Patwardhan *et al.* [14] have proposed a data intensive reputation management scheme for the VANETs. In their model, they use persistent identities, frequency of encounters, and a known set of trustworthy anchored sources to serve as nucleating points for building trust relationships with previously unknown devices. Data is considered to be trustworthy, when there is an agreement among peers (majority consensus) or when it comes from the trustworthy source. During determining the majority consensus, their model does not consider reputation of the peers. Furthermore, authors assumed that each mobile device must have unique persistent identity that violates identity privacy.

C. Chen *et al.* [8] have proposed a trust modelling framework for message propagation and evaluation in the VANETs. In order to model quality of information shared by peers and the trust relationships between peers, they used trust opinions, experience-based trust and role-based trust metrics. Their trust model is binary (either fully trusted or not trusted). Inferring binary trust relationship is not always possible specially when we have incomplete information or when we are in uncertain situations. Furthermore, in their model, privacy and robustness has not been extensively addressed.

3 Proposed Method

As shown in the Figure 1, our proposed method works in three phases. In the first phase, receiver node will calculate its confidence value on each message that comes from the unique sender s_n about an event e. In the second phase, it will calculate the trust value for each unique message for an event e. Note that multiple senders can send same message related to a specific event. In the last phase, method will take the decision of an acceptance of a message, which has the highest trust value. Complete details about each phase are given below.

3.1 Confidence Measurement

Confidence shows the receiver's degree of belief on the data as well as the sender. We measure the confidence (C) value based on the following three parameters:

1. Location closeness (L_c),
2. Time closeness (T_c), and
3. Location verification (L_v).

Location Closeness: Location closeness factor determines the closeness of the sender to the reported event. We model the location closeness L_c as:

$$L_c = \begin{cases} 1 - \dfrac{\min(l_s, l_e)}{\max(l_s, l_e)} & \text{if } |l_s - l_e| < \delta l \\ \quad 1 & \text{otherwise} \end{cases} \tag{1}$$

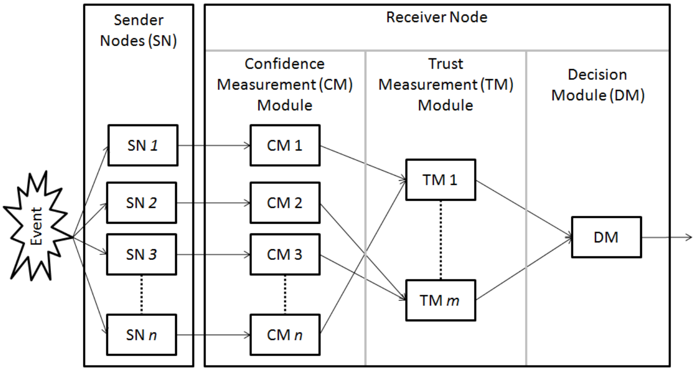


Fig. 1. Proposed Framework

where l_s and l_e represents the location of the source and event respectively. The δl represents a maximum acceptable threshold difference. This location closeness function is developed to keep the following intuitively described requirements.

- Property 1: When the difference between the sender location and event location increases then the location closeness factor also increases.
- Property 2: When the difference between the source location and the event location is more than a pre-defined threshold value then the location closeness factor becomes 1, which means data is not reliable.

The graph in Figure 2 is obtained by implementing the Equation 1 in the Matlab. This graph shows that the location factors increases with the increase in difference between the location of the source and event. This satisfy the property 1. Also, note that when the difference between source and event location is more than the pre-defined threshold value (which in this example is 50 unit), then the location factor becomes 1. This satisfy the property 2.

Time Closeness: Time closeness factor determines the freshness of the data. We model the time closeness T_c as:

$$T_c = \min\left(1, \frac{t_c - t_e}{\delta t}\right) \tag{2}$$

where t_c is the current time and t_e is the event time given in the message; the δt is a threshold time. This time closeness function is developed to keep the following intuitively described requirements.

- Property 1: When the difference between the event time and the current time increases then the time closeness factor also increases.
- Property 2: When the difference between the event time and the current time is more than a pre-defined threshold value then the time closeness factor becomes 1, which means data is outdated.

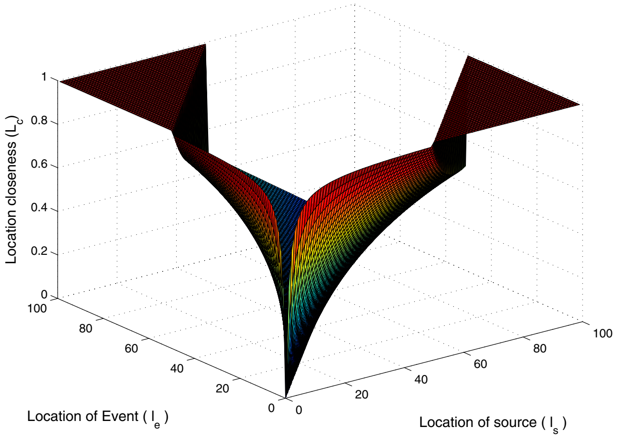


Fig. 2. Analysis of location closeness function: $\delta l = 50$ units

The graph shown in Figure 3 is obtained by implementing the Equation 2. Right side of the graph shows that; with the increase in current time with respect to the event time, the time closeness factor also increases linearly. This satisfy property 1. When current time is greater than 30, the time closeness factor becomes one. This happens because the difference between the current time and event time is more than the threshold value. This satisfy property 2.

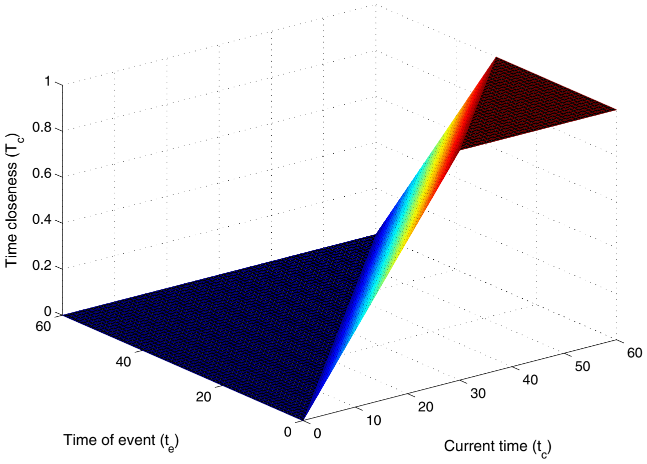


Fig. 3. Analysis of time closeness function: $\delta t = 30$ units

Location Verification: Location verification factor determines whether the sender node has provided its true location or not. Our proposed location verification mechanism (L_v) is described in the Algorithm 1. In this algorithm, first we estimates the region, where the sender is actually located, and then we determines whether the broadcasted coordinates are within the estimated region or not. Detail description of the proposed algorithm is given below.

Let us assume that each vehicle is equipped with a standard embedded device, in such a way that antennas, gains and transmission powers are fixed and known. Let d_{max} is the maximum radio range of the vehicle, and let θ is the angle of arrival of the received packet. How to measure θ is out of the scope of this paper. However, various standard techniques could be employed to measure θ. Whenever a node received a packet, it creates a potential region with the help of θ, and d_{max}, as shown in the Figure 4. Note that receiver's coordinates (x_c, y_c) are reference point. Once the boundaries of possible region are identified (Algo. 1, Lines 2:25), algorithm checks whether the broadcasted coordinates (x_s, y_s) of the sender are within that region or not (Lines 26:30). If (x_s, y_s) are within the identified region, it means the sender has provided true location (Lines 26:27). If (x_s, y_s) are located outside the identified region, it means the sender has provided fake location (Lines 28:29).

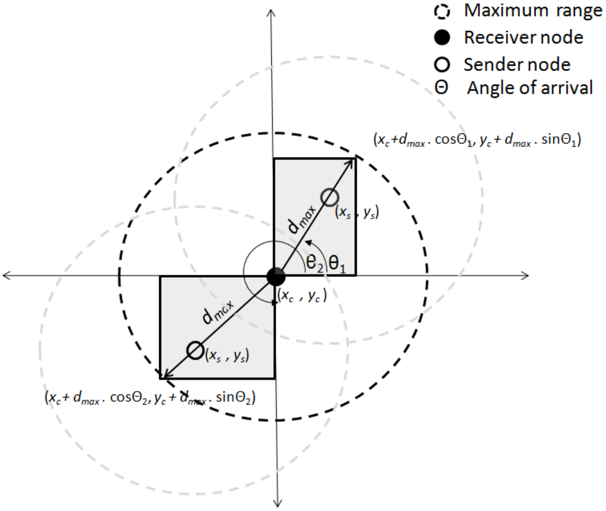


Fig. 4. Region estimation of sender node

Note that, instead of identifying the exact location of the sender, we identify the potential region of the sender. We adopted this approach to achieve simplicity. For better accuracy, other parameters such as signal strength could also be used; however, this will increase the complexity.

Algorithm 1. Location estimation and verification

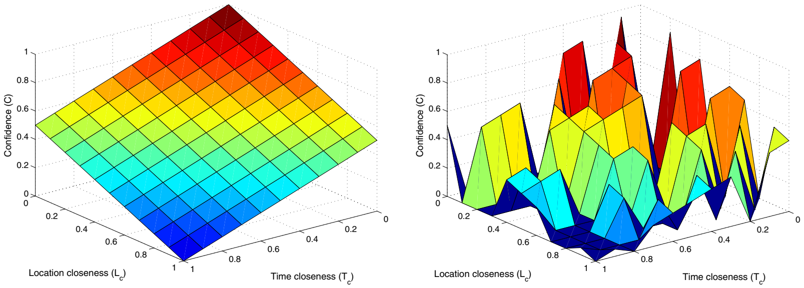
```
function $L_v(\theta, d_{max}, x_s, y_s, x_c, y_c)$
    Let $A$ is a potential region.
    Let $(x_l, y_l)$ are the lower coordinates of $A$.
    Let $(x_u, y_u)$ are the upper coordinates of $A$.
    if $\theta < 90°$ then
        $x_u = x_c + d_{max} \cdot \cos\theta$
        $y_u = y_c + d_{max} \cdot \sin\theta$
        $x_l = x_c$
        $y_l = y_c$
    else if $\theta > 90°$ & $\theta < 180°$ then
        $x_u = x_c$
        $y_u = y_c + d_{max} \cdot \sin\theta$
        $x_l = x_c + d_{max} \cdot \cos\theta$
        $y_l = y_c$
    else if $\theta > 180°$ & $\theta < 270°$ then
        $x_u = x_c$
        $y_u = y_c$
        $x_l = x_c + d_{max} \cdot \cos\theta$
        $y_l = y_c + d_{max} \cdot \sin\theta$
    else
        $x_u = x_c + d_{max} \cdot \cos\theta$
        $y_u = y_c$
        $x_l = x_c$
        $y_l = y_c + d_{max} \cdot \sin\theta$
    end if
    if $(x_l \leq x_s \leq x_u)$ and $(y_l \leq y_s \leq y_u)$ then
        return 1;                          ▷ Sender has provided true location
    else
        return 0;                          ▷ Sender has provided fake location
    end if
end function
```

Once all three factors (time closeness, location closeness and location verification) are determined, we calculate the confidence (C) value on a message x_k as follows:

$$C_{x_k} = \left(1 - \frac{L_c + T_c}{2}\right) \times L_v \tag{3}$$

This equation is developed to satisfy the following intuitively described requirements:

- Property 1: If the location closeness factor increases, then the confidence value decreases.
- Property 2: If the time closeness factor increases, then the confidence value decreases.
- Property 3: If the location verification check is fail, then the confidence level should be zero.



(a) $L_v = 1$

(b) L_v is randomly selected (0 or 1), with equal probability

Fig. 5. Confidence measurement analysis

The graphs shown in Figure 5 are obtained by implementing the Equation 3. These graphs illustrates that the required properties are retained in the Equation 3. The left side of the Figure 5(a) shows that; with the increase in location closeness factor the confidence value decreases. This satisfies property 1. The right side of the Figure 5(a) shows that; with the increase in time closeness factor the confidence value decreases. This satisfies property 2. The Figure 5(b) shows (e.g. the index (1,0.8)) that whenever the location verification method L_{verif} returns zero, the confidence value also becomes zero. This satisfies property 3.

3.2 Trust Measurement

Let X be the set of m unique messages (related to the same event) received from n nodes.

$$X = \{x_1, x_2, \dots, x_m\} \tag{4}$$

For each unique message x_k, we calculate the trust value in the following manner:

$$t_{x_k} = \frac{|x_k|}{n} \times \sum_{i=1}^{|x_k|} C_i \tag{5}$$

where t_{x_k} represents the trust value on message x_k, $|x_k|$ represents the total number of sender nodes who send the message x_k, and $\sum_{i=1}^{|x_k|} C_i$ represents the cumulative confidence value of all the nodes that send message x_k.

Proposition 1: The range of trust value is always between $\left[0, \frac{|x_k|^2}{n}\right]$.

Proof: Let us assume the worst scenario, in which the confidence C evaluation for each peer is zero. Substituting value of C with zero in equation 5 gives the minimum trust value as shown below.

$$t_{x_k} = \frac{|x_i|}{n} \sum_{i=1}^{|x_k|} C_i = \frac{|x_k|}{n} \sum_{i=1}^{|x_i|} 0 = 0$$

Now let us assume the best scenario, in which the confidence C evaluation for each peer is one. If we substitute C with 1 in equation 5, we get the following result.

$$t_{x_k} = \frac{|x_k|}{n} \sum_{i=1}^{|x_k|} C_i = \frac{|x_k|}{n} \sum_{i=1}^{|x_k|} 1 = \frac{|x_k|^2}{n}$$

This gives us the maximum trust value. □

3.3 Decision Logic

At the end of phase 2, we get m trust values as shown below.

$$T = [t_{x_1}, t_{x_2}, \ldots, t_{x_m}] \tag{6}$$

where each trust value corresponds to each unique message related to specific event.

After that, method will take the decision D based on the following logic.

$$D = \textbf{accept}(x_i \in X) | \forall j \;\; t_{x_i} > t_{x_j}, i \neq j \tag{7}$$

It states that accept message x_i that belongs to set X, such that for all values of j, the trust value of the message x_i must be greater than the trust values of the message x_j.

4 Analysis and Evaluation

4.1 Security Resilience Analysis

Definition 1: A message x_k is considered to be *untrustworthy* if:

1. $L_c = 1$ and $T_c = 1$, or
2. $L_v = 0$, or
3. 1 and 2 both.

Definition 2: A node is called *malicious* if it sends *untrustworthy* messages.

Proposition 2: The confidence value of a malicious node is 0.

Proof: From Equation 3, confidence value on message x_k is calculated as:

$$C_{x_k} = \left(1 - \frac{L_c + T_c}{2}\right) \times L_v$$

If x_k is untrustworthy message, then according to the Definition 1, L_c and T_c should be equal to 1. Substituting values of L_c and T_c with 1 in the above mentioned equation gives us the following result.

$$C_{x_k} = \left(1 - \frac{1+1}{2}\right) \times L_v = 0$$

□

Claim 1: Our proposed trust management scheme will not allow malicious nodes to increase the trust value of untrustworthy message.

Proof: Let us assume that, for an event e, node has received two types of messages (x_1 and x_2) from n different nodes. Assume that the message x_1 is received from non-malicious nodes and the message x_2 is received from malicious nodes. Malicious nodes will achieve their objective if:

$$t_{x_2} > t_{x_1}$$

This can also be written as:

$$\frac{|x_2|}{n} \times \sum_{i=1}^{|x_2|} C_i > \frac{|x_1|}{n} \times \sum_{i=1}^{|x_1|} C_i$$

Since, x_2 is received from a malicious node. So, according to the Proposition 2, the confidence value of a malicious node should be 0. Therefore, above inequality will transform into the following:

$$|x_2| \times \sum_{i=1}^{|x_2|} 0 > |x_1| \times \sum_{i=1}^{|x_1|} C_i$$

$$0 > |x_1| \times \sum_{i=1}^{|x_1|} C_i$$

However, this is a contradiction. Hence, it prove that the malicious nodes will not be able to increase the trust value of any untrustworthy message. □

Let M_{x_k} denotes the total number of malicious nodes which send the message x_k. As stated before, let $|x_k|$ represents the total number of sender nodes which send the message x_k. So, $M_{x_k} \leq |x_k|$, and $M_{x_k} > 0$.

Proposition 3: In the presence of malicious nodes, the maximum trust value, the method can assign to the message x_k is $\frac{(|x_k|-M_{x_k})^2}{n}$.

Proof: From Equation 5, we have

$$t_{x_k} = \frac{|x_k|}{n} \times \sum_{i=1}^{|x_k|} C_i$$

If M malicious nodes send message x_k, then the above mentioned equation will transform in the following:

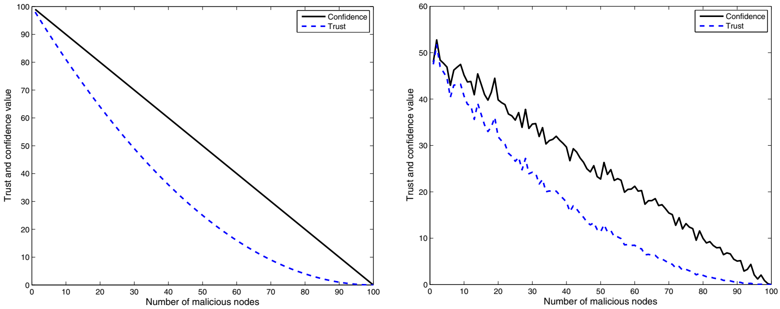
$$t_{x_k} = \frac{|x_k| - M_{x_k}}{n} \times \sum_{i=1}^{|x_k|-M_{x_k}} C_i \tag{8}$$

In this scenario, let us assume the best case, in which the confidence value of all non-malicious nodes is 1. So, $\sum_{i=1}^{|x_k|-M_{x_k}} C_i = |x_k| - M_{x_k}$. Substituting this value in the above-mentioned equation gives the following result.

$$t_{x_k} = \frac{|x_k| - M_{x_k}}{n} \times (|x_k| - M_{x_k}) = \frac{(|x_k| - M_{x_k})^2}{n}. \tag{9}$$

□

Figure 6 shows the behavior of the Equation 8 in two scenarios. In the first scenario, values of the L_c and T_c are set to 0, which means that the confidence value of all non-malicious nodes is 1. In the second scenario, values of the L_c and T_c are randomly selected between 0 and 1. Both graphs show that the trust value decreases with the increase in number of malicious nodes in the network.



(a) $L_c = T_c = 0$

(b) For each message, values of L_c & T_c are randomly selected b/w 0 & 1

Fig. 6. Effect of malicious nodes on trust & confidence values: $N = 100$, $L_v = 1$

Let $t^m_{x_k}$ represents the trust value for a message x_k that is obtained in the presence of malicious nodes.

Claim 2: For message x_k, $t_{x_k} > t^m_{x_k}$.

Proof: Let us prove this claim by contradiction. Assume that

$$t_{x_k} < t^m_{x_k}.$$

From proposition 1, the maximum trust value t_{x_k}, a message x_k can get is $\frac{|x_k|^2}{n}$. From proposition 3, the maximum trust value $t^m_{x_k}$, a message x_k can get is $\frac{(|x_k| - M_{x_k})^2}{n}$. Substituting both values in the above-mentioned inequality gives the following result.

$$\frac{|x_k|^2}{n} < \frac{(|x_k| - M_{x_k})^2}{n}$$
$$= |x_k|^2 < (|x_k| - M_{x_k})^2$$
$$= |x_k| < |x_k| - M_{x_k}$$

However, this is a contradiction, since, $|x_k|$ could not be less than the $|x_k| - M_{x_k}$, because $M_{x_k} > 0$. Thus it proves that, for the message x_k, the trust value obtained in the non-malicious environment will always be greater than the trust value obtained in the malicious environment. □

4.2 Time Complexity Analysis

As stated before, our proposed method works in three phases: 1) Confidence measurement phase, 2) Trust measurement phase, and 3) Decision phase. Let us first derive the time complexity of each phase, and then we will discuss the overall time complexity of the method.

In the confidence measurement phase, for each message, 6 operations (See Eq. 1) are required to calculate location closeness value, 3 operations (See Eq. 2) are required to calculate time closeness value, and k operations are required for location verification. Here, k represents the number of operations that are required to implement Algorithm 1. One can clearly see that the order of complexity of Algorithm 1 is $\mathcal{O}(1)$. After computing L_c, T_c, and L_{verif}, we compute the confidence value on the message x_k. For this, 4 operations are needed (See Eq. 3). So, for a single message, $6 + 3 + k + 4 = k + 13$ operations are required. Let us assume that node has received n messages related to single event e, so the total number of operations required by this phase is: $n(k + 13)$.Note that, here k is constant. So the asymptotic time complexity of this phase is $\mathcal{O}(n)$.

In the trust measurement phase, the number of operations that are required to calculate the trust value of a single message x_k received from $|x_k|$ nodes are: $2 + |x_k|$ (See Eq.5, which has 1 division, 1 multiplication and $|x_k|$ summation). Let us assume that node has received m unique messages from n nodes related to single event e. In that case, total number of operations, which are required to calculate m trust values are:

$$= (2 + |x_1|) + (2 + |x_2|) + \cdots + (2 + |x_m|)$$

Let us assume the worst scenario, in which all n nodes have sent different message. So $n = m$ and $|x_1| = |x_2| = \cdots = |x_m| = 1$. So, total operations that are required to calculate n trust values will be:

$$= (2 + 1) + (2 + 1) + \cdots + (2 + 1) = 3n$$

Hence, the time complexity for the trust measurement phase is also $\mathcal{O}(n)$.

In the decision phase, the proposed method needs to take a decision in favour of the particular message based on the trust values. Let us assume that the decision module has received n trust values for n unique messages related to a single event e. In order to decide, the proposed method will first sort n messages according to their corresponding trust values (from highest to lowest), and then accept the message x_k, which has the highest trust value. So, number of operations mainly depends on a sorting algorithm. There exist many sorting algorithms that run in linear time, e.g., Counting sort, Radix sort, and Bucket sort [18].

Note that the time complexity of all three modules is linear. Therefore, we can confidently say that our proposed method can compute trust and take decisions in real time.

5 Conclusion and Future Work

Due to high mobility and ephemeral nature of the vehicular networks, establishing and managing trust is a challenging task. Furthermore, if we want to ensure privacy of the user, then things become more complex. Existing trust management schemes that are proposed for the vehicular networks do not efficiently deal with the above-mentioned challenges. Therefore, we have proposed a new trust management scheme that overcomes these limitations. Our proposed method is completely decentralized that makes it easy to implement in the VANETs. Moreover, it is resilient against security attacks on the trust model itself. Furthermore, it has linear time complexity, which makes it suitable to use in real-time. Another, unique feature of the proposed method is that it operates in identity anonymous environment, which ensures user privacy.

The network topology and node density changes constantly and rapidly in the vehicular networks. So, what is its impact on the trust management? By performing simulation-based analysis and evaluation, we can find the answer of this question. This will be left to future research.

References

1. Jiang, D., Delgrossi, L.: IEEE 802.11p: Towards an international standard for wireless access in vehicular environments. In: Proc. of IEEE Vehicular Technology Conference, Singapore, pp. 2036–2040 (2008)
2. ETSI: European Telecommunications Standards Institute. News Release (September 2008), `http://www.etsi.org/WebSite/NewsandEvents/2008_09_Harmonizedstandards_ITS.aspx` (retrieved: June 12, 2012)
3. Schoch, E., Kargl, F., Weber, M., Leinmüller, T.: Communication patterns in vanets. IEEE Communications Magazine 46, 119–125 (2008)
4. Altan, O.D., Colgin, R.C.: Threat assessment algorithm for forward collision warning (2004)
5. Maïsseu, B.: IVHW:an inter-vehicle hazard warning system. In: Proc. of the International Workshop on Vehicle Safety Communications, Tokyo, Japan (2003)
6. Nadeem, T., Dashtinezhad, S., Liao, C., Iftode, L.: Trafficview: traffic data dissemination using car-to-car communication. ACM SIGMOBILE Mobile Computing and Communications Review 8, 6–19 (2004)

7. Zhang, J.: A survey on trust management for vanets. In: Proc. of the 2011 IEEE International Conference on Advanced Information Networking and Applications (AINA), Biopolis, Singapore, pp. 105–112 (2011)
8. Chen, C., Zhang, J., Cohen, R., Ho, P.H.: A trust modeling framework for message propagation and evaluation in vanets. In: Proc. of the 2nd International Conference on Information Technology Convergence and Services (ITCS), Cebu, Philippines, pp. 1–8 (2010)
9. Nekovee, M.: Vehicular communications and networks (June 4, 2010), `http://www.radio.feec.vutbr.cz/kosy/soubory/maziar/Nekovee_lecture_2_fulltext.pdf` (retrieved: May 12, 2012)
10. Shaikh, R.A., Jameel, H., d'Auriol, B.J., Lee, H., Lee, S., Song, Y.-J.: Group-based trust management scheme for clustered wireless sensor networks. IEEE Transaction on Parallel and Distributed Systems 20, 1698–1712 (2009)
11. Raya, M., Papadimitratos, P., Gligor, V., Hubaux, J.: On data-centric trust establishment in ephemeral ad hoc networks. In: Proc. of the 27th Conference on Computer Communications (INFOCOM 2008), Phoenix, USA, pp. 1238–1246 (2008)
12. Huang, D., Hong, X., Gerla, M.: Situation-aware trust architecture for vehicular networks. IEEE Communications Magazine 48, 128–135 (2010)
13. Minhas, U.F., Zhang, J., Tran, T., Cohen, R.: Towards expanded trust management for agents in vehicular ad-hoc networks. International Journal of Computational Intelligence: Theory and Practice (IJCITP) 5, 3–15 (2010)
14. Patwardhan, A., Joshi, A., Finin, T., Yesha, Y.: A data intensive reputation management scheme for vehicular ad hoc networks. In: Proc. of the 3rd Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, California, USA, pp. 1–8 (2006)
15. Gómez Mármol, F., Martínez Pérez, G.: Trip, a trust and reputation infrastructure-based proposal for vehicular ad-hoc networks. Journal of Network and Computer Applications 35, 934–941 (2012)
16. Huang, Z., Ruj, S., Cavenaghi, M., Nayak, A.: Limitations of trust management schemes in vanet and countermeasures. In: Proc. of the IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), Toronto, ON, Canada, pp. 1228–1232 (2011)
17. Gerlach, M., Guttler, F.: Privacy in vanets using changing pseudonyms-ideal and real. In: Proc. of the IEEE 65th Vehicular Technology Conference (VTC 2007 Spring), Dublin, Ireland, pp. 2521–2525 (2007)
18. Pandey, H.: Design Analysis and Algorithms. Laxmi Publications Pvt. Ltd. (2008)

Feature and Future of Visual Cryptography Based Schemes

Dhiraj Pandey[1], Anil Kumar[2], and Yudhvir Singh[3]

[1] JSS Noida
[2] CSE Deptt.,
Manipal University Jaipur
[3] MDU Rohtak
University Instt.of Engg.&Technology
{dhip2,dahiyaanil}@yahoo.co.in, yudhvirsingh@rediffmail.com

Abstract. Visual cryptography (VC) is a useful technique that combines the notions of perfect ciphers and secret sharing in cryptography. VC takes a binary image (the secret) and divides it into two or more pieces known as shares. When the shares are printed on transparencies and then superimposed, the secret can be recovered. No computer participation is required. There are various measures on which performance of visual cryptography scheme depends, such as pixel expansion, contrast, security, accuracy, computational complexity, share generated is meaningful or meaningless, type of secret images (either binary or color) and number of secret images(either single or multiple) encrypted by the scheme. In this paper, we will summarize the developments of visual cryptography since its birth in 1994, introduce the main research topics in this area where researchers have been contributing and outline the application of these schemes.

Keywords: Visual cryptography scheme (VCS), pixel expansion, contrast, security, accuracy, computational complexity.

1 Introduction

With the rapid advancement of network technology, most of the information is transmitted over the Internet conveniently. Various confidential data such as military maps and commercial secrets are transmitted over the Internet. While using secret images, security issues should be taken into consideration because hackers may utilize weak link over communication network to steal information that they want. To deal with the security problems of secret images, various image secret sharing schemes have been developed. Visual cryptography was introduced first in 1994 Naor and Shamir [1]. Visual cryptography is a cryptographic technique which allows visual information (e.g. printed text, handwritten notes and pictures) to be encrypted in such a way that the decryption can be performed by the human visual system, without the aid of computers. Visual cryptography scheme eliminates complex computation problem in decryption process, and the secret images can be restored by stacking operation. This property makes visual cryptography especially useful for the low computation load requirement.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 816–830, 2013.

This paper covers the progress of VC, along with the current trends and the various applications for VC. When the data is hidden within separate images (known as shares), it is completely unrecognizable. While the shares are separate, the data is completely unrelated. Each image holds random images and when they are brought together, the secret can be recovered easily. They each depend on one another in order to obtain the original image.

This paper is organized as follows: Section 2 elaborates on the work being done in this area, specifically the most recent improvements. In general, these schemes primarily deal with binary images and noisy random shares. Extended VC is also presented within this section. Section 3 discuss on cheating prevention within VC along with cheating immune VC schemes. These schemes attempt to have some type of authentication or verification method. Grayscale, halftone and color halftone images used with visual cryptography are presented in Section 4. Section 5 elaborates on multiple secret sharing, which involves sharing two or more secrets, typically within a set of two shares. Various applications of visual cryptography are analyzed in Section 6 along with performance analysis in section 7 and the summary is discussed within Section 8, along with the final conclusion.

2 Traditional Visual Cryptography

2.1 Basic Visual Cryptography

Image sharing is a subset of secret sharing because it acts as a special approach to the general secret sharing problem. The secrets in this case are concealed images. Image sharing defines a scheme which is identical to that of general secret sharing. In (k, n) image sharing, the image that carries the secret is split up into n pieces (known as shares) and the decryption is totally unsuccessful unless at least k pieces are collected and superimposed. Visual cryptography was pioneered by Moni Naor and Adi Shamir in 1994 at the Eurocrypt conference [2]. When the k shares are stacked together, the human eyes do the decryption. This allows anyone to use the system without any knowledge of cryptography and without performing any computations whatsoever. This is another advantage of visual cryptography over the other popular conditionally secure cryptography schemes.

Naor and Shamir's initial implementation assumes that the image is a collection of black and white pixels. One disadvantage of this is that the decryption process is lossy; in terms of contrast. Contrast is very important within visual cryptography because it determines the clarity of the recovered secret by the human visual system. The relative difference in hamming weight between the representation of white and black pixels signify the loss in contrast of the recovered secret. Newer schemes that are discussed later deal with grayscale and color images which attempt to minimize the loss in contrast [3] by using digital halftoning. Due to the fact that digital halftoning is a lossy process in itself [4], it is impossible to fully reconstruct the original secret image.

The Hamming weight H (V) of the ORed m-vector V is interpreted by the visual system as follows: A black pixel is interpreted if H (V) <= d and white if H (V) <d-αm for some fixed threshold 1<=d<=m and a relative difference α>0.

The construction of the shares can be clearly illustrated by a 2 out of 2 visual cryptography scheme (commonly known as (2, 2)-VCS). The following collections of 2 * 2 matrices are defined:

$C0$ = all the matrices obtained by permuting the columns of $\begin{bmatrix}1\ 1\ 0\ 0\\1\ 1\ 0\ 0\end{bmatrix}$

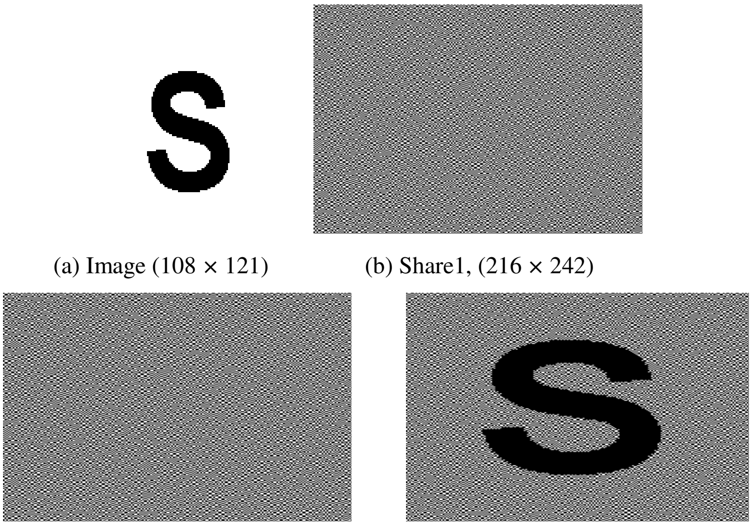
$C1$ = all the matrices obtained by permuting the columns of $\begin{bmatrix}1\ 1\ 0\ 0\\0\ 0\ 1\ 1\end{bmatrix}$

Due to this pixel expansion, one pixel from the original image gets expanded into four pixels. The shares can be generated in the following manner:

1. If the pixel of the original binary image is white, randomly pick the same pattern of four pixels for both shares.
2. If the pixel of the original image is black, pick a complementary pair of patterns.

When the transparencies are superimposed and the sub-pixels are correctly aligned, the black pixels in the combined shares are represented by the Boolean OR of the rows in the matrix.

Below in Figure 1, the implementation and results of (2, 2)-VCS basic visual cryptography are shown. It displays the secret image, the two shares that are generated and the recovery of the secret after superimposing share one and share two.



(a) Image (108 × 121)

(b) Share1, (216 × 242)

(c) Share 2(216 × 242)

(d) $S1 + S2$ (216 × 242)

Fig. 1. The results of a traditional visual cryptography scheme

2.2 Extended Visual Cryptography

The main difference between basic visual cryptography and extended visual cryptography is that a recognizable image can be viewed on each of the shares; once the shares have been superimposed the image on the shares will disappear and the secret message will be visible. This is the basis for the extended form of visual cryptography. An extended scheme proposed by Ateniese et al. [5] is based on an access structure which contains two types of sets. Extended visual cryptography schemes allow the construction of visual secret sharing schemes within which the shares are meaningful as opposed to having random noise on the shares.

In EVCS, the first n shares need to be images of something like a car, boat or, some form of meaningful information. The secret message is normally the $(n + 1)$ th message. This requires a technique that has to take into consideration the color of the pixel in the secret image we want to obtain, so when the n shares are superimposed, their individual images disappear and the secret image can be seen.

Three conditions must be satisfied for encrypting the images. Firstly, images that belong to the qualified set access structure, should, when superimposed, reveal the secret image. Secondly, by inspecting the shares, no hint should be available about what secret is hidden within the shares. Finally, the image within the shares should not be altered in anyway.

2.3 Size Invariant Visual Cryptography

Image size invariant VC was proposed by Ito et al. [6]. Traditional visual cryptography schemes employ pixel expansion, although many researchers have worked on how to improve problem of pixel expansion [7]. Ito's scheme [6] removes the need for this pixel expansion. The scheme uses the traditional (k, n) scheme where m is equal to one. The recovered secret can be viewed as the difference of probabilities with which a black pixel in the reconstructed image is generated from a white and black pixel in the secret image. The most important part of any VC scheme is the contrast. The lower the contrast of the recovered secret, the harder it is to visually recover the secret. The contrast for this scheme is defined as follows: $\beta = |p0 - p1|$, where $p0$ and $p1$ are the probabilities with which a black pixel on the reconstructed image is generated from a white and black pixel on the secret image. Another method to deal with size invariant shares is proposed in [8] in which the frequency of white pixels is used to show the contrast of the recovered image. The scheme is non-expansible. Many researchers have examined size invariant schemes [9,10]. Aspect ratio is a related topic within size invariant schemes. Yang and Chen [11] has presented aspect ratio invariant secret sharing. Aspect ratio invariant secret sharing scheme reduces the number of extra sub pixels needed. A size-adjustable scheme is presented by Yang et al. [12] that allow choosing an appropriate share size. If quality and contrast matter then the size of the shares will increase, for a user's particular application.

2.4 Integrated Review of Basic Schemes

VC remains an important research topic from its inception in 1994. More specifically, the schemes which minimize pixel expansion and also increase the overall contrast, which results in very clear secret recovery. The size adjustable scheme discussed above, which allows the user to specify what size of shares to generate is very interesting work. This allows for a user defined tradeoff between quality and portability of shares. This increases the potential for VC once again, rather than being restricted on a specific scheme which only allows for a certain type of quality. Optimal contrast secret sharing schemes in visual cryptography have been discussed at length because it is an extremely important evaluation metric for any scheme. This is mainly due to how the overall contrast affects the quality of the recovered secret. An approach based on coding theory helps to provide an optimal tradeoff between the contrast and the number of sub pixels. Optimal (2*, n*)-schemes are examined in terms of contrast related to the Hamming distance, as well as the sub pixels tradeoff required for these optimal schemes. A possible option for improving the efficiency of VC is to use the XOR operation [13]. This method will not allow traditional stacking of the shares on transparencies but it will improve the overall share quality. The scheme has favorable properties, such as, good resolution and high contrast. It can be applied to color images as well. The downside to some of these basic forms of VC is that the shares potentially give away the fact that they are encrypted. Extended VC helps with this, producing meaningful shares which have the same pixel expansion as the original basic VC schemes.

3 Cheating Immune Visual Cryptography

The cheating process could cause damage to victims because they will accept a forged image different from the actual secret image as authentic. Many researchers have experimented with the idea of cheating the system and suggested solution for its prevention also. Methods for cheating the basic VC schemes have been presented, along with techniques used for cheating extended VC schemes [14,15,16].

3.1 Authentication Methods

Prevention of cheating via authentication methods [16] has been proposed which focus on identification between two participants to help prevent any type of cheating taking place. Yang and Laih [16] presented two types of cheating prevention; one type used an online trust authority to perform the verification between the participants. The second type involved changing the VC scheme whereby the stacking of two shares reveals a verification logo; however this method requires the addition of extra pixels in the secret. Another cheating prevention scheme is described by Horng et al. [14]. If an attacker knows the exact distribution of black and white pixels of each of the shares of honest participants then they will be able to successfully attack and cheat the scheme. Method which prevents the attacker from obtaining this distribution can be useful to prevent cheating.

3.2 Cheat Prevention

Hu and Tzeng [17] were able to present numerous cheating methods, each of which where capable of cheating Horng et al.'s cheating prevention scheme. They also present improvements on Yang scheme and finally present their own cheating prevention scheme which attempts to minimize the overall additional pixels which may be required. No online trust authority is required and the verification of each image is different and confidential. The contrast is minimally changed and the cheating prevention scheme should apply to any VCS. Hu and Tzeng where also able to prove that both a malicious participant (MP), that is MP $\in P$, and a malicious outsider (MO), MO $/\in P$, can cheat in some circumstances. The MP is able to construct a fake set of shares using his genuine share. After the fake share has been stacked on the genuine share, the fake secret can be viewed. The second cheating method involving an MO is capable of cheating the VC scheme without having any knowledge of any genuine shares. The MO firstly creates a set of fake shares based on the optimal (2, 2)-VCS. Next, the fake shares are required to be resized to that of the original genuine shares size.

3.3 A Traceable Model

A traceable model of visual cryptography [18] was also examined which also helps to deal with cheating. It deals with the scenario when a coalition of less than *k* traitors who stack their shares and publish the result so that other coalitions of the participants can illegally reveal the secret. In the traceable model, it is possible to trace the saboteurs with the aid of special markings. The constructions of traceable schemes for both (*k, n*) and (*n, n*) problems were also presented.

3.4 Quality Evaluation

Most notable improvements on cheating immune VC schemes have been presented within [17] which present examples for traditional and extended schemes. The pixel expansion and contrast reduction are minimal and acceptable due to the overall improvements presented within [17]. The addition of an authentication method, whereby, each participant must verify every other participant is an important improvement. The drop in contrast is very slight when compared to previous schemes. The overall quality that has gone into this scheme is highly impressive and extremely useful.

4 Grayscale, Halftone and Colour Visual Cryptography

It is important to understand how halftoning technologies work, as they are frequently used within many visual cryptography schemes. Halftoning is a print and display technique that trades area for gray-level depth by partitioning an image into small areas in which pixels of different values are purposely arranged to reflect the tone density. There are three main factors that effect these arranged pixels or dot structure. In conjunction, error diffusion techniques coincide with halftone technology. Error diffusion is an adaptive technique that quantizes each pixel according to the input pixel as well as its neighbors. Error diffusion forces total tone content to remain the same and attempts to localize the distribution of tone levels [19]. At each pixel, the errors from its preceding neighbours are added to the original pixel value. This modified value then has a threshold applied to it.

4.1 Grayscale and Halftone Visual Cryptography

This is an extension on Naor and Shamir's original findings in the 2-out-of-2 secret sharing scheme. It also takes extended visual cryptography a step further. The halftoning technique that is used can be applied to colour and grayscale images.

Grayscale halftoning is discussed within this section. Section 4.2 details colour halftone visual cryptography.

Based on the idea of extended visual cryptography, Zhou et al. [20] set about improving these techniques by proposing halftone grayscale images which carry significant visual information. This in itself drastically improves the security model for visual cryptography. Along with Zhou, [21,22,23] present novel techniques by which halftone images can be shared with significant visual meaning which have a higher quality than those presented within [24] by employing error diffusion techniques . These error diffusion techniques spread the pixels as homogeneously as possible to achieve the improvements in the shares overall quality. A halftone scheme [25] was proposed in which the quality of the shares is improved by using contrast enhancement techniques. However the problem with this scheme is that it is not perfectly secure.

The method proposed by Myodo et al. [22] allows natural embedding of grayscale images. The quality of the superimposed image highly depends on its dynamic range and pixel density. The possible pixel density of the superimposed image can be defined as: $max(0, g`_1+g`_2 -1) < ds < min(g`_1, g`_2)$, where $g`_1$ and $g`_2$ are pixel values of the dynamic-range-controlled input images and ds is the pixel density of the superposed image that is estimated with the surrounding pixels. The equation indicates that $g`_1 = g`_2 = 0.5$ gives the widest dynamic range of the superimposed image. Therefore, pixel values of input images should be modified around 0.5 by reducing their dynamic range. Accordingly, each pixel value of a secret image should be restricted between 0 and 0.5. This provides the mechanism for allowing any grayscale natural image to be used as an input.

The conventional method described in [26] uses an error diffusion halftoning technique [27] which works as follows: two grayscale images are used for input along with a secret image. Typically, the secret image cannot be used as an input image so a ternary image is used as input in its place. The output images (that carry the secret) are binary images. Firstly, image 1 is taken and an error diffusion process is applied to it (giving share 1). Image 2 then has an image hiding error diffusion process applied. During this image hiding error diffusion process, pixels from image 2 are modulated by corresponding pixels of share 1 and the secret image in order to embed the secret into the resultant share of image 2 (giving share 2). The secret is recovered by superimposing share 1 and share 2. The previously discussed VC schemes all suffer from pixel expansion in that the shares are larger than the original secret image. Chen et al. [28] present a secret sharing scheme that maps a block in a secret image onto a corresponding equal-sized block in the share image without this pixel expansion. Two techniques which are discussed include histogram width-equalization and histogram depth-equalization. This scheme improves the quality of the reconstructed secret when compared with alternative techniques. Another scheme proposed by Wang et al. [29] uses only Boolean operations. The contrast is also higher than other probabilistic visual cryptography sharing schemes. The area of contrast within halftone and grayscale VC is an interesting one because the contrast determines exactly how clear the recovered visual secret is. Cimato et al. [30] developed a visual cryptography scheme with ideal contrast by using a technique known as reversing, which was originally discussed by [31]. Reversing changes black pixels to white pixels and vice-versa. Viet and Kurosawa's scheme allows for perfect restoration of the black pixels but only almost perfect restoration of the white pixels. Cimato et al. provide their results for perfect restoration of both black and white pixels. Each share also contained a smaller amount of information than Viet and Kurosawa's which makes it a more desirable and secure scheme. Yang et al. [32] also looked at reversing and the shortcomings of Viet and Kurosawa's scheme. Their work presented a scheme that allowed perfect contrast reconstruction based on any traditional visual cryptography sharing scheme.

4.2 Colour Visual Cryptography

Applying visual cryptography techniques to colour images is a very important area of research because it allows the use of natural colour images to secure some types of information. Due to the nature of a colour image, this again helps to reduce the risk of alerting someone to the fact that information is hidden within it. It should also allow high quality sharing of these colour images. Colour images are also highly popular and have a wider range of uses when compared to other image types. In 1996, Naor and Shamir published a second article on visual cryptography "Visual Cryptography II: Improving the Contrast via the Cover Base" [33]. The new model contains several important changes from their previous work; they use two opaque colours and a completely transparent one. The first difference is the order in which the transparencies are stacked. There must be an order to correctly recover the secret. Therefore each of the shares needs to be pre-determined and recorded so recovery is possible. The second change is that each participant has *c* sheets, rather than a single transparency. Each sheet contains red, yellow and transparent pixels. The reconstruction is done by merging the sheets of participant I and participant II, i.e. put the *i*-th sheet of II on top of the *i*-th sheet of I and the $(i + 1)$-th of I on top of the *i*-th of II.

Efficiency within colour visual cryptography is also considered which improves on the work done by [34]. The proposed scheme follows Yang and Laih's colour model. The model considers the human visual system's effect on colour combinations out of a set of colour sub-pixels. This means that the set of stacked colour sub-pixels would look like a specific colour in original secret image. As with many other visual cryptography schemes, pixel expansion is an issue. However Shyu's scheme has a pixel expansion of $log2c$ which is superior to many other colour visual cryptography schemes especially when *c*, the number of colours in the secret image becomes large. An area for improvement however would be in the examination of the difference between the reconstructed colour pixels and the original secret pixels. Having high quality colour VC shares would further improve on the current schemes examined within this survey; this includes adding a lot of potential for visual authentication and identification.

In most colour visual cryptography schemes, when the shares are superimposed and the secret is recovered, the colour image gets darker. This is due to the fact that when two pixels of the same colour are superimposed, the resultant pixel gets darker.

The annoying presence of the loss of contrast makes traditional visual cryptography schemes practical only when quality is not an issue which is relatively rare. Therefore, the basic scheme is extended to allow visual cryptography to be directly applied on grayscale and colour images. Image halftoning is employed in order to transform the original image from the grayscale or colour space into the monochrome space which has proved to be quite effective.

This idea of progressive visual cryptography has recently been extended [35] by generating friendly shares that carry meaningful information and which also allows decryption without any computation at all. Purely stacking the shares reveals the secret.

4.3 Evaluation of Grayscale and Color Schemes

Grayscale, halftone and colour image techniques for visual cryptography provide an important step for the improvement of VC. The best results are obtained when using error diffusion techniques. These results also provide excellent secret recovery because the contrast is high. Using colour images has also improved the potential application for VC, particularly when using computer-specific progressive VC techniques; perfect secret recovery is possible with very high quality colour images and relatively low computational power. However, as discussed, use of computation partially defeats the point of VC. So a tradeoff must be made in order to obtain good recovered secrets and have suitable quality in the meaningful shares.

5 Multiple Secret Sharing in Visual Cryptography

5.1 Basic Multiple Secret Sharing

Multiple secret sharing has the main advantage of being able to hide more than one secret within a set of shares. The multiple secret sharing problems was initially examined by Wu and Chen [36]. They concealed two secrets within two sets of shares $S1$ and $S2$. The first secret is revealed when $S1$ and $S2$ are superimposed. The second becomes available when $S1$ is rotated anti-clockwise 90° and superimposed on $S2$. Due to the nature of the angles required for revealing the secrets (90°, 180° or 270°) and the fact that this scheme can only share, at most, two secrets, it becomes apparent that it is quite limited in its use.

It is also worth noting that another extended form of secret sharing was proposed [37] that is quite similar to the one discussed which involves stacking the transparencies to reveal a different secret each time a new layer is stacked. An improvement on this extended scheme is achieved by reducing the number of subpixels required [38]. Multiple secret sharing was developed further [39] by designing circular shares so that the limitations of the angle ($\theta = 90, 180, 270°$) would no longer be an issue. The secrets can be revealed when $S1$ is superimposed on $S2$ and rotated clockwise by a certain angle between 0° and 360°. A further extension of this was implemented [40] which defines another scheme to hide two secret images in two shares with arbitrary rotating angles. This scheme rolls the share images into rings to allow easy rotation of the shares and thus does away with the angle limitation of Wu and Chen's scheme. The recovered secrets are also of better quality when compared to [39], this is due to larger difference between the black and white stacked blocks. More recently [41] a novel secret sharing scheme was proposed that encodes a set of $x \geq 2$ secrets into two circle shares where x is the number of secrets to be shared. This is one of the first set of results presented that is capable of sharing more than two secrets using traditional visual cryptography methods. The algorithms presented can also be extended to work with grayscale images by using halftone techniques. The expansion is twice the number of secrets to be hidden, so the size of the circle shares increases dramatically when many large secrets are hidden. However, the number of secrets that are contained within the shares still remains a secret unless supplementary lines are added to the circle shares to ease the alignment.

5.2 Quality Evaluation of Sharing Multiple Secrets

Sharing multiple secrets with high quality recovery is very achievable. Depending on the number of secrets a user wishes to hide, this determines the overall size of the shares. The more secrets a user wishes to hide, the larger the resultant shares get. This is one of the shortcomings of multiple secret sharing, the final share size when many large secrets are considered can become unmanageable. Numerous schemes are presented which range from sharing just two secrets to the general case of sharing any number of secrets. Of the schemes presented, circular shares seem to be best in terms of the secrets recovery and contrast. The scheme presented for sharing more than two secrets using standard rectangular shares has issues with contrast while more secrets are added. Using a colour cover image also presents an effective way to share multiple smaller secrets.

Overall, the majority of the multiple secret sharing schemes are successful in effectively hiding two or more secrets with a set of shares. The schemes that roll the secrets into circular shares prove to be the most interesting and effective in terms of sharing many secrets with very high contrast.

6 Visual Cryptography Applications

Visual cryptography applications range from banking industry, satellite imaging to commercial application for preserving collected biometric data.

Practical uses for visual cryptography come in the form of watermarking. Memon and Wong [42] propose various techniques by which these watermarks can be applied to images. Similarly [43] also explores the use of watermarks within visual cryptography. A digital image copyright scheme based on visual cryptography is presented within [44]. It is simple and efficient, both in watermark embedding and retrieval. It is also acceptably robust when the watermarked image is compressed. Robust recovery of the watermark is also possible after the image has been defaced. As with the other schemes previously discussed, this scheme is also key dependant. Without the key, no watermark recovery is possible. One of the most robust ways to hide a secret within natural images is by typically employing visual cryptography based on halftone techniques. The perfect scheme is extremely practical and can reveal secrets without computer participation. VC potentially making it applicable to a wider range of secure applications, such as within the banking industry.

7 Performance Analysis

Various parameters are recommended by researchers to evaluate the performance of visual cryptography scheme. Naor and Shamir [1] suggested two main parameters, pixel expansion m and contrast. Pixel expansion m refers to the number of subpixels in the generated shares that represents a pixel of the original input image. It represents the loss in resolution from the original picture to the shared one.Security is satisfied if each share reveals no information of the original image and the original image cannot be reconstructed if there are fewer than k shares collected. Accuracy is considered to be the quality of the reconstructed secret image and evaluated by peak signal-to-noise ratio (PSNR) measure. Computational complexity concerns the total number of operators required both to generate the set of n shares and to restructure the original secret image .Jen-Bang Feng et al[9] suggested that VCS should support multiple secret to work efficiently. If scheme support only one secret to share at a time to share multiple secret images numerous share have to be generated, transmitted and maintained.

Table 1. Comparison of Visual Cryptography Schemes on the Basis of Number of Secret Images, Pixel Expansion, Image format, Type of Share Generated

Sr. No.	Authors	Year	No. of Secret Images	Pixel Expansion	Image Format	Type of Share generated
1	Naor and Shamir	1995	1	4	Binary	Random
2	Wu and chen	1998	2	4	Binary	Random
3	Hsu et al.	2004	2	4	Binary	Random
4	Wu and cheng	2005	2	4	Binary	Meaningful
5	Chin-chen chang	2005	1	4	Binary	Random
6	Liguo fang	2006	1	2	Binary	Random
7	s. j. shyu	2007	n>=2	2n	Binary	Random
8	w.p.fung	2007	2	9	Binary	Random
9	Jen bang	2008	n>=2	3n	Binary	Random
10	Mustafa Ulutas	2008	2	4	Binary	Random
11	Tzung-Her Chen et al.	2008	2	1	Binary	Random
12	Chen et al.	2008	n(n>=2)	4	Binary, gray, color	Random
13	Zhengxin Fu	2009	4	9	Binary	Random
14	Jonathan Weir et al.	2009	n	4	Binary	Random
15	Xiao-qing Tan	2009	1	1	Binary	Random
16	Verheul Tilborg	1997	1	C*3	Color	Random
17	Yang & Liah	2000	1	C*2	Color	Radom
18	Chin Chen Chang et al.	2002	1	9	Gray	Meaningful
19	Haibo Zhang et al.	2008	1	1	Gray	Random
20	Jonathan Weir	2011	1	4	Binary, gray, color	Hatched

As shown in the Table 1 only few visual cryptography schemes achieve minimum pixel expansion. Less overhead for storage and transmission is required to share multiple secrets while using the scheme [7,9]. Meaningful shares [4, 18] can be helpful to avoid attacks by hacker. Scheme supporting color images and gray [16, 17, 18, 19] are useful in the multimedia environment.

8 Conclusion and Further Direction

Many of the schemes presented work extremely well and the current state of the art techniques have proven to be very useful for many applications, such as verification and authentication.

The following trends have been identified within visual cryptography:

1. Contrast improvement.
2. Share size improvement.
3. Wider range of suitable image types (binary to colour images).
4. Efficiency of VC schemes.
5. Ability to share multiple secrets.

Essentially the most important part of any VC scheme is the contrast of the recovered secret from a particular set of shares. Ideal schemes provide a high contrast when the secret has been recovered. However, a tradeoff is required in some schemes depending on the size of the shares along with the number of secrets which may be concealed. Especially within extended visual cryptography schemes, contrast is of major importance. Making sure the base images completely disappear and a clear secret is recovered which could be another high quality image is vitally important. Some schemes present methods which do not work with printed transparencies and these rely on computation in order to recover the secret. In this respect, high quality secret recovery is possible, however it is preferred if the scheme works with printed transparencies. After all, this is the idea behind VC. Conversely, if an application requires digital recovery of the secrets, then perfect recovery can be achieved via the XOR operation. Having shares that are close to the original secret's size is best, because it results in shares that are easier to manage and transmit. Large secrets with even larger shares become cumbersome. However, at times a tradeoff must be made between the size of the shares and the contrast of the recovered secret. The tradeoff between size and the secret recovery must be suitable so that high quality recovery can take place and must also ensure that the shares do not expand into large, unmanageable sizes.

The use of grayscale and colour images has added value to the field of visual cryptography. Reducing the requirements on input image type so that any kind of image can be used to share a secret is very important. The fact that any image can be used to share a secret within visual cryptography shows a great improvement on the very initial work that required an image to be converted to its binary equivalent before any processing could be done on it. However, the application of the scheme depends greatly on the type of images to be input. Efficiency covers a number of things which have already been discussed, such as contrast and share size. The topic of efficiency also includes how the shares and images have been processed. Numerous methods presented within this paper have improved on prior work and techniques, resulting in schemes that are highly efficient and very simple to implement and use. For the maximum efficiency in recovering the secret, no computer participation should be involved.

Overall, this paper has summarized much of the work done in the area of visual cryptography. There are still many topics worth exploring within VC to further expand on its potential in terms of secret sharing, data security, identification, and authentication.

The previously mentioned trends that have emerged within VC require more attention. This allows VC to remain an important research topic. The focus being, to apply these techniques in conjunction with modern day image hatching techniques which would allow the extension of VC into the currency domain, potentially making it applicable to a wider range of secure applications, such as within the banking industry. The use of these types of shares within the secure printing industry should also be considered. Scanning a share into a computer system and then digitally superimposing its corresponding share could also be considered.

References

1. Shamir, A.: How to share a secret. Communications of the ACM 22(11), 612–613 (1979)
2. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
3. Blundo, C., D'Arco, P., De Santis, A., Stinson, D.R.: Contrast optimal threshold visual cryptography schemes. SIAM Journal on Discrete Mathematics 16(2), 224–261 (2003)
4. Lau, D.L., Arce, G.R.: Modern Digital Halftoning. Marcel Dekker, New York (2000)
5. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Extended schemes for visual cryptography. Theoretical Computer Science 250, 1–16 (1996)
6. Ito, R., Kuwakado, H., Tanaka, H.: Image size invariant visual cryptography. EICE Transactions E82-A(10), 2172–2177 (1999)
7. Tzeng, W.G., Hu, C.M.: A new approach for visual cryptography. Designs, Codes and Cryptography 27(3), 207–227 (2002)
8. Yang, C.N.: New visual secret sharing schemes using probabilistic method. Pattern Recognition Letters 25(4), 481–494 (2004)
9. Yang, C.N., Chen, T.S.: New size-reduced visual secret sharing schemes with half reduction of shadow size. IEICE Transactions 89-A(2), 620–625 (2006)
10. Yang, C.-N., Chen, T.-S.: Visual secret sharing scheme: Improving the contrast of a recovered image via different pixel expansions. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4141, pp. 468–479. Springer, Heidelberg (2006)
11. Yang, C.N., Chen, T.S.: Aspect ratio invariant visual secret sharing schemes with minimum pixel expansion. Pattern Recognition Letters 26(2), 193–206 (2005)
12. Yang, C.N., Chen, T.S.: Size-adjustable visual secret sharing schemes. IEICE Transactions 88-A(9), 2471–2474 (2005)
13. Tuyls, P., Hollmann, H.D.L., van Lint, J.H., Tolhuizen, L.M.G.M.: XOR-based visual cryptography schemes. Designs, Codes and Cryptography 37(1), 169–186 (2005)
14. Horng, G., Chen, T., Tsai, D.S.: Cheating in visual cryptography. Des. Codes Cryptography 38(2), 219–236 (2006)

15. Naor, M., Pinkas, B.: Visual authentication and identification. In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 322–336. Springer, Heidelberg (1997)
16. Yang, C., Laih, C.: Some new types of visual secret sharing schemes, vol. III, pp. 260–268 (December 1999)
17. Hu, C.M., Tzeng, W.G.: Cheating prevention in visual cryptography. IEEE Transactions on Image Processing 16(1), 36–45 (2007)
18. Biehl, I., Wetzel, S.: Traceable visual cryptography. In: Han, Y., Quing, S. (eds.) ICICS 1997. LNCS, vol. 1334, pp. 61–71. Springer, Heidelberg (1997)
19. Kang, H.R.: Digital Color Halftoning. In: Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, WA, USA (1999)
20. Zhou, Z., Arce, G.R., Crescenzo, G.D.: Halftone visual cryptography. IEEE Transactions on Image Processing 15(8), 2441–2453 (2006)
21. Myodo, E., Sakazawa, S., Takishima, Y.: Visual cryptography based on void-and cluster halftoning technique. In: ICIP, pp. 97–100 (2006)
22. Myodo, E., Takagi, K., Miyaji, S., Takishima, Y.: Halftone visual cryptography embedding a natural grayscale image based on error diffusion technique. In: ICME, pp. 2114–2117 (2007)
23. Wang, Z., Arce, G.R.: Halftone visual cryptography through error diffusion. In: ICIP, pp. 109–112 (2006)
24. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Extended capabilities for visual cryptography. Theoretical Computer Science 250(1-2), 143–161 (2001); 102 Weir, J., Yan, W.
25. Nakajima, M., Yamaguchi, Y.: Extended visual cryptography for natural images. In: WSCG, pp. 303–310 (2002)
26. Fu, M.S., Au, O.C.: A novel method to embed watermark in different halftone images: data hiding by conjugate error diffusion (dhced). In: ICME 2003, Washington, DC, USA, pp. 609–612. IEEE Computer Society, Los Alamitos (2003)
27. Ulichney, R.A.: Digital Halftoning. MIT Press, Cambridge (1987)
28. Chen, Y.F., Chan, Y.K., Huang, C.C., Tsai, M.H., Chu, Y.P.: A multiple-level visual secret-sharing scheme without image size expansion. Information Sciences 177(21), 4696–4710 (2007)
29. Wang, D., Zhang, L., Ma, N., Li, X.: Two secret sharing schemes based on Boolean operations. Pattern Recognition 40(10), 2776–2785 (2007)
30. Cimato, S., De Santis, A., Ferrara, A.L., Masucci, B.: Ideal contrast visual cryptography schemes with reversing. Information Processing Letters 93(4), 199–206 (2005)
31. Viet, D.Q., Kurosawa, K.: Almost ideal contrast visual cryptography with reversing. In: Okamoto, T. (ed.) CT-RSA 2004. LNCS, vol. 2964, pp. 353–365. Springer, Heidelberg (2004)
32. Yang, C.-N., Wang, C.-C., Chen, T.-S.: Real perfect contrast visual secret sharing schemes with reversing. In: Zhou, J., Yung, M., Bao, F. (eds.) ACNS 2006. LNCS, vol. 3989, pp. 433–447. Springer, Heidelberg (2006)
33. Naor, M., Shamir, A.: Visual cryptography ii: Improving the contrast via the cover base. In: Crispo, B. (ed.) Security Protocols 1996. LNCS, vol. 1189, pp. 197–202. Springer, Heidelberg (1997)
34. Yang, C.N., Laih, C.S.: New colored visual secret sharing schemes. Designs, Codes and Cryptography 20(3), 325–336 (2000)
35. Fang, W.P.: Friendly progressive visual secret sharing. Pattern Recognition 41(4), 1410–1414 (2008)

36. Wu, C., Chen, L.: A study on visual cryptography. Master's Thesis, Institute of Computer and Information Science, National Chiao Tung University, Taiwan, R.O.C (1998)
37. Katoh, T., Imai, H.: An extended construction method for visual secret sharing schemes. IEICE Transactions J79-A(8), 1344–1351 (1996)
38. Yang, C.N., Chen, T.S.: Extended visual secret sharing schemes: Improving the shadow image quality. IJPRAI 21(5), 879–898 (2007)
39. Wu, H.C., Chang, C.C.: Sharing visual multi-secrets using circle shares. Computer Standards & Interfaces 28, 123–135 (2005)
40. Hsu, H.C., Chen, T.S., Lin, Y.H.: The ringed shadow image technology of visual cryptography by applying diverse rotating angles to hide the secret sharing. Networking, Sensing and Control 2, 996–1001 (2004)
41. Shyu, S.J., Huang, S.Y., Lee, Y.K., Wang, R.Z., Chen, K.: Sharing multiple secrets in visual cryptography. Pattern Recognition 40(12), 3633–3651 (2007)
42. Memon, N., Wong, P.W.: Protecting digital media content. Communications of the ACM 41(7), 35–43 (1998)
43. Luo, H., Pan, J.S., Lu, Z.M.: Hiding multiple watermarks in transparencies of visual cryptography. Intelligent Information Hiding and Multimedia Signal Processing 1, 303–306 (2007)
44. Hwang, R.J.: A digital image copyright protection scheme based on visual cryptography. Tamkang Journal of Science and Engineering 3(2), 97–106 (2000)
45. Hassan, M.A., Khalili, M.A.: Self watermarking based on visual cryptography. Proceedings of World Academy of Science, Engineering and Technology 8, 159–162 (2005)
46. Sleit, A., Abusitta, A.: A visual cryptography based watermark technology for individual and group images. Systemics, Cybernetics and Informatics 5(2), 24–32
47. Chuang, S.C., Huang, C.H., Wu, J.L.: Unseen visible watermarking. In: ICIP(3), pp. 261–264. IEEE, Los Alamitos (2007)
48. Hou, Y.C., Chen, P.M.: An asymmetric watermarking scheme based on visual cryptography. In: WCCC-ICSP 5th International Conference on Signal Processing Proceedings, vol. 2, pp. 992–995 (2000)
49. Praun, E., Hoppe, H., Webb, M., Finkelstein, A.: Real-time hatching. In: ACM SIGGRAPH 2001, pp. 579–584. ACM, New York (2001)
50. Yan, W.Q., Jin, D., Kankanhalli, M.S.: Visual cryptography for print and scan applications. In: Proceedings of International Symposium on Circuits and Systems, Vancouver, Canada, pp. 572–575 (May 2004)

A Novel Framework for Users' Accountability on Online Social Networks

Gambhir Mohit[1], Doja M.N.[2], and Moinuddin[3]

[1] Jamia Millia Islamia, New Delhi, India
mohitgambhir@gmail.com
[2] Dept. of Computer Engineering, Jamia Millia Islamia, New Delhi, India
[3] Delhi Technological University, New Delhi, India

Abstract. Social networking sites are a rage among public now a days. Individuals' keep themselves abreast of latest developments around them through these Online Social Networks (OSNs). With so many activities on OSNs, many a times users tend to reveal the information that may not be appeasing or morally acceptable by other users. Quite possible, though unknowingly, any user may enter into malpractices of spreading hatred among people by posting unethical and unacceptable material. Through this paper, the author has tried to resolves zooming issues of socially unacceptable postings by providing a new framework for controlling the user's actions on OSNs and thereby trying to minimize the menace of notorious activities.

Keywords: Social networks, security, privacy, spamming, phishing, malpractice, credibility.

1 Introduction

The face of technology is observing a frequent change in current scenario of globalization. Every new technology concept opens a lot many areas of research and have been a trust area of many researchers like Yager [1], Adar and Re [2], and Borgatti, Mehra, Brass and Labianca in [3]. The Online Social Networks (OSNs) open a new vista for individuals to interact with virtually anyone across the globe depending on certain conditions as set forth by the particular service provider. According to a report presented by Madden and Zickuhr [4], almost 65% of adults use online social networking sites in day-to-day routine; this clearly reflects how far these online social networking sites have reached in influencing our daily life.

The users of OSNs are provided with various facilities to share personal and other information such as date of birth, residential address, marital status, personal photographs, interests, hobbies, thoughts and much more. Such information as posted by the user is made viewable to other persons who may or may not be part of the OSNs. Interestingly, a user usually tends to share what he or she perceives from the surrounding world and develops thinking accordingly. Sometimes users post various types of information (on online social networks) according to their interest and mentality that may not be appeasing to the viewers of such post. Additionally, as per

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 831–841, 2013.

known human psychology described in McRaney [5], users seek information from outside world to confirm themselves before taking any decision. Quite realistic, mostly individuals rely on online resources to collect information before taking big decisions of their lives. Many big organizations and people from different domains and age groups such as politicians, business tycoons, technicians, educationists, teenagers and even children depend on online information.

Today's scenario brings variety of people to a common platform of social networking to gather information that can be useful for their personal, professional and social spheres of life. Such conglomeration of variety of people develops a social culture of a particular social networking website. Many social websites, such as Facebook, allow users to connect and interact globally for their social fellowships, business purposes, ecommerce, fun and many more purposes. The ultimate idea behind all such interactions is Information Exchange (IE), not just to serve livings personally, professionally but also to move along with the varying scenario of society and technology.

Let's consider a common scenario of the democratic government wherein politicians are elected and ruled based on choice of people. The politicians may need to know or change views of general public before planning to join any political party for their safe play in the contemporary political system. This can be done mainly through online social blogs or social networking sites that may provide information posted by various groups of people regarding their needs, choices and thought processes. Certainly, the social networking sites can be joined by any person or political party or oppositions to influence public by delivering more impactful messages for themselves and for others' as stated in Gorshkov[6].

Undoubtedly, the information posted by anybody when spread among public creates a common belief that turns out to be menacing, peacemaker or just informatory for people. Further, many people post their personal experiences in the form of videos or text messages that can influence many who seek practical insights of the related matter at hand. However, for any information seeker, finding genuine information and relying on the found information is a real challenge.

Typically, users tend to hide their actual personal information from public viewing or write fake information that poses wrong impression of such users on viewers of the profile and thus the viewers of such fake profile can take wrong decisions. For example, if an individual such as a marketer or a recruiter spends considerable time to find target users with a particular profile then such fake profiles (of the users) provide no guarantee that the individual will take correct decision in selecting target users. Thus the online available information is hard to be trusted and if breached once, then the level of trust is difficult to regain.

Many times, users post inflammatory material under the pretext of freedom of speech that can bring in legal action because of such post. This may not only lead to embarrassing social predicament for the user who is responsible for such post but also for the service provider of the social networking site who allowed such free content on the social website. A popular case held at Delhi court issued summons to various foreign-based social networking sites, including Facebook and Google, to face criminal charges for allegedly hosting objectionable contents and directed them to appear before the court. It asked the ministry of external affairs to get the summons served on these companies. The court direction came after the counsel, appearing for Facebook India, said over ten out of 21 companies named as accused in the case were foreign-based and that the court would have to issue process to serve the summons on them.

The court was hearing a private complaint filed by a journalist Vinay Rai against these firms for allegedly web-casting objectionable contents. The summons was issued to the sites including Facebook, Microsoft, Google, Yahoo and YouTube. "Let the process (to serve the summons) on (foreign-based) accused be sent through the MEA as per the process," Metropolitan Magistrate Sudesh Kumar said in [7].

An additional woeful plight is that many clean-handed users who are at no fault may also need to bear the consequences of judgment that may be laid by the court due to web-casting of obnoxious contents.

Further, due to ever increasing dependency on online information pool, reliability on web information, such as information available on many online social networking sites, is an extreme necessity. As per a report ***"less than 10% of India's population of 1.2 billion is online, according to the Internet and Mobile Association of India trade group. A Google executive said last year that the company expects India's Web user base to grow by 200 million people to reach at least 300 million by 2014, largely driven by increased Internet use on mobile phones"*** [8].

Author conducted a survey on 460 persons. Survey contains 20 questions ranging from general information of persons regarding social networking site viz. their frequency of usage of online social network sites, style of password setting, any compromise with security in existing environment to detect their interest for the new environment having more security and trust. Persons include from different professions, areas and age. The findings of survey is presented as below:

Q1. Have you made "User Category List" in your OSN site? If yes, please specify it?
A. 3% of the persons made their list according to the age group, 71% persons prepared the list according to friend relevancy such as family, classmate, college mate, neighbor etc. and 26% persons told about any other criteria but not specify that.

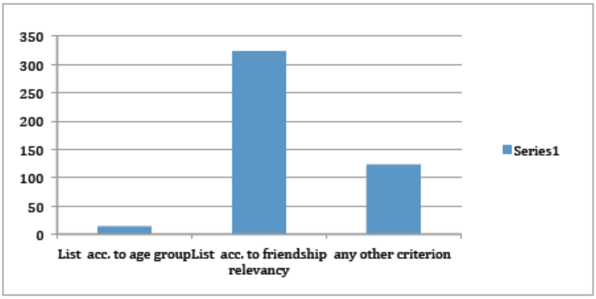


Fig. 1. Reflecting Response of Q1

Q2. How open you are to register (one time) for a central repository that will have control over whole web and will protect you from any repudiation or identification theft?

A. 10% of the persons said YES, 20% of persons said NO, 40% persons said SOMEWHAT and 30% persons said DONOT KNOW.

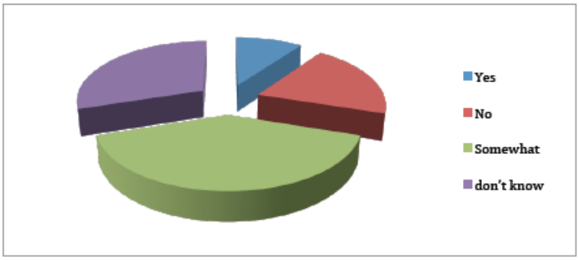


Fig. 2. Reflecting Response of Q2

Q3. Would you prefer a technology that enables more authentication and non-repudiation on websites?

A. 30% of the persons said YES, 37% of persons said NO, 33% persons said DON'T KNOW.

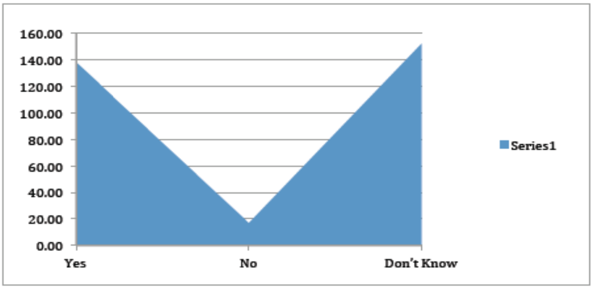


Fig. 3. Reflecting Response of Q3

Based on the aforementioned challenges, it is concluded that an efficient novel system needs to be provided to enhance trustworthiness of online information by safeguarding user interest and fundamental rights. Also, such novel system should be able to make a user, who posts information, accountable for his/her post by protecting the interest of the service provider (Online Social Networking sites) and other users of the networking site. Further, the users who post information should be trustworthy for viewers to believe on the information posted by such users. Thus, the novel system should efficiently be able to determine the trustworthiness of any user. To present such a system with aforementioned features and to enhance reliability of social networks, the author proposes a novel framework as explained below in subsequent sections

2 Proposed Framework of OSNs

An online social networking site is based on a structure that enables users to interact with each other by sharing their profile, pictures, and information with other users and also to post comments on other users' profile. Any social networking site can be defined as a hub for individuals to establish social relationships with each other. Each user of the social networking site makes social relations by starting with creating connections and making friends on the social networking site. Friends are considered as trusted users of the networking site who may share their personal information, professional information, interest and hobbies and can post other information related to real or virtual world. Not just friends (who are connected), other users (i.e., both registered and unregistered users) are also facilitated to spread or view any information through the social networking sites.

Social networking sites can facilitate users with wide range of tools for people to build a sense of community in an informal and voluntary way. Online users interact with each other, contribute information to the common information space, and participate in different interactive activities (e.g., photo uploading, tagging, etc.). Further, the Online Social Networks (OSNs) contain specific components that allow people to: define an online profile, list their connections (e.g., friends, colleagues), receive notifications on the activities of those connections, participate in group or community activities, and control permission, preference and privacy settings. Specifically, proposed framework corresponds to a novel system that can be implemented for online social networks.

The novel system facilitates each user of an online social network to bucket information (that the user is willing to post on the social network) into a suitable category. The suitable category is the category that relates majorly with the content of the information (that the user is willing to post) and intent of the user. Such facilities not only help users to organize information but also help information seekers to find suitable information in minimum span of time. Not only this, the current system enhances accountability on the user's part for the information posted by the user and thus, the user who posts information will be more careful about the content for posting on the online social network. Due to this, advantageously, the social network shall not be held liable for the reprobate acts (such as posting objurgatory information/remark) of any of the users of the social network. Additionally, the system can be an easy aid in legal field and can provide an insight for legal workers to reduce cyber crimes by following the user who is responsible for any illegal or anti-social activities on social sites.

The aforementioned and various other advantages of the proposed system are explained along with the implementation of the system in further sections of this paper. Further, the Online Social Networks (OSNs) contain specific components that allow people to: define an online profile, list their connections (e.g., friends, colleagues), receive notifications on the activities of those connections, participate in group or community activities, and control permission, preference and privacy settings. Specifically, proposed framework corresponds to a novel system that can be implemented for online social networks. The novel system facilitates each user of an online social network to bucket information (that the user is willing to post on the social network) into a suitable category. The suitable category is the category that relates majorly with the content of the information (that the user is willing to post) and intent of the user. Such facilities not only help users to organize information but also help information seekers to find suitable information in minimum span of time. Not only this, the current system enhances accountability on the user's part for the information posted by the user and thus, the user who posts information will be more careful about the content for posting on the online social network. Due to this, advantageously, the social network shall not be held liable for the reprobate acts (such as posting objurgatory information/remark) of any of the users of the social network. Additionally, the system can be an easy aid in legal field and can provide an insight for legal workers to reduce cyber crimes by following the user who is responsible for any illegal or anti-social activities on social sites. The aforementioned and various other advantages of the proposed system are explained along with the implementation of the system in further sections of this paper.

A unique architecture of the proposed system is depicted in Figure 4. The following layers have formed the entire system:

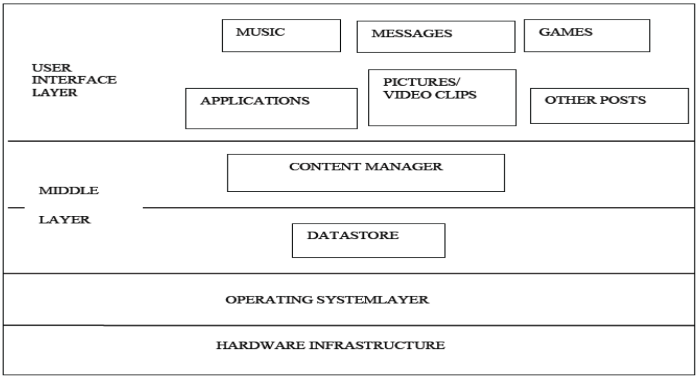


Fig. 4. Proposed Framework

- **A user interface layer:** This layer provides an interface to users for enabling users to interact with services offered by the social networking website. For example, the user interface layer facilitates users to post messages, pictures, videos, comments information and so forth. Further, users of the social networking sites can be facilitated to play music, interact with various applications, such as games. Typically, the user interface layer controls a display for enabling the user to interact with other users, post information (such as audio, video or text) and share information with other users of social networking website. The input provided by the users (of the social networking website) is managed/processed by a middle layer of the architecture.
- **Middle Layer:** Middle layer comprises mainly two components: content manager and data store. The content manager is for analyzing the content (information) when the user starts entering it in an input box that is made visible (on a display page) by the user interface layer. Based on the characters that the user types in an input box (provided to the user through the user

interface layer), one or more suitable categories can be determined and displayed as suggestions, to the user, so as to enable the user to select at least one category from the suggested categories.

Further, the suggestions for the suitable category can be provided based on intent of the user behind posting the information. The intent of the user can be determined based on the profile of the user and past record of activities of the user.

The data store is responsible for two main tasks: one task is for efficiently storing the information of social graphs and for handling increased database loads. Another main task includes storing information items of a social networking site. Data Stores can be Multimedia Databases, User Profiles Databases etc. The middle layer can also be referred to as data management layer.

- **Operating system layer** provides support system for implementation of functionality of other layers. Further, this layer provides an interface between users' activities and the hardware infrastructure.

Thereby, if the user selects wrong category (from the provided suggestions) for the information then an administrator of the system corrects the selected category and accordingly provide ranking (trust score) to the user. If the user posts any information in any wrong category or receives complaint about the posted information from any other user of the social network, then the administrator provides a low rank or negative trust score to the user. The trust score of the user provides an insight about the credibility of the user. Further, such insight about the credibility of the user helps other users to decide whether to add the user in their contact list or not. Also, the content posted by any user with low rank (trust score) may be considered as less credible for other users of the social network. Further research is going on to enhance the system's functionality so as to make the system more efficient in providing ranking or trust score to the user.

According to the proposed framework, it has been made mandatory for a user to login to any system using his global id such as being provided by OPENID. The OPENID will help in recognizing the user since it is the single id through which user can login to majority of the service providers like email, social networking sites and other online services. The OPENID is a concept by which the users don't have to remember multiple usernames and passwords since single username password combination provided access to various different sites. Further, a universal id concept also helps in recognizing the user in a much more credible way than that the conventional login systems as mentioned by Thibeau[9], Maler and Reed[10].

Once a user enters his or her login credentials as per the proposed framework, the credentials are matched to a central database so as to decide whether to grant access or not. Once the credentials are found OK, the user proceeds to access the service. The framework has been designed in a way to make user more accountable for his/her activities online. The user has to select an appropriate category for the message post that he/she is posting online. This concept is explained further in implementation section of the proposed system.

3 Implementation of the Proposed System

For any information (such as free information that a user may put on any social networking website), the user himself/herself has been made accountable (in the proposed system) by relieving the service provider (or the social networking website) from receiving blames due to anti-social or illegal posts of one or more users of the social networking site. The system corresponding to proposed framework is implemented by an online social network as shown in Figure 5.

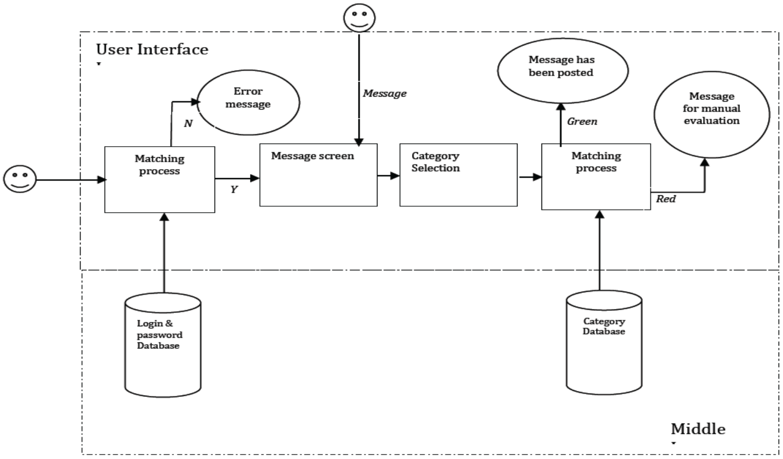


Fig. 5. Flow of Data in Proposed Framework

Assuming a user is registered with the online social network and possesses necessary details, such as user name and password, to login into the system to gain various services of the social network. The user name and the password of each user of the system can be stored, along with profile information of the user, in a database of the middle layer. Initially, any registered user needs to pass through an authentication stage. As shown, a user can provide authentication details that need to be matched with the details stored in the database for enabling the user to enter into the system of the social network. In case, the user is unable to authenticate himself/herself, the user cannot enter into the system to gain services corresponding to the social network and accordingly, an error message is displayed to the user.

On entering into the system, the user is provided with an input box on a display page wherein the user can enter message details that the user wants to post. When the user starts entering characters of the message in an input box (can also be referred to as ‘message box’) then a crawler might be utilized to index suitable information on the World Wide Web (WWW) and in the database of the system in order to determine suitable categories for the typed message (or any number of characters of the message that the user is willing to post). In this, the context of the message may be analyzed, semantically, to provide suggestions for categories. Specifically, the information indexed on the WWW and in the database can further be analyzed in light of the entered characters (related to the message), profile of the user, past record of the activities of the user, to determine suitable categories and accordingly suggestions for relevant categories can be provided to the user.

Additionally, further categories can be determined based on the intent of the user that are determined from the profile of the user and the past records of the activities of the user.

Further, the following categories, Table 1, are utilized by the system and stored by the database of the system. These categories are ever evolving based on the new content and information that a user posts or is willing to post.

The user selects one of the suggested categories that are provided as the suggestions to the user. For example, the user get many categories 'on the fly' while typing in the message box. Such categories allow the user to select one of the suggested categories. Additionally, the user may be provided with some other categories that may be less related to the content typed (by the user) in the message box. If the user selects any of the 'other categories' other than main suggested categories (of the system) for the message typed (in the message box) by the user then the message is analyzed by the system administrator to determine if the category suits to the context of the message.

Table 1. Various categories utilized by the System

Academic	Government	Sports
Adult Themes	Hate/Discrimination	Television
Adware	Health	Travel
Alcohol	Humor	Video Sharing
Auctions	Jobs/Employment	Visual Search Engines
Automotive	Movies	Tobacco
Business Services	Music	Weapons
Dating	News/Media	Religious
Drugs	Non-profits	Research/Reference
Ecommerce/Shopping	Nudity	Sexuality
Educational Institutions	Politics	Software/Technology
Games	Pornography	Email

Further, it is determined if the user has chosen a correct category, by matching the selected category from the suggested categories. If the selected category for the message is not matched with the suggested categories, then the message undergoes a manual evaluation. In one case, an administrator provides a trust score that may define the credibility of the user. Otherwise, the administrator may give a negative score for the user if the user has categorized his/her message wrongly by selecting a wrong category for his/her message. This trust score makes the user accountable for his/her postings and thus restricts him/her from posting any illicit or vulgar post that may be banned from public distribution.

4 Conclusion

Based on the above solution as provided by the proposed system, a user may be able to post the content (information) in an appropriate category by selected the appropriate category from the suggested list of categories. Thus, the user himself or herself is not required to spent time in thinking and deciding much about creating a suitable category for the information that the user is willing to post. Also, due to categorized information, the information seekers can easily select the target category to find any required information.

For example, due to such facility of categorization a student can easily find suitable categories to seek relevant information without requiring any filtering work from all the posts of different categories. Further, based on such categorization, users of a particular age group may be barred from entering into a specific group. For example, students below the age of 18 may be barred from accessing the category related to adults. Also, due to strict categorization of the information based on various factors, such as analysis of the content (that the user types in the provided message box/input box), profile of the user and the past records of activities related to the users (as mentioned above), the user is held liable if the information is posted explicitly in any unsuitable category which is not suggested. Accordingly, the system administrator may rate the user negatively and provide a negative trust score that may be viewable by the other users. Due to this, any user and his/her content can gain or lose credibility in the eyes of other users based on selection of correct or incorrect category by the user. Moreover, in case of any post by an anti-social element, the service provider may not be held liable and can easily catch hold of the responsible user for the related anti-social activity on the social network. This further, provides an insight for legal workers to keep an eye on anti-social elements from the social network websites.

It may be appreciated that the applications and advantages are not limited to the above explanation and many more applications may be understood while implementing the proposed system. Further, the research work is still continued to make the system further more efficient.

References

1. Yager, R.R.: Granular Computing for Intelligent Social Network Modeling and Cooperative Decisions. In: International IEEE Conference: Intelligent Systems (2008)
2. Adar, E., Re, C.: Managing Uncertainty in Social Networks. Data Engineering Bulletin 30, 23–31 (2007)
3. Borgatti, S.P., Mehra, A., Brass, J.D., Labianca, G.: Network Analysis in the Social Sciences. Science 323(5916), 892–895 (2009)
4. Madden, M., Zickuhr, K.: 65% of Online Adults use Social Networking Sites. In: a project of PewResearchCenter (August 26, 2011), http://pewinternet.org/Reports/2011/Social-Networking-Sites.aspx
5. McRaney, D.: You Are Not So Smart - Book Trailer – Procrastination, http://www.youtube.com/watch?feature=player_embedded&v=DJ2T4-rUUcs

6. Gorshkov, I.: The Internet and Social Information Networks in Contemporary Politics. In: International Scientific Conference Networks in the Global World: Structural Transformations in Europe, June 22-24, St. Petersburg, US (2012), http://www.ngw.spbu.ru/node/283
7. Foreign Social Networking Sites Summoned by Court. In: Hindustan Times, New Delhi (January 13, 2012), http://www.hindustantimes.com/India-news/NewDelhi/Foreign-social-networking-sites-summoned-by-court/Article1-796494.aspx
8. Sharma, A.: Facebook, Google to Stand Trial in India. The Wall Street Journal, Technology (March 13, 2012), http://online.wsj.com/article/SB10001424052702304537904577277263704300998.html
9. Thibeau, D.: Open Trust Frameworks for Open Government: Enabling Citizen Involvement through Open Identity Technologies (August 2011), http://openid.net/government/
10. Maler, E., Reed, D.: The Venn of Identity: Options and Issues in Federated Identity Management. IEEE Security and Privacy 6, 16–23 (2008), http://doi.ieeecomputersociety.org/10.1109/MSP.2008.50

Network Security Using ECC with Biometric

Dindayal Mahto and Dilip Kumar Yadav

Department of Computer Applications
National Institute of Technology, Jamshedpur
Jharkhand, PIN- 831014, India
{dindayal.mahto,dkyadav1}@gmail.com

Abstract. The popular asymmetric cryptography is RSA but most of the RSA–based hardware and software products and standards require big cryptographic keys length for higher security level. The existing asymmetric cryptography algorithms need the storage of the secret keys. Stored keys are often protected by poorly selected user passwords that can either be guessed or obtained through brute force attacks. This is a major weakness of the crypto-system. Combining biometrics with cryptography is seen as a possible solution. This paper discusses the network security using Elliptic Curve Cryptography with contactless palm vein biometric system. It provides more security with less key length and also there is no need to store any private key anywhere. It focuses to create and share secret key without transmitting any private key so that no one could access the secret key except themselves.

Keywords: Elliptical Curve Cryptography (ECC), Biometric, Palm Vein, MD5, Rivest Shamir Adleman (RSA).

1 Introduction

We are living in cyber age, where most of the information is produced with the help of computers and computer networks, which provides platform to do e-commerce tasks, online banking, and sharing of information and many more, and while more than two parties communicate to each other then they worry about confidentiality, data integrity, non-repudiation and privacy etc. [1]. In order to mitigate these issues, we can apply cryptography with biometrics. Cryptography is a kind of secret writing by which two parties can communicate with secret messages [2]. Most of the researches have demonstrated that biometric is the ultimate solution for identification and authentication, since it is proved as reliable and universally acceptable identification/authentication methods in many application areas [3].

Due to the popularity of biometrics and cryptography, the information security is becoming as a common demand in all applications area. Biometric is referred as automatic system that uses measurable, physical or physiological characteristics or behavioral traits to recognize the identity of an individual. Biometrics offers greater security in identification/ authentication system. However, the security level of the network can be further enhanced using cryptography and biometrics.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 842–853, 2013.

To secure the communication currently there are two popular kinds of cryptographic protocol namely symmetric key and public key protocol. In symmetric key protocol such as Data Encryption Standard (DES), International Data Encryption Algorithm (IDEA) and Advanced Encryption Standard (AES) [2], a common key is used by both sender and receiver for encryption and decryption. This system provides high speed but have the drawback that a common key must be established for each pair of participants. In public key protocol there are two keys, public key and private key by which message can be encrypted and decrypted. One is kept private by owner and used for decryption. The other key is published to be used for encryption. Some of the most useful example of the public key cryptography is RSA, ElGamal and Digital Signature Algorithm (DSA) [4]. Although, these algorithms of asymmetric crypto-systems are slower than that of the symmetric crypto-systems but they provide high level security. Due to comparative slowness of the public key cryptography algorithms, dedicated hardware support is desirable. RSA is used in most of the network and standards that uses public key cryptography for encryption, decryption and digital signature. The length of the keys for RSA has been increased in recent years, and this is putting a heavier load on the application of RSA. It creates extra computation cost and processing overhead. However, ECC compared to RSA, offers higher security per bit with smaller key size. It provides higher security per bit. Since ECC has smaller key size, hence it also reduced the computation power, memory and bandwidth.

Therefore, in this paper a model has been proposed for network security using ECC with biometric. In this model, keys are generated from palm vein biometric which are used for encryption and decryption in the ECC for identification and authentication.

This paper is organized as follows. In section 2, we provide the review of the elliptic curve cryptography, why we use elliptic curve cryptography instead of RSA or other cryptography system, the implementation method of ECC and its mathematical operation and method for finding all points on the elliptic curve on which we have to encrypt the message. We describe the Elliptic Curve Diffie-Hellman Algorithm (ECDH) in this section for generating key. In section 3 we describe about the biometric and importance of palm over the biometric, why we use palm instead of iris, finger, face, retina or other biometric. In section 5 we describe how can we encrypt and decrypt the message by the help of palm as a private key. In section 6 we describe the result and discussion. We conclude the paper in section 7.

2 Elliptic Curve Cryptosystem

In 1985, Neil Koblitz [4] and Victor S. Miller [5] independently proposed the use of elliptic curve cryptography. Since 1985, there have been a lot of studies concerning elliptic curve cryptography. The use of ECC is very inviting for various reasons [1, 3, 6, 7]. The first and probably most important reason is that ECC offers better security with a shorter key length than any other public-key cryptography. For example, the level of security achieved with ECC using a 160-bit key is equivalent to conventional public key cryptography (e.g. RSA) using a 1024-bit key [4]. There are huge importances of shorter key lengths especially in applications having limited memory resources because shorter key length requires less memory for key storage purpose. Elliptic curve cryptosystems also require less hardware resources than conventional public-key cryptography. Now at the security level ECC is more secure than RSA. RSA can be cracked successfully, uses 512 bits and for ECC the number of bits is 97, respectively. It has been analyzed that the computation power required for cracking ECC is approximately twice the power required for cracking RSA. ECC provides higher level of security due to its complex mathematical operation. Mathematics used for ECC is considerably more difficult and deeper than mathematics used for conventional cryptography. In fact this is the main reason, why elliptic curves are so good for cryptographic purposes, but it also means that in order to implement ECC more understanding of mathematics is required. A short introduction to mathematics behind elliptic curve cryptosystems is given in this paper; however, this paper should give a good overall picture of ECC and its implementation issues.

2.1 Mathematics Behind ECC

Cryptographer noticed that elliptic curves behaved conveniently when operations were performed with prime modulus. That means cryptographer elliptic curve is in the form $y^2 \bmod p = (x^3 + ax + b) \bmod p$ where $4a^3 + 27b^2 \neq 0$ and p is a prime number and a, b is the parameter of the curve, here variables and coefficient are all restricted to elements of a finite field. There are two families of elliptic curve are used in cryptography application [8, 9, 10].

1. Elliptic Curves over $GF(2^m)$
2. Elliptic Curves over Z_p.

In Elliptic curves defined over $GF(2^m)$, the variables and co-efficient all take on values in $GF(2^m)$ and in calculation performed over $GF(2^m)$.

In Elliptic Curves over Z_p, we use a cubic equation in which the variables and co-efficient all take on values in the set of integers from 0 through (p-1) and in which calculations are performed modulo p.

This paper is based on the Elliptic curves over Z_p.

For example, let us take our elliptic curve is

$$y^2 \bmod 11 = (x^3 + ax + 2) \bmod 11 \tag{1}$$

2.2 Arithmetic Operation in ECC

The rule of mathematical operation on elliptic curve is different from the rule conventional mathematical operations. If we want to add two points of elliptic curve then we have to follow given below rule. For this and all arithmetic operation there are some rules which are as follows [8, 9,10].

The rules for addition over $E_p(a, b)$. For all points P, Q Є $E_p(a, b)$:

Rule 1: P + O (Infinity) = P

Rule 2: If P = (x_1, y_1), then P + $(x_1, -y_1)$ = O.

Rule 3: *If P = (x_1, y_1) and Q = (x_2, y_2) with $P \neq -Q$,* then R = P + Q = (x_3, y_3) is determined by the following rules:

$$x_3 = (\lambda^2 - x_1 - x_2) \bmod p \tag{2}$$

$$y_3 = (\lambda(x_1 - x_3) - y_1) \bmod p \tag{3}$$

where,

$\lambda = ((y_2 - y_1) / (x_2 - x_1)) \bmod p$, if $P \neq Q$

and,

$\lambda = ((3x_1^2 + a) / 2y_1) \bmod p$, if $P = Q$

Rule of Multiplication: It is defined as repeated addition.

Suppose P is a point on elliptic curve P = (x_1, y_1)

Thus 8*P =P+P+P+P+P+P+P+P

=2P+2P+2P+2P

=4P+4P

2.3 Points on ECC

For any operation on elliptic curve, first of all we have to find the all point of that curve [10]. Thus for finding the point on the curve firstly we have to chose any elliptic curve. Suppose $y^2 \bmod p = (x^3 + ax + b) \bmod p$ is an elliptic curve where $4a^3 + 27b^2 \neq 0$. Then points on this curve are the set $E_p(a, b)$ consisting of all pairs of integers (x, y), which satisfy the above equation together with the point Zero. Method for finding the points on the curve is as follows:

Points on ECC

Step1. Determine the L.H.S of elliptic curve for all (x, y) € Z_p.

Step2. Determine the R.H.S of elliptic curve for all x, y € Z_p.

Step3. Choose the Pair of corresponding value of x and y as a pair for all x, y € Z_p for which L.H.S. = R.H.S.

Step4. All pairs of such (x, y) are the point on the curve.

Example

If in above curve, value of p=11, a=1, b=1, then points on the elliptic curve are (0,1),(2,0),(3,3),(3,8),(4,5) etc.

2.4 ECDH (Elliptic Curve Diffie-Hellman Algorithm)

Elliptic curve Diffie-Hellman algorithm is the Diffie-Hellman algorithm for the elliptic curve [3, 8]. The original Diffie-Hellman algorithm is based on the multiplicative group modulo *p*, while the elliptic curve Diffie-Hellman (ECDH) protocol is based on the additive elliptic curve group. We assume that the underlying field *GF* (*p*) is selected and the curve *E* with parameters *a*, *b*, and the base point *P* is set up. The order of the base point *P* is equal to *n*. The standards often suggest that we select an elliptic curve with prime order, and therefore, any element of the group would be selected and their order will be the prime number *n*. At the end of the protocol the communicating parties end up with the same value *K* which is a point on the curve. A part of this value can be used as a secret key to a secret-key encryption algorithm.

Suppose there are two users Alice and Bob. According to the Diffie-Hellman the key generation and exchange is as follows.

Key generation and key exchange

Step 1: Alice uses his palm vein feature for his private key d_A which less than n.

Step 2: Alice generates a public key $P_A = d_A * G$; the public key is a point in $E_p(a, b)$.

Step 3: Bob similarly uses his palm vein features for his private key d_B which is less than n.

Step 4: Bob computes a public key $P_B = d_B * G$.

Step 5: Alice generates the secret key

$k = d_A * \quad P_B$.

Step 6: Bob generates the secret key

$k = d_{\mathbf{B}} * \quad P_A$.

By exchanging the key through this method both Bob and Alice can communicate safely. Bob can use the secret value he computed to build an encrypting key. When Alice gets the message from Bob, she uses the secret value she computed to build the decrypting key. It is the same secret value, so they use the same key. Thus what Bob encrypts Alice can decrypt.

3 Why PALM?

Palm vein authentication uses the vascular patterns of an individual's palm as personal identification data. Compared with a finger [18] or the back of a hand, a palm has a broader and more complicated vascular pattern and thus contains a wealth of differentiating features for personal identification. The palm is an ideal part of the body for this technology; it normally does not have hair which can be an obstacle for photographing the blood vessel pattern, and it is less susceptible to a change in skin color, unlike a finger or the back of a hand. The deoxidized hemoglobin in the vein vessels absorb light having a wavelength of about 7.6 x 10.4 mm within the near-infrared area [19]. When the infrared ray image is captured, unlike the image seen in Fig.1, only the blood vessel pattern containing the deoxidized hemoglobin is visible as a series of dark lines (Figure 5). Based on this feature, the vein authentication device translates the black lines of the infrared ray image as the blood vessel pattern of the palm (Figure 6). The palm vein sensor (Figure 7) captures an infrared ray image of the user's palm. The lighting of the infrared ray is controlled depending on the illumination around the sensor, and the sensor is able to capture the palm image regardless of the position and movement of the palm.[20]

Palm vein offers contactless authentication and provides a hygienic and noninvasive solution, thus promoting a high-level of user acceptance. Fujitsu believes that a vein print is extremely difficult to forge and therefore contributes to a high level of security, because the technology measures hemoglobin flow through veins internal to the body.[20]

4 Previous Works

The main problem of asymmetric cryptography is the management of private key. No one should be able to access someone else's private key. They need to store in such a place which is protected from unauthorized accessing. This is vulnerable for attacking by hackers. This creates big problem in asymmetric cryptography. Thus it can be solved by the use of biometric template. Private Key can be generated directly by the biometric template. Since private key can be generated dynamically from one's biometric template, so there is no need to store private key anymore and network becomes more secure and safe. But there are very little work has been done in the field of ECC with the help of biometric. Some of the suggested approaches are given. [1], [22], [23]. However these biometrics have lots of issues regarding training, capturing image, easily obscured by eyelashes, eyelids, lens and reflections from the cornea, lack of existing data deters ability, cost, voice can be captured while uttering the password, a camera can photograph an iris from across the room, and fingerprints left on surfaces can be lifted hours later [23] etc. For some individuals, the iris image capturing is very difficult. Iris recognition system requires lots of memory to be stored. It is easily absurd by eyelash, eyelids, lens and reflection from the cornea. People are not much familiar with iris recognition system yet, so there are lots of myths and fears related to scanning the eye with light source. Iris recognition system works on the basis of acquisition of iris image, but acquisition of an iris image needs more training and attractiveness than most other biometrics. It cannot be verified by human too. The most problem with iris recognition system is its expensiveness. Lifang et al.[22] have generated cryptographic key from user's face features and then the key has been applied in DES algorithm for encryption and decryption purposes and same way Fabian et al.[23] have generated cryptographic key from user's voice while speaking a password, but no further implementation of key has been described on their paper.

As palm vein print is extremely difficult to forge and therefore contributes to a high level of security, because the technology measures hemoglobin flow through veins internal to the body. We are generating cryptography keys from user's palm vein and then the generated key are used as user's secret keys for ECC.

Hence in proposed method we are using palm vein as a secret key instead of other biometric.

5 Proposed Work

In this paper we are using palm vein features of senders' and receivers' for generating secret keys, then the keys are used in Elliptic Curve Cryptography to provide network security while sending the information from sender to receiver and vice versa.

5.1 Method for Generating Public Key and Private Key

First of all users' palm features are scanned through Palm Vein scanner and then same are filtered for registrations purpose known as enrollment and later palm features are used for authentication.

Fig. 1. ATM with palm vein pattern authentication sensor unit

Fig. 2. Palm vein access control unit

Fig. 3. ATM for convenience stores with downsized palm vein pattern sensor unit

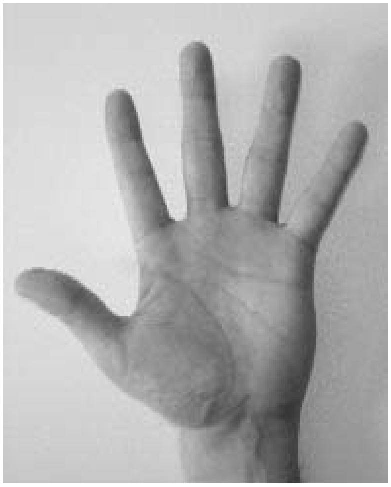


Fig. 4. Visible ray image

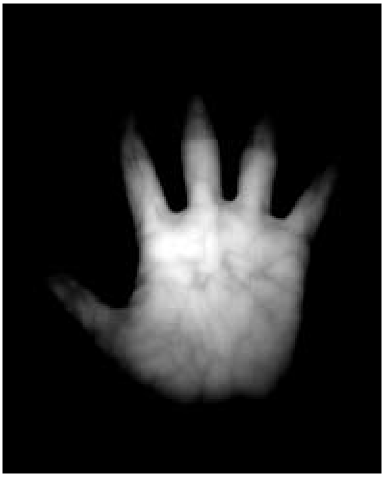


Fig. 5. Infrared ray image

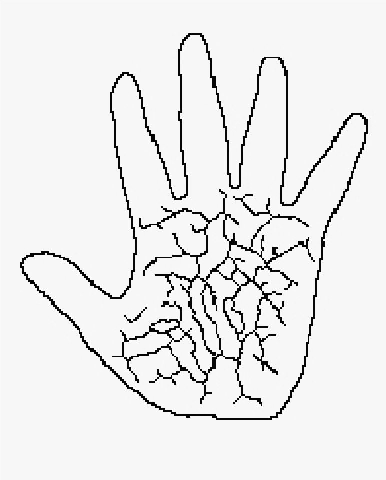


Fig. 6. Extracted vein pattern

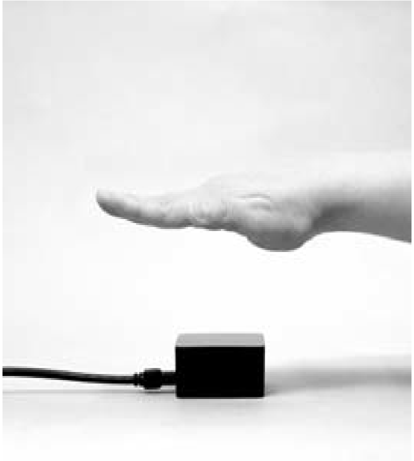
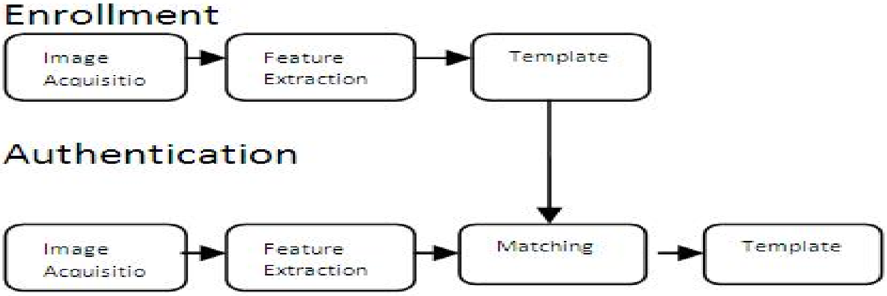


Fig. 7. palm vein sensor

Palm vein authentication process consists of two essential procedures: enrolment and authentication. Taking the following steps completes each procedure:



To generate private key, we take the palm vein of the user and generate its hash value by the help of MD5 cryptographic hash function [9]. This resultant hash value is the private key of the user. Suppose this value is d_A for use Alice and d_B for user Bob.

Now to generate public key in elliptic curve cryptosystem by the help with this private key is as follows:-

Step1: Both user choose the same large prime ‘p’ and the elliptic curve parameter ‘a’ and ‘b’ such that

$$y^2 \bmod p = (x^3 + ax + b) \bmod p;$$
$$\text{where, } 4a^3 + 27b_2 \neq 0$$

Step 2: Now choose any one point G(x, y) from this elliptic curve. This point is called the base point of the curve.

Step3: Compute $P_A = d_A * G(x, y)$

This P_A is called the public key of user Alice.

To generate public key of user Bob same operation can be performed by the help with private key of user Bob.

5.2 Message Encryption

Suppose user Alice wants to send a message to user Bob, then first task in this system is to encode the plaintext message m to be sent as a point P_m (x, y). It is the point P_m that will be encrypted as a cipher text and subsequently decrypted. After mapping of points [17] with user message characters on elliptic curve, they can encrypt the message by following steps

Step 1: Suppose Alice encodes the message m as $P_m = (x, y)$

Step 2: Alice takes his private key from his palm vein feature suppose it is k and produces C_m consisting of the pair of points:

.

$$C_m = \{k * G, P_m + k * P_B\}$$

Here C_m is a cipher text, Alice sends this cipher text to Bob.

5.3 Message Decryption

For Message decryption Bob has to do following procedures.

Step 1: Bob multiplies the first point in the pair by his secret key and subtracts the result from the second point:

$$P_m + k * P_B - d_B * k * G$$
$$= P_m + k(d_B G) - d_B(k * G)$$
$$=Pm$$

Step 2: The message P_m is the required message of Bob, which is sent by Alice.

6 Result and Discussion

Traditional methods for implementing public key infrastructure, encryption and decryption techniques face lots of problem such as key management, key storing, key privacy etc. Our proposed approach can handle such problems. Here we are using palm vein features as a private key so that there is no need to store any private key and also palm vein has lots of merits over other biometrics (i.e., it is most user friendly and cheaper too). Palm vein recognition also has some outstanding features like universality, permanence, uniqueness and accuracy. As we are using ECC, so we can achieve high level security with very shorter key size. Thus it also solves the key size problem. ECC requires very complex mathematical operation (because of elliptic curve Diffie-Hellman problem, which is harder than discrete logarithmic problem) therefore security strength per bit is also very high.

We have implemented ECC portion of this proposed work in MATLAB R2008a under Microsoft Windows platform. It asks all related parameters and then generates cipher-text in one graph and decrypted-text in another graph.

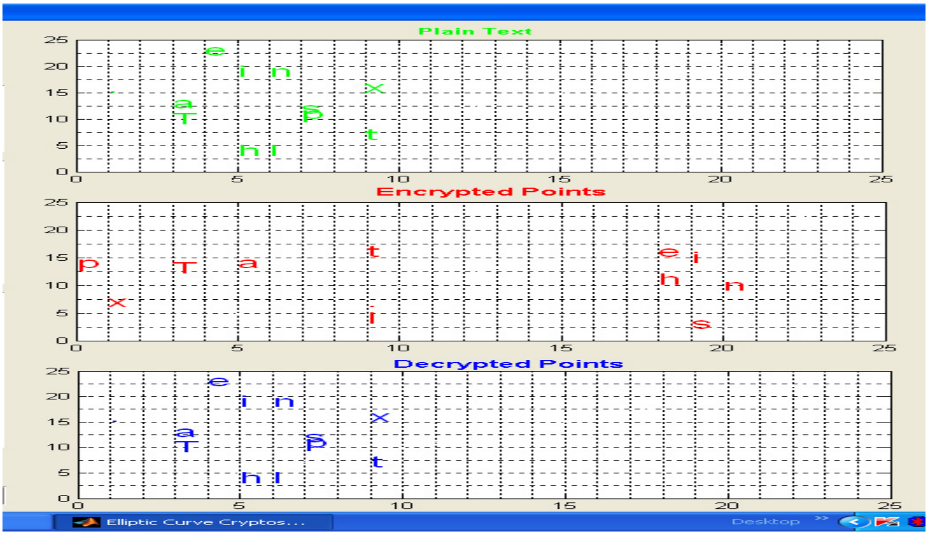


Fig. 8. Elliptic Curve Cryptography Software Implemented in MATLAB (r2008a)

7 Conclusions

In this paper, network communication becomes very secure with the help of ECC and palm vein biometric. The main advantage of ECC is that it requires very less key size and gives high level of security with cheapest biometric recognition system and there is no need to store any private key anywhere. Palm vein authentication technology offers contactless authentication and provides a hygienic and non-invasive solution, thus promoting a high-level of user acceptance. A vein print is extremely difficult to forge and therefore contributes to a high level of security, because the technology measures hemoglobin flow through veins internal to the body. Thus the proposed model provides a very secure network communication system.

References

1. Mohammadi, S., Abedi, S.: ECC based Biometric Signature: A new approach in electronic banking security. In: International Symposium on Electronic Commerce and Security (ISECS 2007), pp. 763–766 (2008), doi:10.1109/ISECS.2008.98
2. Stallings, W.: Cryptography and Network Security Principles and Practices, Edition Fourth. Pearson Prentice Hall (2007)
3. Nandini, C., Shylaja, B.: Efficient Cryptographic key Generation from Fingerprint using Symmetric Hash Functions. International Journal of Research and Reviews in Computer Science (IJRRCS) 2(4) (August 2011)
4. Mel, H.X., Baker, D.: Cryptography Decrypted. Addision-Wesley (2011)
5. Koblitz, N.: Elliptic Curve Cryptosystem. Mathematics of Computation (48), 203–209 (1987)
6. Miller, V.S.: Uses of Elliptic Curve in Cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986)
7. Prasanna Ganesan, S.: An Asymmetric Authentication Protocol for Mobile Devices Using Elliptic Curve Cryptogphy. In: ICACC, pp. 107–109.
8. Zhou, X.: Elliptic Curves Cryptosystem Based Electronic Cash Scheme with Parameter Optimization. In: Pacific-Asian Conference on Knowledge Engineering and Software Engineering (KESE 2009), pp. 182–185 (2009), doi:10.1109/KESE.2009.55.
9. Kumar, M.: Cryptography and Network Security, Krishna Prakashan Media (P) Ltd. 2nd edn. (2007)
10. Anoop, M.S.: Elliptic Curve Cryptography, An implementation tutorial, Tata Elexsi Ltd., Thiruvananthapuram, India
11. Doshe, C., Lange, T.: Arithmetic of Elliptic Curves. In: Cohen, H., Frey, G. (eds.) Handbook of Elliptic and Hyper Elliptic Curve Cryptography, ch. 13. Chapman and Hall/CRC, Taylor and Francis Group (2006)
12. Ahmad Jhat, Z., Hussain Mir, A., Rubab, S.: Palm Texture Feature for Discrimination and Personal Verification. In: Third international Conf. on Emerging Security, System and Technologies (SECURWARE 2009), pp. 230–235 (2009), doi:10.1109/SECURWARE. 2009.42
13. Udb –Din, H., Al-Jaber, A.: Securing online shoping using biometric personal authentication and stagenography. In: ICTTA 2006, pp. 233–238 (2006)
14. Woodward, J.D., Orlans Jr., N.M., Higgins, P.T.: Biometrics The ultimate reference. Dreamtech Press (2003)
15. NSTC on Biometrics, http://www.questBiometrics.com
16. Biometric-Comparison, http://biometric.pbworks.com/w/page14811349/advantagedisadvantage
17. Nanawati, S., Thieme, M., Nanavati, R.: Biometrcs Identity Verification in a networked world, 1st edn. Willey Computer Publishing (2002)
18. Rao, O.S.: Efficient mapping method for elliptic curve cryptosystems. International Journal of Engineering Science and Technology 2(8), 3651–3656 (2010)

19. Miura, N., Nagasaka, A., Miyatake, T.: Extraction of Finger-Vein Patterns Using Maximum Curvature Points in Image Profiles. In: Proceedings of the 9th IAPR Conf. on Machine Vision Applications (MVA 2005), Tsukuba Science City, Japan, pp. 347–350 (2005)
20. Bio-informatics Visualization Technology committee, Bio-informatics Visualization Technology, p. 83. Corona Publishing (1997)
21. Watanabe, M., Endoh, T., Shiohara, M., Sasaki, S.: Palm vein authentication technology and its applications. Fujitsu Laboratories Ltd., 1-1, Kamikodanaka 4- chome, Nakahara-ku, Kawasaki, 211-8588, Japan
22. Zhanga, P., Hub, J., Lic, C., Bennamound, M., Bhagavatula, V.: A pitfall in fingerprint bio-cryptographic key generation. Computers & Security 2(4) (August 2011)
23. Wu, L., Liu, X., Yuan, S., Xiao, P.: A Novel Key Generation Cryptosystem Based on Face Features. In: Precedings of the ICSP 2010. IEEE (2010)
24. Monrose, F., Reiter, M.K., Li, Q., Wetzel, S.: Cryptographic Key Generation from Voice. In: Proceedings of the 2001 IEEE Symposium on Security and Privacy (May 2001)

Hybrid Key Management Technique for WSN's

Ravi Kishore Kodali and Sushant Chougule

Department of Electronics and Communications Engineering
National Institute of Technology, Warangal
Warangal, 506004 Andhra Pradesh, India

Abstract. Wireless sensor networks are envisaged in military, commercial and healthcare applications, where data security is an important aspect. Security of the data in the network is based on the cryptographic technique and the way in which encryption and decryption keys are established among the nodes. Managing the keys in the network includes node authentication, key agreement and key update phases which poses an additional overhead on network resources. Both Symmetric and Asymmetric key techniques when applied separately in WSN fails to provide a scheme suitable for wide range of applications. Hybrid key management scheme is scalable alternative to match security requirements of WSN with minimum overhead on available resources. Heterogeneous WSN is considered in which ID based key establishment and polynomial based key pre-distribution scheme are proposed for higher and lower level of hierarchy respectively. The results of the proposed hybrid key management scheme indicate reduced resource overhead and improved security level.

Keywords: WSN, Elliptic curves, symmetric key pre-distribution, IBC.

1 Introduction

Wireless sensor networks (WSN's) are being used in wide range of military and commercial applications. A WSN consists of tiny resource constrained sensor nodes and special monitoring device termed as base station. Sensor nodes act as the skin, which collect the data from surrounding environment and forward to the base station, the brain of the network, controls the data flow. WSN technology is expected to play crucial role in near future as the means of global data communication. 'Internet of things'[1] idea, proposed recently, considers WSN as the basic element to gather data. The collected data is then made globally available by connecting many small WSNs to the Internet. In such scenario, information collected by WSN's would have transcended value compared to its previous small scale application. Consequently, security of the data also plays a vital role.

Confidentiality, Integrity and Authenticity of the data collected are the main issues in sensor network security. Wireless nature of the network along with the lack of computational ability of sensor nodes poses many challenges in the implementation of security protocol for WSN. RSA-1024 and AES cryptographic

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 854–865, 2013.

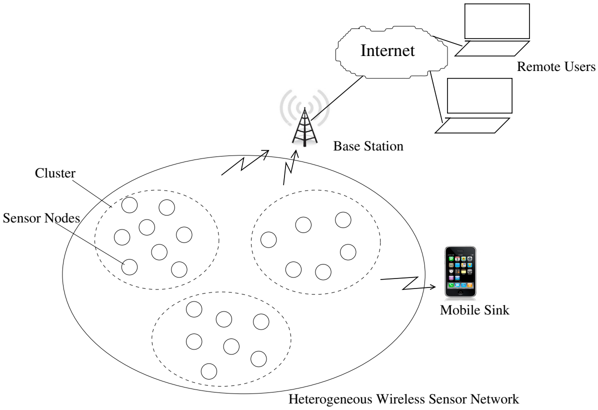


Fig. 1. Wireless Sensor network scenario

standards are widely used in secure internet transactions[2]. But computational cost of these schemes is not feasible over sensor nodes. Elliptic curve cryptography is proved to be the most suitable asymmetric key technique for WSN because of small key size$(160 - bit)$ with equivalent security[3]. Effective applicability of these cryptographic schemes in the network depends on the key management technique used.

Key Management technique is the backbone of any network security scheme. Secured channel for data transmission in a WSN is provided by key establishment protocol. The design of key management protocol mainly focusses on the consumption of resources like memory, energy and processing time by the scheme, resilience against various attacks, communication overhead and scalability. At the system level, the demands from key management technique being resilience against node capture, forward secrecy, backward secrecy, node revocation on intrusion detection and security against network level attacks. Both symmetric and asymmetric key techniques used for computer networks fail to satisfy these WSN specific security requirements.

In this paper, a hybrid key management technique is proposed for heterogeneous wireless sensor network. First, clusters are formed based on the location of WSN nodes. A suitable cluster head selection algorithm can be used to elect and update the cluster head. A cluster head is assumed to have the same hardware resources as those of other nodes. The base station communicates with a cluster head and establishes secure connection using Identity Based Cryptography (IBC). All the cluster heads are securely connected using IBC, which is more secure due to the elliptic curve discrete logarithmic problem $(ECDLP)$. Nodes inside a cluster use polynomial based pairwise key pre-distribution scheme and avoid extra computational burden. Section II discusses the related work in the field of WSN key management. Section III provides mathematical background for IBC and Section IV presents the proposed hybrid key management scheme.

2 Related Work

Key management techniques proposed for WSN attempt to seek perfect connectivity and resilience against node capture attack[4]. This means that each node should be able to communicate with every other node in the network and if a node gets captured, secured connections of other nodes should remain intact. The simplest way to establish key management in a WSN is to make use of single master key for the entire network. It provides full connectivity and scalability, but a single node compromise can expose the whole network. To circumvent this problem, pairwise keys can be pre-distributed in each node so that capturing of a node will affect only single node keeping all the other connections secure. But for a WSN, consisting N nodes, each node needs to store (N-1) pairwise keys to achieve full connectivity. Apart from stringent memory requirements, this technique limits the scalability of the WSN.

To provide a trade-off between connectivity and resilience against node capture attack, Eschenauer and Gligor[4] first proposed probabilistic key pre-distribution scheme. In this scheme, prior to deployment, a subset of the keys from a large key pool is stored in each node. Each key is tagged with unique identifier. Nodes broadcast key identifiers to their neighbours and pairwise key with the nodes having at least one common key. Nodes that are unable to establish a direct pairwise key, enters into secure path discovery phase. The node capture attack affects non-captured nodes as captured node contains common keys with a given probability.

Improvement to this scheme is Q-composite random key distribution[5], which requires nodes to contain at least Q common keys to establish pairwise key. This technique reduces the probability of compromising secured link between non captured nodes by the factor Q. Another improved random key pre-distribution uses hash function H[6]. For node i key from key pool is hashed $(i-1)$ times. For establishing pairwise key between nodes A and B having keys $K_A = H^{i_a}\left(K_i\right)$ and $K_B = H^{i_b}\left(K_i\right)$ respectively shares i_a and i_b value. If $i_a < i_b$ node B can easily calculate symmetric key as

$$K_{AB} = H^{i_a - i_b}\left(K_B\right) \tag{1}$$

Polynomial based pair wise key distribution scheme[7] provides more resilience against node capture attack with less memory requirement. Polynomial $p\left(x,y\right)$ of degree t and having coefficients over $GF\left(q\right)$ is used to establish keys between the nodes. The polynomial has the property $p\left(x,y\right) = p\left(y,x\right)$

$$p\left(x,y\right) = \sum_{0 \leq i,j \leq t}^{t} a_{ij} x^i y^j , \tag{2}$$

where a_{ij} are the elements of symmetric matrix A of order $t \times t$. Node with identity i stores $p\left(i,y\right)$ and to establish pair wise key with the node having identity j calculates stored polynomial over point j, $k_{i,j}$. Similarly node j computes pair wise key $p\left(j,y\right)$ over point i, $k_{j,i}$. Because of symmetry property of A, $k_{i,j} = k_{j,i}$.

Matrix A is the secret information in the network and $(t+1)/2$ nodes has to compromised to calculate A.

Improvement in key pre-distribution scheme can be obtained by combination of probabilistic key pre-distribution, Q-composite key generation and Polynomial pool based key pre-distribution scheme[8]. Proposed schemes have threshold property which means that network security is maintained if number of nodes captured is less than some threshold. Comparing communication overhead, memory requirement, connectivity and security aspects improvement achieved by combination of different schemes is highlighted.

Key pre-distribution schemes are based on security vs connectivity tradeoff. Hence to achieve both security and connectivity with minimum resource overhead many researchers have focussed on asymmetric key establishment techniques suitable for Wireless sensor networks. Public key infrastructures(PKI)[2] used in computer networks requires Certification authority(CA) to bind the public key of user to its identity. Mechanism to handle large certificates and computationally intensive Digital Signature algorithms are too complex to implement on resource constrained WSN.

Shamir[9] first introduced Identity based encryption scheme which uses unique ID of the device as its public key. For computer networks, this ID can be email address or IP address. In the context of WSN, ID can be assigned by network deploying party to ensure its uniqueness. Identity based key management scheme does not require CA but another entity termed as Private key generator (PKG) is used to generate private keys from node's ID. Research on ID based key techniques for WSN focus on Pairing based cryptography (PBC) to establish pairwise key between the sensor nodes. ID-based key management scheme is implemented in MANET with key refreshment technique[10]. Apart from *Setup*, *Extract*, *Encrypt* and *Decrypt* phases in IBE, $Refresh$ phase is added to update private keys after certain amount of time. This achieves Forward secrecy and dynamic key management. Taking this work further, $Refresh$, *Recover* and *Revocation* phases are added in ID-based key management technique for WSN[11]. In their scheme more than one base stations are used to generate private key. In effect, this scheme achieves forward secrecy, backward secrecy, intrusion detection and resilience against base station capture attack. To achieve dynamic network topology cluster formation and group key management techniques are used along with key update and Revocation mechanism[12].

WSN nodes are energy constrained devices. The need of pairing algorithm and its implementation on ARM processor is studied[13] using Pairing functions from MIRACL[14] library. Pairing is considered as the most power consuming operation. Its results show that $0.444J$ power is consumed by the pairing algorithm alone. Energy consumption and execution time of point operations over super singular elliptic curve are also presented. TinyPBC[15] is another pairing algorithm for ID-based Non-Interactive Key distribution in WSN's. It demonstrates how sensor nodes can exchange keys in authenticated and non-interactive way. Paper shows that MICA2 sensor nodes with ATmega128L micro-controller $(8-bit/7.3828MHz)$ computes pairings in $5.5s$ time. K. McCusker [16] presented symmetric key distribution scheme based on Identity based cryptography

(IBC). The idea is to use asymmetric key algorithm (IBC) for authenticated key agreement and then encryption can be performed using symmetric keys generated. An accelerator hardware for Tate pairing achieves running time of $1.75ms$ and energy consumption of $0.08mJ$. These are the best result in the field of ID-based key management scheme for WSN.

3 Mathematical Background

Concept of ID-based cryptography relies on the fact that device in the network can be uniquely identified by ID assigned to it. To calculate pairwise key using ID of the device pairing based cryptography can be used. It should be noted that Identity based cryptography can be implemented in WSN only after making use of bilinear pairing properties.

Bilinear Pairing: Let G be an additive cyclic group of order n and let G_T be the multiplicative group. bilinear pairing is a computable, non-degenerate mapping function,

$$e : G \times G \to G_T$$

Bilinearity property: $\forall P, Q \in G$, and $\forall a, b \in Z*$

$$e\left([a]\,P, [b]\,Q\right) = e\left([a]\,P, Q\right)^b = e\left(P, [b]\,Q\right)^a = e\left(P, Q\right)^{ab}$$

3.1 Weil Pairing

Weil pairing maps points on elliptic curve over $GF\left(q\right)$ to the root of unity in extension field $GF\left(q^k\right)$. This implication transforms Elliptic curve discrete logarithmic problem $(ECDLP)$ in $GF\left(q\right)$ to discrete logarithmic problem (DLP) in $GF\left(q^k\right)$. Let E: Elliptic curve over prime field GF_q, Point $P, Q \in r-$torsion group. Weil pairing mapping function can be computed as

$$e\left(P, Q\right) = \frac{f_P\left(A_Q\right)}{f_Q\left(A_P\right)}$$

To calculate Weil Pairing mapping function Miller's algorithm is used twice. Following Explicit formulas are used in Miller Algorithm[3] $E : y^2 = x^3 + ax + b$
$P_1, P_2 \in E$, $P_1 = (x_1, y_1)$ and $P_2 = (x_1, y_1)$
Let $P_3 = P_1 + P2$
For $x_1 = x_2, y_1 = -y_2$,

$$V : X - x_1 = 0 \tag{3}$$

For $P_1 = P_2$ Slope is

$$\lambda_1 = \frac{3x_1^2 + a}{2y_1} \tag{4}$$

Tangent line is

$$T : Y - (\lambda_1 X + y_1 - \lambda_1 x_1) = 0 \tag{5}$$

Algorithm 1. Miller Algorithm for Weil Pairing

```
$f_1 \leftarrow V_{P+R}(Q)/L_{P,R}(Q)$
$f \leftarrow f_1$
$Z \leftarrow P$
**for** $i \leftarrow t-1$ to $0$ **do**
    $f \leftarrow f^2 T_Z(Q)/V_{2Z}(Q)$
    $Z \leftarrow 2Z$
    **if** $r_i = 1$ **then**
        $f \leftarrow f * f_1 * V_{P+R}(Q)/L_{P,R}(Q)$
        $Z \leftarrow Z + P$
    **end if**
**end for**
```

For $P_1 \neq P_2$, slope is

$$\lambda_2 = \frac{y_2 - y_1}{x_2 - x_1} \tag{6}$$

Line equation is

$$L : Y - (\lambda_2 X + y_1 - \lambda_2 x_1) = 0 \tag{7}$$

3.2 Tate Pairing

Weil pairing using Miller algorithm is not suitable for resource constrained hardware platforms. Instead Tate pairing over super singular elliptic curves in binary field can be used to implement pairing. η_T algorithm[17] computes Tate pairing over super singular curve $E(GF(2^m)) : y^2 + y = x^3 + x + b$.

Algorithm 2. η_T Algorithm

```
Input:$P, Q$
Output:$e(P, Q)$
Let $P = (x_P, y_P)$ and $Q = (x_Q, y_Q)$
$f \leftarrow 1$
**for** i← 1 to $m$ **do**
    $u \leftarrow x_p^2$
    $g \leftarrow (u+1).(x_P + x_Q) + u + y_P + y_Q + (u + x_Q + 1)s + t$
    $f \leftarrow f.g$
    $x_P \leftarrow u, y_P \leftarrow y_P^2, x_Q \leftarrow \sqrt{x_Q}, y_Q \leftarrow \sqrt{y_Q}$
**end for**
return $f^{q^2-1}$
```

3.3 Pairwise Key Establishment

Pairwise key generation between two nodes A, B can be verified using bi-linearity property as follows:
Let K_i : Private key of node i,

ID_i: Identity of node i,
e : bilinear mapping function,
K_{ij}: pair wise key for nodes i
j, H: Hash function,
s: Master secret key

$$K_{AB} = e(K_A, H(ID_B))$$

$$K_{AB} = e(s.H(ID_A), H(ID_B))$$

$$K_{AB} = e(H(ID_A), s.H(ID_B))$$

$$K_{AB} = e(K_B, H(ID_A))$$

$$K_{AB} = K_{BA}. \tag{8}$$

4 Scheme

The scheme proposes hybrid key management technique formed by the combination of symmetric and asymmetric key primitives. The main aim of the scheme is to achieve maximum secured connectivity and minimize energy and memory overhead over the entire network. Distributing the computational load among the different nodes, overall performance of the network, in terms of security, energy usage and connectivity, can be improved. We consider that all same nodes are similar with respect to their energy, memory and computational resources. Base station (BS) is assumed to have more computational power. Network architecture is heterogeneous. Network is clustered after the deployment and cluster head (CH) is selected by the base station. Identity based key management scheme is used to establish secure connection among different cluster heads and between base station and cluster head. This key establishment is dynamic and done on-line. Sensor nodes inside the cluster are securely connected by making use of polynomial based key pre-distribution primitives. Following subsections describes detailed implementation of the scheme.

4.1 Setup

Before deployment of the network cryptographic primitives need to be stored in sensor nodes and base station to establish keys in the network.

- $ID_i \leftarrow$ ID of the node i stored by deploying authority
- $e \leftarrow$ Tate pairing function over binary field $GF\,(2^m)$
- $s \leftarrow$ Master key stored in BS
- $p(i, y) \leftarrow$ bivariate polynomial calculated using symmetric matrix M, stored in node i
- $H \leftarrow$ Hash function stored in each node to map Identity of node to point on elliptic curve
- $E \leftarrow$ Elliptic curve parameters

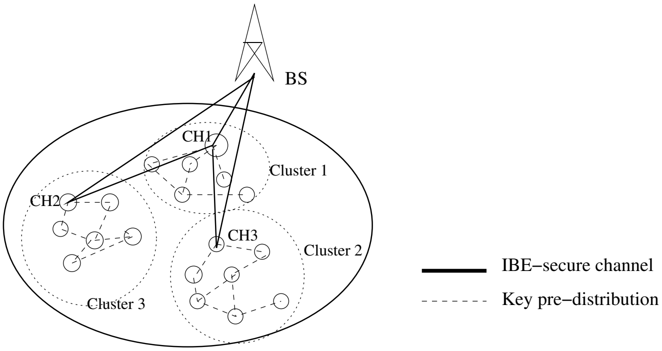


Fig. 2. Hybrid Key establishment scheme for WSN

Hash function H is nothing but Koblitz encoding method on elliptic curve[3].It is assumed that sensor nodes are deployed randomly on the field and nodes are not tamper resistant. Instead of high cost tamper resistant nodes, nodes are replaced when they found to be captured. To avoid the replay attack in the network, Time Stamp(TS) is concatenated to each message.

4.2 Secure Cluster Formation

After deployment base station selects the cluster head one by one. Number of cluster heads to be formed is programmed into base station. First cluster head is randomly selected. Following communication takes place:

$$BS \rightarrow CH_i : s.ID_{BS}\text{:Public key of BS}$$

$s.ID$ multiplication takes place over elliptic curve, hence to deduce s is computationally exhaustive task by the implication of $ECDLP$.Cluster head calculates pairwise key as follows:

$$K_{CH_i,BS} = e(s.ID_{BS}, ID_{CH_i}) \tag{9}$$

At the same time BS can also calculate pairwise key using bilinear pairing property:

$$K_{BS,CH_i} = e(s.ID_{CH_i}, ID_{BS}) \tag{10}$$

Selected cluster head sends *Hello* packet to its neighbour nodes and adds the nodes to its group upon response from them. Group list, encrypted using pairwise key calculated previously, is sent to BS. BS station stores the list and selects next cluster head which is not present in the stored list. In this way all clusters are formed in secured way.

4.3 Key Agreement Phase

Case 1: I two nodes inside the same cluster want to establish pairwise key, bivariate polynomial $p(x, y)$, given by equation (2) stored in the node is used.

For example node i and j are in same cluster. Pairwise key k_{ij} calculated as,

$$k_{ij} = p(i, y).M.p(j, y)' \tag{11}$$

Similarly, node j calculates k_{ji} and by the symmetry property of bivariate polynomial,

$$k_{ij} = k_{ji} \tag{12}$$

Case 2: When one cluster head wants to communicate with another cluster head, it request pairwise key to the base station. In this case BS act as PKG and send pairwise key to both the cluster heads through previously established secure channel. *Case 3:* If node in one cluster wants to communicate to the node in another cluster, three step key agreement is performed. First pairwise key between cluster head and node is established using bivariate polynomial. In the next step, using ID based key encryption two cluster heads are securely connected. In the last step again pairwise key between cluster head and destination node is established.

4.4 Key Update

Cluster heads are changed periodically. New cluster head is decided by the nodes in cluster depending upon the pre-defined threshold level of energy. Newly elected cluster head publish its ID to BS. BS and new cluster head generates fresh pairwise key using Tate pairing function given by Algorithm [2]. Also base station broadcast new cluster head ID to other cluster heads. Key update mechanism also suits the energy constraints of the nodes. Single node energy is avoided and at the same time key refreshment is also achieved.

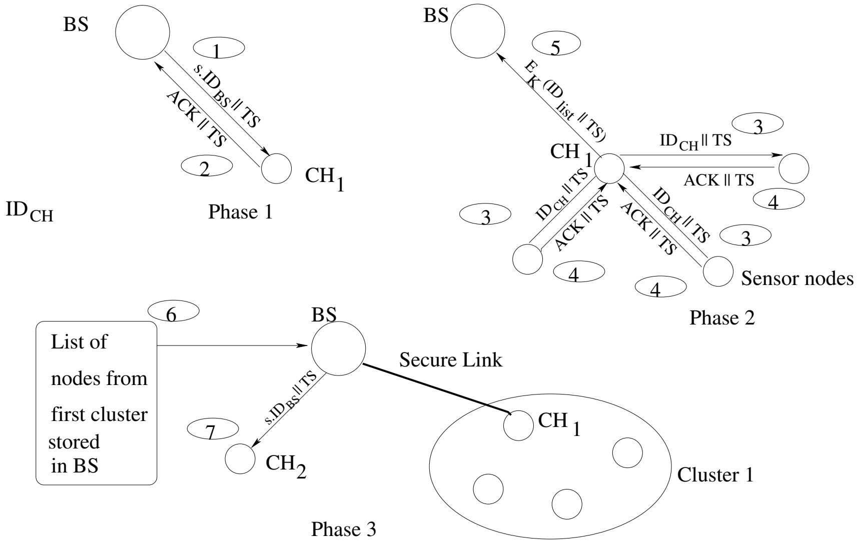


Fig. 3. Key establishment phases

4.5 Revocation

BS manages the list of authenticated cluster heads while cluster head holds the list of authenticated nodes in the cluster. As soon as number of nodes captured in cluster goes above certain threshold, cluster head reports security threat to the BS and that cluster is removed from authenticated cluster heads list. As nodes are less in cost new nodes are installed instead of using costly tamper resistant nodes.

5 Results and Discussion

Hybrid key management technique provides scalable security option for large wireless sensor networks. ID-based key technique provides secured connection at the cost of high computational requirement. Comparison of hybrid key scheme and ID-based key scheme[18] with respect to algorithm execution time and energy consumption. Instead of considering resource consumption by single node resources consumed throughout the network are analysed. For time calculation worst case scenario, in which all keys are established different times, is assumed.
Number of nodes: 500
Number of clusters: 20
Numner of nodes in cluster: 25

Table 1. Comparison with ID-based key scheme

Key Management scheme	Timing	Energy
IBK scheme (using Tate pairing)	1330s	31.365J
Hybrid key scheme	53.2s	1.255J

Different key management related issues of the proposed hybrid key technique are discussed as follows.
Scalability: Cluster head formation mechanism adopted in the scheme allows large sensor nodes to be deployed with minimum overhead on the memory and energy resources. Cluster head formation mechanism is secure and new clusters can be easily added without causing any threat to security to expand the network.
Forward and backward secrecy: Because of periodic key update, new nodes can not detect previous messages. Revocation phase take care that old nodes in the network should not be able to read new messages in the network. *Communication overhead:* Non interactive key establishment using ID based cryptography minimizes communication overhead.
Memory overhead: As polynomial pool based key technique is used at cluster level less number of polynomial coefficients are need to be stored. If m cluster are formed for n number of nodes, memory overhead is reduced by factor $\frac{m}{n}$

Energy consumption: Cluster head consumes most of the energy and there is a chance of single node energy drain out. Energy consumption is distributed among all the nodes as cluster head is update periodically. Most of the node expect cluster heads uses polynomial based key technique which requires less energy compared to IBE. Also energy is conserved by avoiding communication between cluster heads.

6 Conclusion

Key management technique designed by combination of symmetric and asymmetric key primitives over different levels of hierarchy proves to be an effective solution for resource constrained networks. Security of the overall network is improved compared to polynomial based key pre-distribution scheme with minimized memory overhead. At the same time, energy consumption due to computationally intensive IBK scheme is limited to the nodes in higher level of hierarchy. Also, Cluster head rotation policy avoids energy drain of single node and distribute energy overhead among different nodes. Energy consumption and execution time results calculated for the proposed schemes shows that considerable amount of energy and time overhead can be reduced by the application hybrid key management scheme.

References

1. Atzori, L., Iera, A., Morabito, G.: The internet of things: A survey. Computer Networks 54(15), 2787–2805 (2010)
2. William, S., et al.: Cryptography and Network Security, 4/e. Pearson Education India (2006)
3. Hankerson, D., Menezes, A., Vanstone, S.: Guide to elliptic curve cryptography. Springer (2004)
4. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 41–47. ACM (2002)
5. Chan, H., Perrig, A., Song, D.: Random key predistribution schemes for sensor networks. In: Proceedings of 2003 Symposium on Security and Privacy, pp. 197–213. IEEE (2003)
6. Shan, T., Liu, C.: Enhancing the key pre-distribution scheme on wireless sensor networks. In: IEEE Asia-Pacific Services Computing Conference, APSCC 2008, pp. 1127–1131. IEEE (2008)
7. Blom, R.: An optimal class of symmetric key generation systems. In: Beth, T., Cot, N., Ingemarsson, I. (eds.) EUROCRYPT 1984. LNCS, vol. 209, pp. 335–338. Springer, Heidelberg (1985)
8. Rasheed, A., Mahapatra, R.: Key predistribution schemes for establishing pairwise keys with a mobile sink in sensor networks. IEEE Transactions on Parallel and Distributed Systems 22(1), 176–184 (2011)
9. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985)

10. Balfe, S., Boklan, K.D., Klagsbrun, Z., Paterson, K.G.: Key refreshing in identity-based cryptography and its applications in manets. In: IEEE Military Communications Conference, MILCOM 2007, pp. 1–8. IEEE (2007)
11. Saab, S., Kayssi, A., Chehab, A.: A decentralized energy-aware key management scheme for wireless sensor networks. In: 2011 International Conference for Internet Technology and Secured Transactions (ICITST), pp. 504–508. IEEE (2011)
12. Jian-wei, J., Jian-hui, L.: Research on key management scheme for wsn based on elliptic curve cryptosystem. In: First International Conference on Networked Digital Technologies, NDT 2009, pp. 536–540. IEEE (2009)
13. Doyle, B., Bell, S., Smeaton, A., Mccusker, K., O'Connor, N.: Security considerations and key negotiation techniques for power constrained sensor networks. The Computer Journal 49(4), 443–453 (2006)
14. Scott, M.: Miracl–multiprecision integer and rational arithmetic c/c++ library. Shamus Software Ltd., Dublin, Ireland (2003), http://www.shamus.ie
15. Oliveira, L., Scott, M., Lopez, J., Dahab, R.: Tinypbc: Pairings for authenticated identity-based non-interactive key distribution in sensor networks. In: 5th International Conference on Networked Sensing Systems, INSS 2008, pp. 173–180 (June 2008)
16. McCusker, K., O'Connor, N.E.: Low-energy symmetric key distribution in wireless sensor networks. IEEE Transactions on Dependable and Secure Computing 8(3), 363–376 (2011)
17. Barreto, P., Galbraith, S., hÉigeartaigh, C., Scott, M.: Efficient pairing computation on supersingular abelian varieties. Designs, Codes and Cryptography 42(3), 239–271 (2007)
18. Szczechowiak, P., Kargl, A., Scott, M., Collier, M.: On the application of pairing based cryptography to wireless sensor networks. In: Proceedings of the Second ACM Conference on Wireless Network Security, pp. 1–12. ACM (2009)

DRMWSN-Detecting Replica Nodes Using Multiple Identities in Wireless Sensor Network

Nagaraj Ambika and G.T. Raju

Dayananda Sagar College of Engineering,
Research Scholar of Bharathiar University, Bangalore, India
Prof & Head, Dept of CS & Engg, RNSIT, Bangalore
ambika.nagaraj76@gmail.com, drgtraju_rnsit@yahoo.com

Abstract. Wireless sensor network are prone to different types of attacks due to lack of supervision. Trusting the data received from the network becomes quite difficult. Implementing prevention and detection techniques provide strong impediment to the network from getting compromised. In this paper both the techniques are being utilized providing better reliance over the data being communicated. Pair-wise keys and group wise keys are being generated, which also provides identity of the nodes at that instant of time. This technique deters wormhole attack to a large extent.

Keywords: prevention and detection technique, steganography, group-key generation, pair-wise key generation.

1 Introduction

Wireless sensor network are low cost nodes deployed in unattended areas which sense, collect and transmit data to the base station. These nodes are used widely in many applications like habitat monitoring [1], [2] [3], forest fire detection [4], [5], military applications [6] and so on. These nodes are prone to failures due to the low manufacturing cost. These nodes are limited in power and hence the battery power needs to be utilized carefully. This pursues them to change the topology frequently. As these nodes are utilized to send some classified material, the chances of these nodes getting hacked are quite discernible. An adversary can take control of the nodes or can modify the information sent in the course of its travel. The adversaries can pose itself as one of the nodes and try to gather all the data sent by other nodes. To counteract this type of attacks, data is encrypted accompanied by authenticating the nodes. To bring this act in to play, the nodes can either prevent itself getting compromised or the base station/ neighboring node/ cluster head can detect the compromised node and can cease to send data to the compromised node. Using both the techniques together would help to elevate the security of the network by guarding the data from unauthorized acsessors.

Many preventive and detection algorithms [17] are being suggested. This paper is based on graph theory. Pair-wise key is generated using bipartite graph. Group keys are generated using multipartite graph and Hungarian algorithm. The paper also

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 866–878, 2013.

utilizes steganography, a technique used to transmit message hidden inside another text message. Steganography [9] is a technique through which the data which is communicated does not come to the notice of its adversaries.

This paper is being divided into sections. Section 1&2 gives a description of the graph theory and set theory concepts utilized in the paper. Section 3 provides related work done in the similar field. Section 4 provides a detail description of proposed model. Section 5 furnishes the simulated results of how the suggested protocol escalates security in the network. Section 6 provides the conclusion and suggests the further work in this field.

2 Preliminaries and Notations

Consider the entire network as a subset of clusters. The cluster is represented as a sub-graph and the union of all the sub-graph constitutes the network. Each node in the cluster is considered as a vertex and the keys generated are considered as edges.

The nodes are pre-deployed with keys which are mutually exclusive from each other. The nodes can generate pair-wise key or a group key. Two nodes from the cluster are chosen to generate pair-wise key.

Table 1. Notations used in DRMWSN

Notation	Meaning
D	Detector
BS	Base station/ sink
$N_i \rightarrow N$	Ni node in the cluster broadcast HELLO message
C_i	ith Cluster in the network
$A \rightarrow B$:msg	A sends message to B
$BS \rightarrow N$: msg	Base station broadcast message to the network N
$K \rightarrow K_i \| K_j$	Generation pair-wise key
$K \rightarrow K_i \| K_j \| K_m \| K_n \| K_o$	Generation of group key
$K(N_i)$	Key of node N_i
G	Entire network considered as graph
S	Steganography message broadcasted by the base Station
ADDR(OFFSET_ADDR)	OFFSET_ADDR considered as the starting address
ADDR(STEGO_MSG)	Address dispatched by the base station using steganography
EN_DATA(N_i)	Encrypted data of node N_i

Definition 1: let B be a bipartite graph if there is $N_1,N_2...N_n \subseteq G_i$ where I = {1,2…….n}. let $N_i = \{K_1, K_2,.......K_n\}$be the set of keys stored in the node such that $K(Ni) \cap K(Nj) = \emptyset$.

Definition 2: let B be a multipartite graph if there is N1,N2...Nn $\subseteq$ Gi where I = {1,2…….n}. let Ni $=\{K_1, K_2,.......K_n\}$be the set of keys stored in the node such that $K(N_i) \cap K(N_j) \cap ....K(N_m) = \emptyset$.

Definition 3: Consider the bipartite graph B, let F: $K(N_i) \rightarrow K(N_j)$ be a mapping between the unmatched keys of the nodes.

Definition 4: let $S= S_1 \cup S_2 \cup ..... \cup S_n$ be the message broadcasted by the base station to the network . Calculate (F: $K(N_i) \rightarrow K(N_j)) \rightarrow ((ADDR(K(N_i)) \cup$ offset_addr) $\subseteq$ S)

3 Related Work

In [23] a framework is provided which is used to study the security of key pre-distribution schemes. During the key pre-distribution phase, key information is assigned to each node, such that after deployment, neighboring sensor nodes can find a secret key between them. key pre-distribution phase generates G and D matrices, followed by the selection of a key space. Then, in the key agreement phase, after deployment, each node discovers whether it shares any key space with its neighbors. To achieve this, each node broadcasts a message containing the node's ID, the indices of key spaces it carries, and the seed of the column of G it carries. The paper proposes a new key pre-distribution scheme which substantially improves the resilience of the network compared to previous schemes, and give an in-depth analysis of our scheme in terms of network resilience and associated overhead.

In [19] Erdos and Renyi component theory is utilized. This theory shows inspite of small node's degree the network remains connected. The paper evaluates relation between connectivity, memory size and security.

In [17] authentication-based intrusion prevention and energy-saved intrusion prevention is being utilized to improve security of cluster-based sensor network. The member nodes take turn to monitor the cluster head.

[20] presents a deterministic key distribution scheme based on Expander Graphs. It shows how to map the parameters (e.g., degree, expansion, and diameter) of a Ramanujan Expander Graph to the desired properties of a key distribution scheme for a physical network topology.

4 Intruder Model

The intruder's main intention is to take control of the entire network. It can accomplish this task by taking control of all the nodes in the network. The intruder with the help of a compromising node can mask itself as one among them and divert all the traffic towards itself, thereby draining the energy of the network and can mislead the base station by modifying the data being sent by other nodes of the network. Other uncompromised nodes in the network will be unable to find the intention of the intruder and will hence share some secret information within them. If other nodes utilize the same encryption key distributed by the compromise node, the intruder after blocking the data can decrypt data easily. This activity not only affect the cluster in which the compromised node reside, but also the other nodes in the network as the compromised node can advertise itself as a node near to the base station. Uncompromised nodes in-turn divert all their data to this compromised node. The intruder can make a replica of the compromised nodes and place them in different location and can take the control of the entire network.

5 DRMWSN Model

5.1 System Model

The paper utilizes Tinynode 584 to be distributed in the required environment. It is a low powered OEM module providing simple and reliable way to add wireless communication to sensors. TinyNode 584 is optimized to run TinyOS and packaged as a complete wireless subsystem with 19 configurable I/O pins offering up to 6 analog inputs, up to 2 analog outputs.

Table 2. illustrates the current consumption of TinyNode

Mode	Energy Consumption
Sleep, time off	0.004 mA
Sleep, time on	0.007 mA
µC only	2 mA
Receive	16 mA
Transmit(0dbm)	25 mA
Transmit(10dbm)	46 mA

5.2 Key Distribution and Deployment of Sensors in the Field

Set of keys are generated by the base station. All the sensors are embedded with subset of keys. Care is taken that the keys are mutually exclusive. Each node has an offset address(signifies the actual starting address i.e. the node masks over the actual address to calculate the starting address) , which differs from the rest of the other cluster members.

5.3 Formation of Cluster

To authenticate each other, unique id’s are embedded inside sensors. The sensors after deployment broadcast HELLO message. The sensors which respond to the broadcasted message are pooled to form a cluster. They authenticate each other by utilizing encrypted unique ID stored inside them. After the cluster is formed, the cluster members choose the cluster head and the subsequent node depending on the energy stored in each node.

$$N_i \rightarrow N:msg$$

5.4 Communications between Cluster Members and Base Station

Steganography is an art and science used to embed secret data inside the cover data by utilizing embedded algorithm and steganography key. This paper utilizes this technique, where the base station broadcast message to the entire network by embedding the key inside the cover text.

$$BS \rightarrow N : msg$$

5.5 Generating of Key

5.5.1 Generation of Pair-Wise Key

The nodes in the network have to interpret the message sent by the base station. The nodes have to obtain the location of key , identify which node should participate in generation of the key,& global time. Within the time limit the key has to be generated and distributed to other members of the cluster. The nodes authenticate each other and form a pair-wise key.

$$START\_ADDR \rightarrow OFFSET\_ADDR$$

$$K_i \rightarrow ADDR(STEGO\_MSG)$$

$$K \rightarrow K_i \,\|\, K_j$$

$$K \rightarrow C_i$$

Every node differs in their offset address. The offset address is added to the location obtained from the steganography message broadcasted by the base station. Hence the mapping between the keys is decided by the base station. The base station will have a prior idea as to which key is being utilized in the pair-wise key[18] generation. If any nodes do not respond within the time limit, the detector isolates the node and keeps the node under observation. If Ith node respond with Ki key within the time limit and jth node does not respond , then Ki is declared as the key to encrypt the messages. considering $K_j = \emptyset$

$$K \rightarrow K_i \ \phi$$

After several observations, the suspected node if termed as compromised node, its ID is removed and the detector D sends a message to the base station, which in turn broadcast the message to the network

$$D \rightarrow BS:msg$$

$$BS \rightarrow N{:}msg$$

The generated key is being distributed to the cluster members and the members of the cluster in turn encrypt the data to be transmitted using the key. The message is forwarded to the next cluster head by the present cluster head.

$$N_C \rightarrow EN\_DATA(N_i) \parallel EN\_DATA(N_j) \parallel EN\_DATA(N_j) \parallel EN\_DATA(N_m) \parallel EN\_DATA(N_n) \parallel EN\_DATA(N_o)$$

5.5.2 Generating Group-Key

All the nodes in the cluster participate forming a group key utilized to encrypt the sensed data. After receiving the broadcasted message from the base station, all the nodes should start to form a group key. If the key from any one of the nodes is not transmitted during the time limit, the node comes under suspicion that it may be a compromised node. After some amount of observations the suspected node is confirmed to be compromised or uncompromised node. If the detector D concludes the suspected node to be compromised node as a compromised node, the detector sends the report to the base station. The base station in turn broadcast the message to the network.

$$START\_ADDR \rightarrow OFFSET\_ADDR$$

$$K_i \rightarrow ADDR(STEGO\_MSG)$$

$$K \rightarrow K_i \parallel K_j \parallel K_m \parallel K_n \parallel K_o$$

$$K \rightarrow C_i$$

$$D \rightarrow BS{:}msg$$

$$BS \rightarrow N{:}msg$$

If one of the key is not been in the participation, then that key is assumed to be null and the rest of thenodes participate to generated the group key. The key is been distributed to the rest of the cluster members to encrypt the data.

$$K \rightarrow K_i \parallel K_j \parallel K_m \parallel k_n \parallel K_o$$

$$N_C \rightarrow EN\_DATA(N_i) \parallel EN\_DATA(N_j) \parallel EN\_DATA(N_j) \parallel EN\_DATA(N_m) \parallel EN\_DATA(N_n) \parallel EN\_DATA(N_o)$$

6 Security Analysis

The following scenario is considered.

6.1 Cluster Head/Subsequent Node Is Compromised

Subsequent node and the cluster head authenticate each other, hence if any one of the node is compromised the report is generated by the other and sent to the base station. The base station broadcast the blacklisted node to other nodes in the network. The cluster which has the compromised node, chooses another node in its place. If both the cluster head and subsequent nodes are compromised, the node will not be able to read the hidden message broadcasted by the base station. Either the node will not transmit the packets in the scheduled time or will utilize different keys for encryption. If the base station suspects that the cluster head and subsequent node are compromised, it informs other nodes in the network to black list them. Other nodes in the cluster elect their cluster head and subsequent nodes among themselves.

6.2 Cluster Members Are Compromised

If any of the cluster members are compromised, the cluster head will know at the time of authentication. The cluster head sends the report to the base station. The base station in turn, broadcasts the message to other nodes in the network. The nodes in the network mark this node as blacklisted node and do not accept or forwards packets to this node.

7 Simulated Results

The work is simulated using NS2. Considering the length of the encryption key to be 128 bits, tiny node can store around 512 keys. Hence a cluster can generate 512 combination of keys. The following are the outcomes of the paper.

Table 3. Simulated Results

	Pair-wise keys	Group keys
Distribution of nodes	Uniform	Uniform
Number of nodes in the network	Multiples of 1000	Multiples of 1000
Probability of detection of compromised nodes	>=0.4	>=0.99
Probability of false alarm	<=0.1	<=0.001
Probability of data integrity	>=0.8	>=0.98
Probability of reliable data reaching base station	>=0.89	>=0.98

7.1 Energy Consumption

Energy is one of the important features in wireless sensor network. As the sensors are deployed in unattended areas, the battery cannot be recharged. Sensors either run out of battery power or come under the control of the adversaries. In both the cases, the network will get shortage of sensors which in turn will not provide accurate readings. Hence to avoid the loss, the battery power has to be consumed very carefully. but at the same time a prevention mechanism has to be implemented so that the sensors are provided with a protective shield to safe guard themselves. This paper provides detection and prevention model which better protection against the adversaries. This paper utilizes 24% more energy than a normal pre- distribution key use in wireless sensor network (PKWSN). Utilizing this technique helps the sensors to shield itself from the adversaries and in any case if any sensor gets compromised, the rest of the nodes in the cluster is prevented from getting compromised.

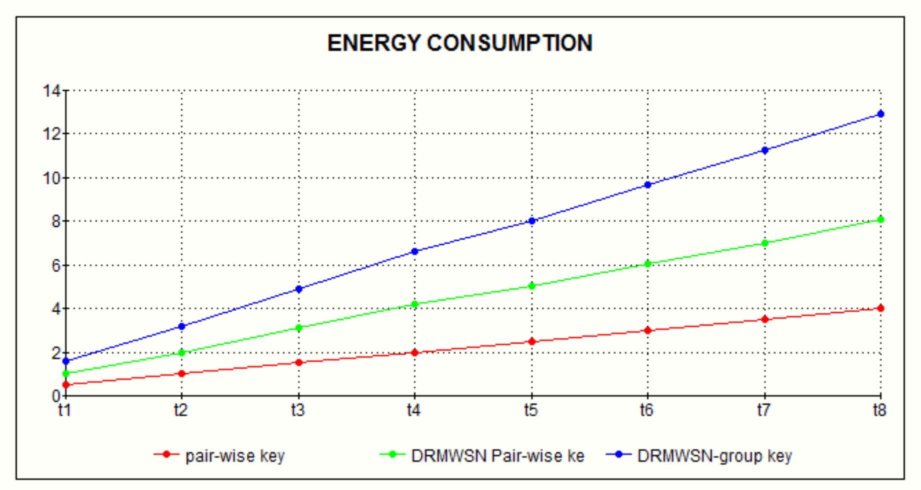


Fig. 1. Energy Consumption in PKWSN and DRMWSN

7.2 Sybil Attack

One of the priority security issues in any kind of network would be authentication. Unless the receiver is an authorized node to receive the data, there is a large possibility that the data can get modified leading to breach integrity. In a large network consisting of multiples of thousands of nodes, it is quite a difficult task to keep a track of data being transferred from one node to another. Hence every cluster have to shield itself from such types of attack. The primary task to avert such type of attacks is to device a strong authentication technique which makes it unique from other clusters. The base station will find it intricate if each node has a unique authentication code. The process simplifies if each cluster has an authentication code which cannot be replicated by adversary. This paper utilizes a technique where in any two nodes of the cluster involving in generating pair-wise key will make the network secure. This also provides authentication of the source (cluster). If the mapping goes wrong in any one of the cluster, the base station will be able to track it easily and consider it as malicious node by evaluating the report sent by the detector of the cluster. This paper provides 22.3% more security than a regular pre-distribution key in network (PKWSN). The fig 2, provides a pictorial representation of PKWSN and DRMWSN against Sybil attack [14], [15], [16]. X-axis denotes the time and Y-axis denotes percentage of nodes deployed.

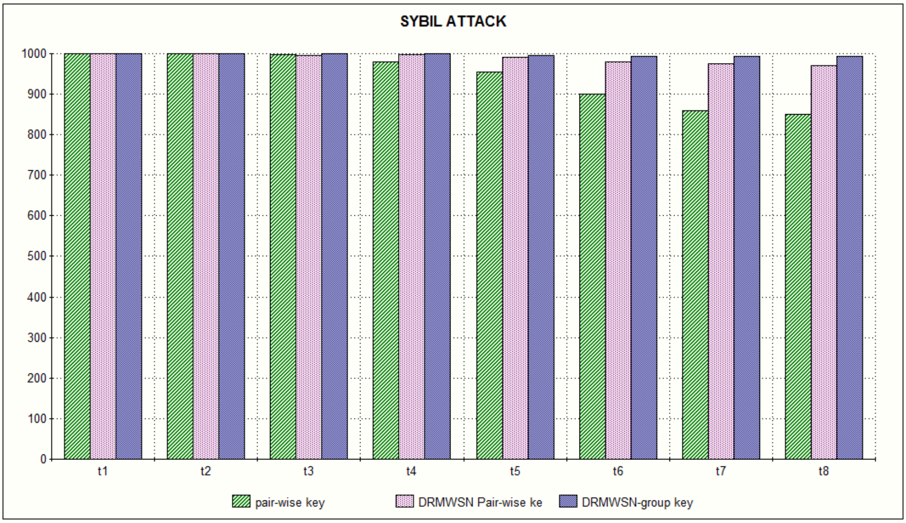


Fig. 2. Illustration of Sybil attack in PKWSN and DRMWSN

7.3 Sinkhole Attack

Sinkhole attack[11], [12], [13], [24],[26] is a type of attack, which attracts all the traffic towards itself posing as one of the vested nodes in the network to which nodes could forward the data. It disguises itself as the one of the nodes nearer to the base station. If the node is positioned nearer to the base station, it acquires almost all the packets moving towards the base station. The adversary can modify the packets or can even deny forwarding the packets towards the base station. This paper reduces this attack by 8.4%. Fig 3 illustrates the working of DRMWSN against sinkhole attack and provides a comparison to PKWSN. The nodes in the network will change the encryption key when ever intimated by the base station which keeps the data safe from the adversary. The nodes will be on continuous check by detector. As the cluster members are authenticated continuously after every session, the compromised nodes can be detected and notified.

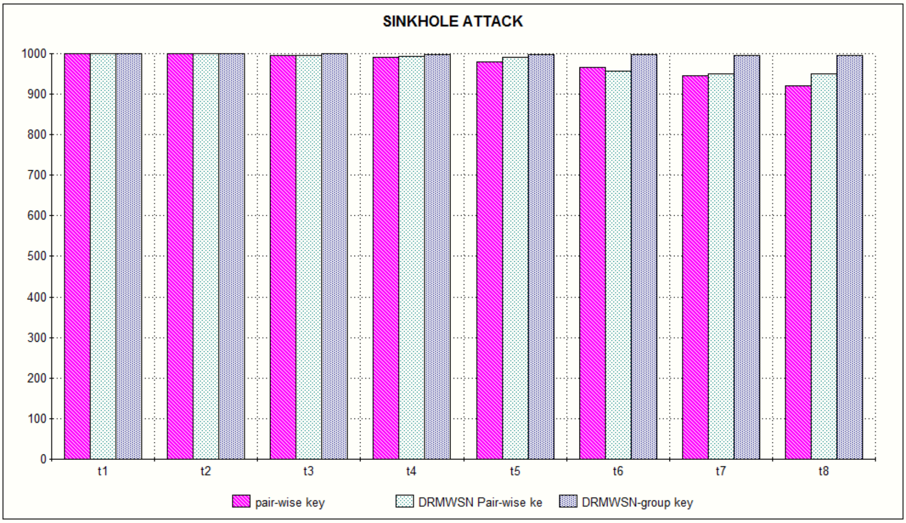


Fig. 3. Illustration of Sinkhole attack in PKWSN and DRMWSN

7.4 Wormhole Attack

Wormhole attack [7], [8], [25], [9], [10] is a kind of attack where the adversary tunnels the data from one location to another and retransmits the data. Due to this activity the base station will not get correct readings from the exact location. Added to this, the base station will not receive the required information in time. This paper secures the network by 16.3% against wormhole attack. The base station will have a prior knowledge of the keys being stored in each sensor node and as the keys keep on changing the base station will be able to detect the compromised node in the network.

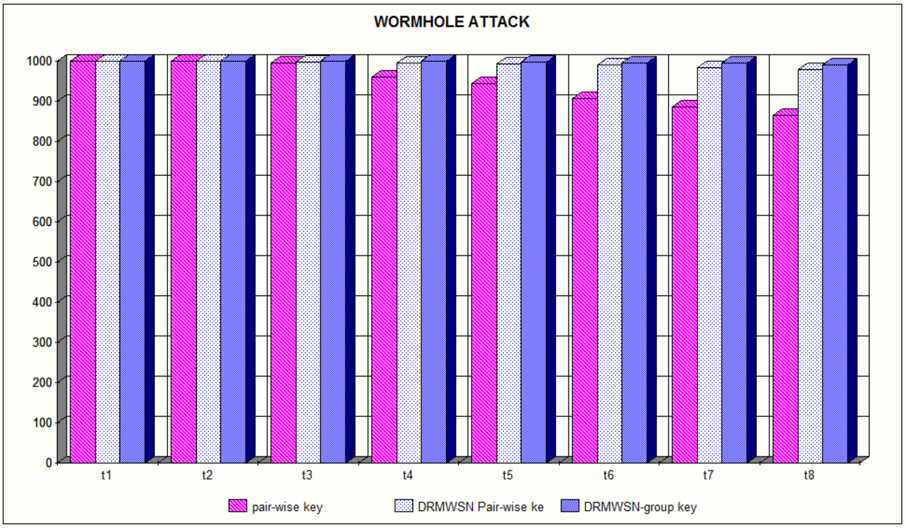
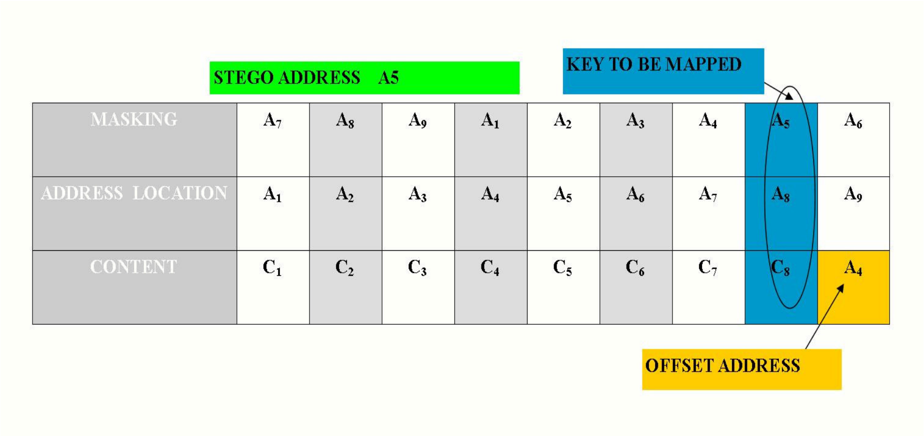


Fig. 4. Illustration of Wormhole attack in PKWSN and DRMWSN



MASKING	A_7	A_8	A_9	A_1	A_2	A_3	A_4	A_5	A_6
ADDRESS LOCATION	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
CONTENT	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	A_4

Fig. 5. Implementation of DRMWSN on nodes

8 Conclusion

This paper increases the security of the network to a larger extent. The paper is based on graph theory and set theory. The pair-wise key/group key is established by the nodes of the cluster and being circulated to other members of the cluster. The keys are unique among themselves and changing of the encryption keys after every broadcast by the base station makes it more resilient against different kinds of attack. The base station utilizes steganography technique to transmit data which it needs to communicate to its nodes in the network. This in turn makes the transmission safer.

References

1. Mainwaring, A., Culler, D., Polastre, J., Polastre, J., Polastre, J.: Wireless Sensor Networks For Habitat Monitoring. In: Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, doi:10.1145/570738.570751
2. Szewczyk, R., Osterweil, E., Polastre, J., Hamilton, M., Hamilton, M., Hamilton, M.: Habitat monitoring with sensor networks. Magazine of Communications of the ACM - Wireless Sensor Networks 47(6) (2004)
3. Naumowicz, T., Freeman, R., Kirk, H., Dean, B., Calsyn, M., Liers, A., Braendle, A., Guilford, T., Schiller, J.: Wireless Sensor Network for habitat monitoring on Skomer Island. In: IEEE 35th Conference on Local Computer Networks (LCN), pp. 882–889 (2010), doi:10.1109/LCN.2010.5735827
4. Hefeeda, M., Bagheri, M.: Wireless Sensor Networks for Early Detection of Forest Fires. In: IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems, pp. 1–6. Simon Fraser Univ., Surrey (2007), doi:10.1109/MOBHOC.2007.4428702
5. Mal-Sarkar, S., Sikder, I.U., Konangi, V.K.: Application of wireless sensor networks in forest fire detection under uncertainty. In: 13th International Conference on Computer and Information Technology (ICCIT), pp. 193–197 (2010), 10.1109/ICCITECHN.2010.5723853

6. Lee, S.H., Lee, S., Song, H., Lee, H.S.: Wireless sensor network design for tactical military applications: Remote largescale environments. In: Military Communications Conference, pp. 1–7. IEEE (2009), doi:10.1109/MILCOM.2009.5379900
7. Hu, Y.-C., Perrig, A., Johnson, D.B.: Wormhole attacks in wireless networks. IEEE Journal on Selected Areas in Commnication 24(2), 370–380 (2005)
8. Zhao, Z., Wei, B., Dong, X., Yao, L., Gao, F.: Detecting Wormhole Attacks in Wireless Sensor Networks with Statistical Analysis. In: International Conference on Information Engineering (ICIE), pp. 251–254 (2010), doi:10.1109/ICIE.2010.66
9. Chen, H., Lou, W., Sun, X., Wang, Z.: A secure localization approach against wormhole attacks using distance consistency. Journal EURASIP Journal on Wireless Communications and Networking - Special Issue on Wireless Network Algorithms, Systems, and Applications 2010, doi:10.1155/2010/627039
10. Modirkhazeni, A., Aghamahmoodi, S., Modirkhazeni, A., Niknejad, N.: In: The 7th International Conference on Networked Computing (INC), pp. 122–128 (2011)
11. Sharmila, S., Umamaheswari, G.: Detection of Sinkhole Attack in Wireless Sensor Networks Using Message Digest Algorithms. In: International Conference on Process Automation, Control and Computing (PACC), pp. 1–6 (2011), doi:10.1109/PACC.2011.5978973
12. Krontiris, I., Giannetsos, T., Dimitriou, T.: Launching a Sinkhole Attack in Wireless Sensor Networks; The Intruder Side. In: IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, WIMOB 2008, pp. 526–531 (2008), doi:10.1109/WiMob.2008.83
13. Ngai, E.C.H., Liu, J., Lyu, M.R.: An efficient intruder detection algorithm against sinkhole attacks in wireless sensor networks. Journal Computer Communications 30(11-12) (2007), doi:10.1016/j.comcom.2007.04.025
14. Ssu, K.-F., Wang, W.-T., Chang, W.-C.: Detecting Sybil attacks in Wireless Sensor Networks using neighboring information. The International Journal of Computer and Telecommunications Networking 53(18) (2009), doi:10.1016/j.comnet.2009.07.013
15. Newsome, J., Shi, E., Song, D., Perrig, A.: The sybil attack in sensor networks: analysis & defenses (January 2004)
16. Xiu-Li, R., Wei, Y.: Method of Detecting the Sybil Attack Based on Ranging in Wireless Sensor Network. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2009), doi:10.1109/WICOM.2009.5302573
17. Su, C.-C., Chang, K.-M., Kuo, Y.-H., Horng, M.-F.: The new intrusion prevention and detection approaches for clustering-based sensor networks. In: IEEE Conference on Wireless Communication and Networking, vol. 4, pp. 1927–1932 (2005), doi:10.1109/WCNC.2005.1424814
18. Liu, D., Ning, P., Li, R.: Establishing pairwise keys in distributed sensor networks. ACM Transactions on Information and System Security 8(1) (2005), doi:10.1145/1053283.1053287
19. Hwang, J., Kim, Y.: Revisiting random key pre-distribution schemes for wireless sensor networks. In: Proceedings of the 2nd ACM Workshop on Security of Ad Hoc and Sensor Networks (2004), doi:10.1145/1029102.1029111
20. Camtepe, S.A., Yener, B., Yung, M.: Expander Graph based Key Distribution Mechanisms in Wireless Sensor Networks. In: IEEE International Conference on Communication, pp. 2262–2267 (2006), doi:10.1109/ICC.2006.255107
21. Sajedi, H., Jamzad, M.: Secure cover selection steganography. In: Park, J.H., Chen, H.-H., Atiquzzaman, M., Lee, C., Kim, T.-h., Yeo, S.-S. (eds.) ISA 2009. LNCS, vol. 5576, pp. 317–326. Springer, Heidelberg (2009)

22. Turner, C.: A steganographic computational paradigm for wireless sensor networks. In: International Conference on Innovations and Information Technology, pp. 258–262 (2009), doi:10.1109/IIT.2009.5413637
23. Du, W., Deng, J., Han, Y.S., Varshney, P., Katz, J., Khalili, A.: A Pairwise Key Predistribution Scheme for Wireless Sensor Networks. The ACM Transactions on Information and System Security (TISSEC) 8(2), 228–258 (2005)
24. Chen, C., Song, M., Hsieh, G.: Intrusion detection of Sinkhole attack in largescale wireless sensor network. In: WCNIS 2010, pp. 711–716 (2010)
25. Labraoui, N., Gueroui, M., Aliouat, M.: Secure DVHop localization scheme against wormhole attacks in wireless sensor networks. In: European Transactions on Telecommunications, doi:10.1002/ett.1532
26. Krontiris, I., Giannetsos, T., Dimitriou, T.: Launching a Sinkhole Attack in Wireless Sensor Networks; The Intruder Side. In: WiMob 2008, pp. 526–531 (2008)

Security Improvement in Group Key Management

Manisha Manjul*, Rakesh Kumar, and Rajesh Mishra

School of ICT, Gautam Buddha University, Greater Noida, India
{manisham,rmishra}@gbu.ac.in

Abstract. Multicast is a one to the group communication which have various challenges such as group key management, multicast receiver access control, multicast finger printing and multicast source authentication. Various protocols introduced by many researchers to minimize the lacks such as computational, communication, message size and storage overheads for group key management, but these proposed methods still have some lack as discussed above, while rekeying cost is also not less. Therefore to provide a solution of existing problem after leaving a group, there is a need for efficient and improved mechanism for group key management.

Keywords: group key management, multicast security, rekeying, communication overheads.

1 Introduction

Computer network is basically the combination of computers and different type of devices that are interfaced by various resources via communication channels that provide the communications among users and allows them facility of sharing the resources [23]. Multicast is one of the service that provide different type of ways for communication such as one to many, many to one, many to many.

Multicast refers to the transmission of a message from one sender to multiple receivers or from multiple Senders to multiple receivers [14]. The advantage of multicast is that, it enables the desired applications to service many users without overloading a network and resources in the server. If the same message is to be sent to different destinations, multicast is preferred to multiple unicast. Group Communication introduces the challenging issues relating to group confidentiality and key management, when a source that sends data to a set of receivers in a multicast session. The security of the session is managed by two main functional, first is Group Controller (GC) responsible for authentication, authorization and access control, where as second is known a Key Server (KS) responsible for the maintenance and distribution of the required key material [11].

Security is the one of the issue in the multicast. There are four type of multicast security such as multicast receiver access control, multicast source authentication, multicast fingerprinting and group key management. All these multicast security have

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 879–893, 2013.

some issues and researchers provided solution for multicast security issues [8]. In this paper the authors have focused on group key management on security related issues. As we discuss the type of multicast communication, there are three types of multicast communication *i.e.* one-to-many, many-to-many and many-to-one. In each type of communication the idea is the same; the sender directs all datagram to a single IP address once and the datagram are delivered to every member of the multicast group [22]. One-to-many communication there exists only one sender and more than one hosts that are listening to the senders datagram. This is natural way for all types of on demand and file distribution services. One of the examples of one-to-many communications is telecast movies and all kinds of TV material.

Many-to-many multicast communication occurs usually when we are dealing with group communication. Video conferencing and other conferencing services are the example of many to many communications. More specifically these might include online gaming and online mentoring systems.

Many-to-one type of communication is very useful when we are providing high availability resource discovery and data collection services to large amount users. Auctioning services is the example of many to one communication. Multicasting is achieved with special routers, which keep track of all the networks within its routing domain that contain multicast/host group members. The routers do not have to keep track of all the members of multicast group. They just need to know the networks towards which they should copy the multicast datagram. In principle, the sender doesn't have to keep track of all the recipients either. But when we keep in mind the nature of most multicasting services, in practice there has to be some entity on behalf of service provider that registers all the receiving parties.

Multicast communication suffers from receiver access problem due to forward secrecy, backward secrecy. The group key management is an efficient mechanism to handle this situation. But there are many factors which effect the communication [6, 12, 13, and 14], computation overhead, message size, storage overhead, these factors are as following:

a. Heterogeneous nature of the group membership affects the possible type of encryption algorithm to be used, and the length of the key that can be supported by an end user.
b. The cost of setting up and initializing the entire system parameters, such as selection of the group controller (GC), group announcement, member join and initial key distribution.
c. Administrative policies, such as those defining which members have the authorization to generate keys.
d. Required level of performance of parameters, such as session sustainability, and key generation rates.
e. Required additional external support mechanisms, such as the availability of a certificate authority (CA).

Therefore it is required efficient group key management approach to secure the system and reduce the overhead in the existing approach [9]. Existing key graph [17] proposes the extension of the binary key tree to 4-ary key tree and 4-ary key tree overcome the problem of re-keying in terms of height of the key tree. Using a greater degree reduces the height of the key tree and, as a result, improves re-keying performance. Performance of re-keying measured in terms of computation overhead, communication overhead, message size and storage overhead. Really, optimal results are gained when the tree has a degree of 4. In the figure 1(a) illustrates the logical key tree with two nodes when there are seven joining members (u_1 through u_7). When u_8 joins, the key server first attaches it to node $K_{1,2}$ as shown in figure 1(b), and then, changes the group key K_G and the node key $K_{1,2}$ to $K`_G$ and $K`_{1,2}$ respectively. For delivering them, each new key is encrypted with the previous one (K_G and $K_{1,2}$ respectively), and a set of them are sent by multicast for existing members. For the new member they are sent by unicast being encrypted with its session key. On the other hand, when a member leaves the group, new keys are encrypted by their corresponding child keys, and a set of them are sent for remaining members by multicast. For example, when member u_8 leaves the group shown in figure 5(b), the key server changes $K`_{1,2}$ and $K`_G$ to $K``_{1,2}$ and $K``_G$ respectively. Then, it delivers $K``_{1,2}$ for $\{u_5,u_6,u_7\}$ being encrypted by $K_5$, $K_6$ and $K_7$, and $K``_G$ for $\{u_1,u_2,u_3,u_4\}$ and $\{u_5,u_6,u_7\}$ being encrypted by $K_{1,1}$ and $K``_{1,2}$ respectively. A set of these keys are sent by one multicast message.

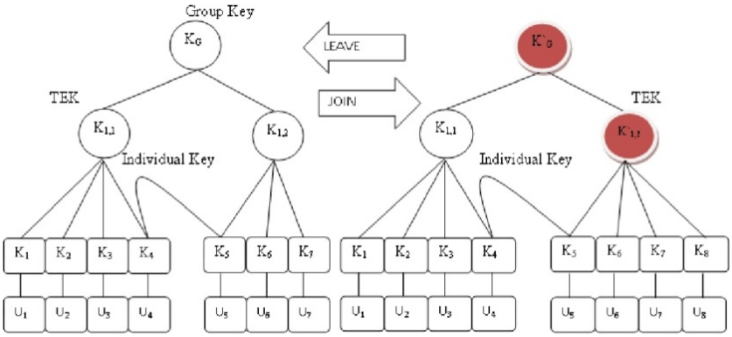


Fig. 1. Logical Key Tree for Key Graph (a) 4-Ary Tree for Seven Members (b) 4-Ary Tree for Eight Members

2 Proposed Protocol

It is based on the idea of key graph that manages the whole group on the basis of logical 4-ary key tree or key tree is the extended version of binary tree. In this protocol, the authors have divided whole group in several subgroups and every subgroup organized in a logical key hierarchy as in 4-ary key tree which reduce the complexity for a member join or leave from O (m) to O (log4n/m). The members in each subgroup contribute with each other to generate the subgroup key. This process delegates the key update process at a leave Process from the key server side to the member side. The proposed protocol works in a hierarchy of two levels of controllers;

the first for the group controller (GC) [3, 7] and the second is the subgroup controller (SC). The GC shares a symmetric key with all SCs which are trusted entities. The role of the SCs is to translate the data coming to their subgroups. Each SC works as the server of its subgroup. Figure 2: Illustrate the structure of proposed protocol.

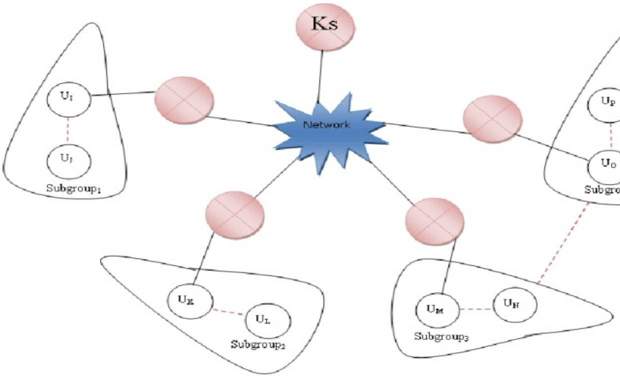


Fig. 2. Network Structure of Our Approach

The main objective of this protocol is to management a symmetric key between all group members in order to preserve the security of group communication [10]. In case of dynamicity occurs in the group membership by joining or leaving the group, the group key should be updated to maintain backward secrecy and forward secrecy. The structure of the subgroup hierarchy in the proposed protocol is shown in figure 6. The subgroup is organized in a hierarchy like the LKH approach [18] and KS is the key of the group key. For the Process of the proposed protocol are following:

i. In this approach key server is a trusted entity which responsible for generate required keys and for distributing those keys to valid group members as shown in the figure 6 and Every member of the group has IGMP membership [1, 2, 16], when a new member joins a group; it sends an IGMP membership report message to its neighboring router to have the multicast data delivered from a multicast sender. Other side, the member sends a join request message to the key server to obtain the group key by which the multicast data is encrypted. This is different from other LKH approaches, in term to handle a large number of members efficiently; our approach divides group members into subgroups. For example 256 members are divided into 16 subgroups as shown in figure 3.

ii. Our approach applies the concept of key tree in LKH to the subgroups. In the logical key tree, leaf nodes correspond to subgroups, not individual members. Similar to other LKH approaches, the root node corresponds to the group key, and the intermediate nodes correspond to traffic encryption key (TEK) used for key transfers.

iii. The division of group members into subgroups is performed so that a balanced tree is constructed. In this case, by dividing n (256) members into subgroups whose size is m (16) members, we will have $\lceil n/m \rceil$ subgroups, and the height of the tree will be $\log_4 \lceil n/m \rceil$. For example the group divided into 16 subgroups from 1 to 16 subgroups and height of the tree is 3 as shown in figure 3.

iv. When a member joins a group, it is allocated to a subgroup. At this time, the member obtains the following three kinds of key information from the key server.

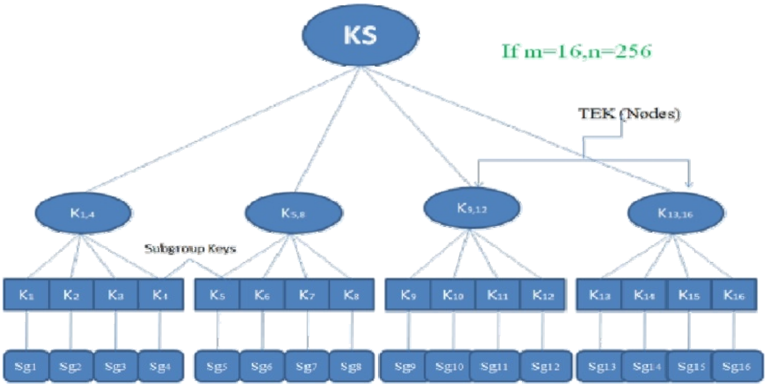


Fig. 3. Logical Key Tree in Proposed Approach

2.1 Design Methodology

Following steps have been involved in the designing stage.

2.1.1 Key Generation

As mentioned above, we use the modular exponential function as a one-way function. Since p is a large prime and g is the primitive element of multiplicative group $\boldsymbol{Z_p^*}$ it is computationally difficult to determine α given g and gα (mod p). Based on this property, the subgroup keys, the node keys and the group key are organized as follows.

First of all, the member secret α_j^i is selected under the condition that $2 \leq \alpha_j^i \leq p - 2$ and gcd (α_j^s,p-1)=1.

The server secret the server secret α_j^s is selected under the condition that is selected under the condition that $2 \leq \alpha_j^s \leq p - 2$.

Using those secrets, the subgroup key for subgroup j is calculated by Kj ≡ $g\alpha_j^1 \alpha_j^2 \ldots\ldots\ldots\ldots \alpha_j^m \alpha_j^s$(mod p). The node keys and the group key are organized by multiplying the exponents [20, 24] of its two child node keys (or the subgroup keys) in the logical key tree.

In order to illustrate the algorithm for re-keying, we use a simple example of multicast group divided into 16 subgroups; subgroup 1 to 16 with m members and subgroup 16 with m -1 members respectively. Figure.3 depicts the logical key tree for this group. The members of subgroups 1,2,3,4 own subgroup keys K1, K2, K3 and K4 respectively, node key K1, 4. The members of subgroups 5,6,7,8 own subgroup keys K5, K6, K7 and K8 respectively, node key K5, 8. The members of subgroups 9,10,11,12 own subgroup keys K9, K10, K11 and K12 respectively, node key K9, 12. The members of subgroups 13,14,15,16 own subgroup keys K13, K14, K15 and K16 respectively, node key K13,16 and group key KG. In this process key server used pre-computational function (PK) for calculating key when member join or leave the group and by using this pre-computational function process, we have minimized the computational cost during key generation and the keys are calculated as follows:

$\mathrm{K1} \equiv g \propto_1^1 \,.....\, \propto_1^{m-1} \propto_1^{m} \propto_1^{!s}$ (mod p)
$\mathrm{K2} \equiv g \propto_2^1 \,.....\, \propto_2^{m-1} \propto_2^{m} \propto_2^{!s}$ (mod p)
$\mathrm{K3} \equiv g \propto_3^1 \,.....\, \propto_3^{m-1} \propto_3^{m} \propto_3^{!s}$ (mod p)
$\mathrm{K4} \equiv g \propto_4^1 \,.....\, \propto_4^{m-1} \propto_4^{m} \propto_4^{!s}$ (mod p)
$\mathrm{K5} \equiv g \propto_5^1 \,.....\, \propto_5^{m-1} \propto_5^{m} \propto_5^{!s}$ (mod p)
$\mathrm{K6} \equiv g \propto_6^1 \,.....\, \propto_6^{m-1} \propto_6^{m} \propto_6^{!s}$ (mod p)
$\mathrm{K7} \equiv g \propto_7^1 \,.....\, \propto_7^{m-1} \propto_7^{m} \propto_7^{!s}$ (mod p)
$\mathrm{K8} \equiv g \propto_8^1 \,.....\, \propto_8^{m-1} \propto_8^{m} \propto_8^{!s}$ (mod p)
$\mathrm{K9} \equiv g \propto_9^1 \,.....\, \propto_9^{m-1} \propto_9^{m} \propto_9^{!s}$ (mod p)
$\mathrm{K10} \equiv g \propto_{10}^1 \,.....\, \propto_{10}^{m-1} \propto_{10}^{m} \propto_{10}^{!s}$ (mod p)
$\mathrm{K11} \equiv g \propto_{11}^1 \,.....\, \propto_{11}^{m-1} \propto_{11}^{m} \propto_{11}^{!s}$ (mod p)
$\mathrm{K12} \equiv g \propto_{12}^1 \,.....\, \propto_{12}^{m-1} \propto_{12}^{m} \propto_{12}^{!s}$ (mod p)
$\mathrm{K13} \equiv g \propto_{13}^1 \,.....\, \propto_{13}^{m-1} \propto_{13}^{m} \propto_{13}^{!s}$ (mod p)
$\mathrm{K14} \equiv g \propto_{14}^1 \,.....\, \propto_{14}^{m-1} \propto_{14}^{m} \propto_{14}^{!s}$ (mod p)
$\mathrm{K15} \equiv g \propto_{15}^1 \,.....\, \propto_{15}^{m-1} \propto_{15}^{m} \propto_{15}^{!s}$ (mod p)
$\mathrm{K16} \equiv g \propto_{16}^1 \,.....\, \propto_{16}^{m-1} \propto_{16}^{!s}$ (mod p)
$\mathrm{K1,4} \equiv g(\pi_i \alpha_1)(\pi_i \alpha_2)(\pi_i \alpha_3)(\pi_i \alpha_4)$ (mod p)
$\mathrm{K5,8} \equiv g(\pi_i \alpha_5)(\pi_i \alpha_6)(\pi_i \alpha_7)(\pi_i \alpha_8)$ (mod p)
$\mathrm{K9,12} \equiv g(\pi_i \alpha_9)(\pi_i \alpha_{10})(\pi_i \alpha_{11})(\pi_i \alpha_{12})$ (mod p)
$\mathrm{KG} \equiv (\mathrm{PK}_{13,16}\mathrm{PK}_{9,12}\mathrm{PK}_{5,8})^{\alpha_{16}^1 \cdots\cdots \alpha_{16}^{m-1}} \alpha_{16}^{!s}$ (mod p)

2.1.2 Join Process

We now use to explain how re-keying is done when a new member joins the multicast group. In this process key server used pre-computational function for calculating key when member join or leave the group and this pre-computational function process minimized the computational cost during key generation. The procedure is as follows:

i. When key server receives a join request, it authenticates the member [5]. This may be done by the conventional approach such as remote authentication dial in user service (RADIUS) [4, 21], and we do not discuss this procedure. If required, the key server assigns the session key, and sends it to the member.

ii. The key server determines the subgroup for the new member and assigns the identity within the subgroup. In this example, the new member belongs to subgroup 16 and its identity is m. At this time, the path set for subgroup16, the keys K13, K14, K15, K16 and KG need to be changed to new ones.

iii. The key server assigns member secret α_{16}^{m} to u_{16}^{m}, and calculates its inverse value α_{16}^{-m} as well.

iv. The key server changes the server secret assigned to subgroup 16 from α_{16}^{s} to $\alpha_{16}^{!s}$.

v. The key server updates K16, K13,16 and KG to K`16, K`16 and K`G using α_{16}^{m} and $\alpha_{16}^{!s}$ in the following way.

$$K`16 \equiv g \propto_{16}^{1} \ldots\ldots \propto_{16}^{m-1} \propto_{16}^{m} \propto_{16}^{!s} \qquad \text{(mod p)}$$
$$K`13,16 \equiv g(\pi_i\alpha_{13})(\pi_i\alpha_{14})(\pi_i\alpha_{15})(\pi_i\alpha_{16}) \qquad \text{(mod p)}$$
$$K`G \equiv (PK_{13,16}PK_{9,12}PK_{5,8})^{\alpha_{16}^{1} \ldots\ldots\ldots\ldots \alpha_{16}^{m-1}} \propto_{16}^{m} \alpha_{16}^{!s} \qquad \text{(mod p)}$$

vi. The key server encrypts {K`16, K`13,16, K`G},and the inverse values of the other members in that subgroup, $\alpha_{16}^{-1} \ldots\ldots\ldots \alpha_{16}^{-m-1}$ by Kp_{16}^{m} than it sends this encrypted message through unicast to $\propto_{16}^{m}$. It has been given below:

$S\xrightarrow{\text{Unicat}}\{\propto_{16}^{m}\}$:{(K`16, K`13,16, K`G, $\alpha_{16}^{-1} \ldots\ldots\ldots \alpha_{16}^{-m-1}$)$Kp_{16}^{m}$}.

vii. Server encrypts α_{16}^{-m}, and K`16 by K16for subgroup 16, K`13,16 by K13,16 for subgroup 13,14,15, K`G by KG for subgroup 1 to 12,and distributes these encrypted keys through multicast for existing members. This process describe as following:

$S\xrightarrow{\text{Multicast}}$ {Existing Members}

{(α_{16}^{-m}, K`16) K16,( K`13,16) K13,16 ,(K`G) KG}.

In this process, each updated key is encrypted by the previous one for existing members, and as a result, only the members who know the corresponding previous keys can decrypt the encrypted message containing the new keys.

2.1.3 Leave Process

When user u_{16}^{m} leaves the group then all member of the group affected by this change and key server changes the group key or path key such as K`16 to K``16, K13,16 to K``13,16 and KG to K``G. According to our protocol, these updated keys do not need to be sent to the remaining members. Instead, the key server just prepares one message for subgroup 16 indicating u_{16}^{m} leaves and delivers α_{16}^{-m} for subgroup 1 to 15.

The value of α_{16}^{-m} is encrypted into multiple copies by K15 and K13,16, for subgroup 15 and 1 to 14 respectively. The key server sends this message through multicast. This process describe as following:

$S\xrightarrow{\textit{Multicast}}$ {Remaining members}

:{(α_{16}^{-m}) K15, (α_{16}^{-m}) K13,16}.

When the remaining members receive this message, they decrypt it by the corresponding keys and then Use u_{16}^{m} to update those keys.

$$K``16 \equiv (K`16)\ \alpha_{16}^{-m} \qquad (\mathrm{mod}\ p)$$

$$\equiv g\ \alpha_{16}^{1} \ldots\ldots \alpha_{16}^{m}\alpha_{16}^{-m}\alpha_{16}^{16} \qquad (\mathrm{mod}\ p)$$

$$K``_{16} \equiv g\ \alpha_{16}^{1} \ldots\ldots \alpha_{16}^{m-1}\alpha_{16}^{16} \qquad (\mathrm{mod}\ p)$$

$$K``_{13,16} \equiv (K`_{16})\ \alpha_{16}^{-m} \qquad (\mathrm{mod}\ p)$$

$$\equiv g(\pi_i\alpha_{13})(\pi_i\alpha_{14})(\pi_i\alpha_{15})(\pi_i\alpha_{16}) \qquad (\mathrm{mod}\ p)$$

$$\equiv g(\pi_i\alpha_{13})(\pi_i\alpha_{14})(\pi_i\alpha_{15})(\pi_{i-1}\alpha_{16}) \qquad (\mathrm{mod}\ p)$$

$$K``_G \equiv (K`_G)\ \alpha_{16}^{-m} \qquad (\mathrm{mod}\ p)$$

$$\equiv (PK_{13,16}PK_{9,12}PK_{5,8})^{\alpha_{16}^{1}\ldots\ldots\ldots\ldots\alpha_{16}^{-m}}\alpha_{16}^{m}\alpha_{16}^{16} \qquad (\mathrm{mod}\ p)$$

$$\equiv (PK_{13,16}PK_{9,12}PK_{5,8})^{\alpha_{16}^{1}\ldots\ldots\ldots\ldots\alpha_{16}^{-m}}\alpha_{16}^{16} \qquad (\mathrm{mod}\ p)$$

As we notice, the key server does not need to generate new keys (TEK and group key) after a leave. Instead, it just sends the inverse value of leaving member to remaining members. Then, the remaining members update the necessary keys. In this way, updating the keys after a leave is shifted to member's side which improves the efficiency of re-keying at leave.

3 Comparison with Exist One

In the comparison, n denotes the number of members in the group, the number after a join and before a leave in a strict numbers, called group size. We define m as the number of members in a subgroup, called subgroup size only for our proposal. We also show some numerical results for the overhead by changing the group sizes from 16 to 1048576. To evaluate our proposal, we use subgroup size is i.e. 256.

We are using a binary tree for LKH and OFT, and our proposal is based on 4-ary key tree. In case of simple app. height of the tree h=2 and height LKH=OFT=$\log_2$n. The height of the key tree for our proposal under the condition of n ≤ m will be equal to 1 and under the condition of n>2m will be log4 $\lceil n/m \rceil$. In general, the number of node keys is proportional to the height of the key tree in LKH based protocols, in other words, a protocol with smaller height has fewer nodes along the path. In our approach we have minimized the height of the tree along the key path and number of key generation, encryption/decryption also minimized. Due to this reason the performance of the system will be improve.

3.1.1 Computational Overhead

Computational overhead depending on the Key generation overhead and encryption/decryption overhead as following:

3.1.1.1. Key Generation Overhead. It is the overhead at the key server and member node along the path to the root at each join or leave. The number of key generations at the key server is almost equal to the height of the key tree. First of all, Simple App. has the smallest overhead at the key server both join and leave process. Our approach minimizes number of key generation at the key server both join and leave process as compare to LKH and OFT. By contrast, because of smaller size of h_{sg}, the key server generates fewer keys at join. Most importantly, the key server does not need to generate new keys for the members at leave.

On the other hand in simple application and LKH, a member node does not generate any keys by it at each event, but in OFT the new member at join and a remaining member node along the path at leave need generate new node keys by mixing two hash values. At a member leave process the group and subgroup controller doesn't generate any keys. Instead it multicasts the identity of the leaving member to all the group and subgroup members to be factored from the subgroup key by using the leaving member's inverse value. Figure 4(a) and 4(b) shows comparative result of number of key generation overhead on the basis of group size and number of key generated at the key server.

From the figure 4(a) one can notice that the proposed protocol has minimized overhead at the join process because the key server reduced the height of key tree by using 4-ary key tree. From the figure 4(b) one can notice that the proposed protocol has the smallest overhead at the leave process because the key server doesn't generate any keys in that case.

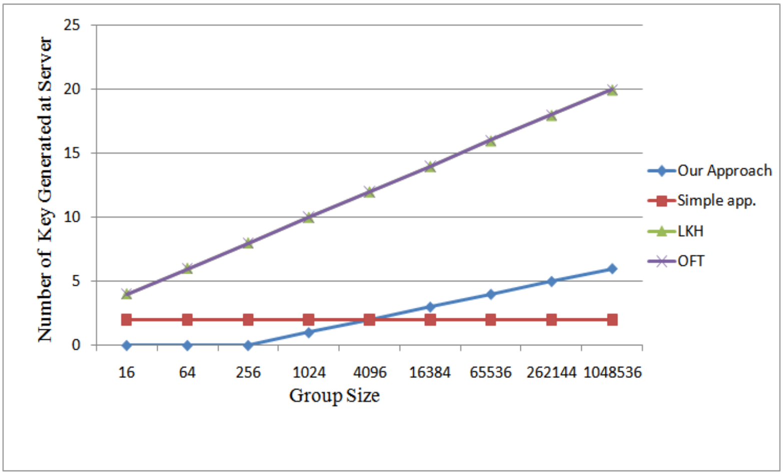


Fig. 4. (a): Key Generation Overhead at the Key Server during Join Process

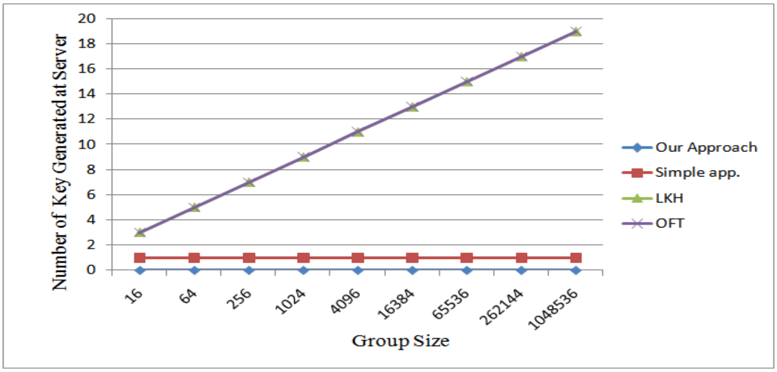


Fig. 4. (b): Key Generation Overhead at the Key Server during Leave Process

Figure 5(a) and 5(b) shows comparative result of number of key generation overhead on the basis of group size and number of key generated at the member node.

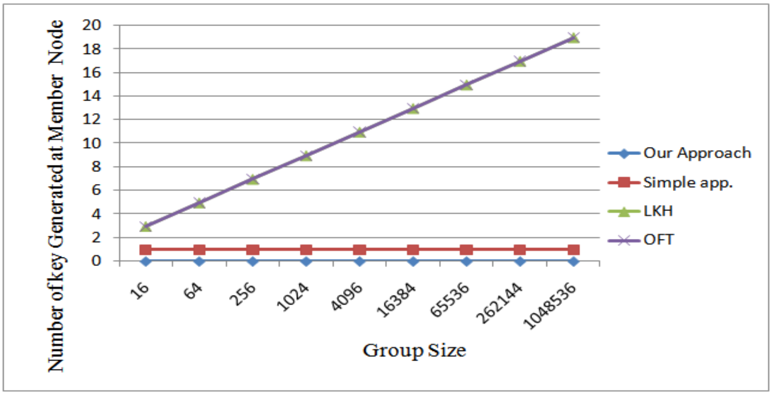


Fig. 5. (a): Key Generation Overhead at the Member Node during Join Process

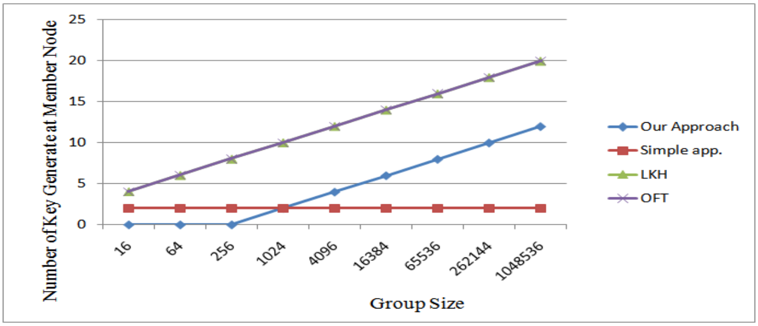


Fig. 5. (b): Key Generation Overhead at the Member Node during Leave Process

The key generation overhead for our protocol is 0 at join, but proportional to the height of the key tree at leave as shown in the table 1 at member node. In fact, a member node renews the node keys along the path to the root by modular exponentiation.

3.1.1.2. Encryption/Decryption Overhead. Encryption overhead at the key server (left side) and decryption overhead at a member node (right side) at each join and leave. The values for join process at the key server are the sums of the number of encryptions for existing members and a new member. Although the overhead at the key server for Simple App. is the smallest value at join, it is the largest one of all at leave. In simple App. the key server has to perform n-1encryptions for the remaining members at leave, when n are the number of members. This is the problem we mentioned in Section 4; in which other protocols have tried to solve this problem by introducing the 4-ary key tree.

At the key server, LKH involves two separate encryptions per node, one for each of its two children, compared to OFT which involves one encryption per node. Therefore, even with the same height of LKH = height of OFT, the encryption overhead for LKH is larger than of that for OFT. By contrast, because of the small height of subgroup (hsg) and small height of group hg, the key server performs fewer encryptions at join and leave for our protocol compared with LKH, OFT. The encryption /decryption formula for proposed approach and previous approaches as shown in the table 3, according to our approach number of encryption will be minimize at the key sever at the time of both join and leave process. On the other hand number of decryption at the member node also minimized at the joining time as compared to LKH and OFT.

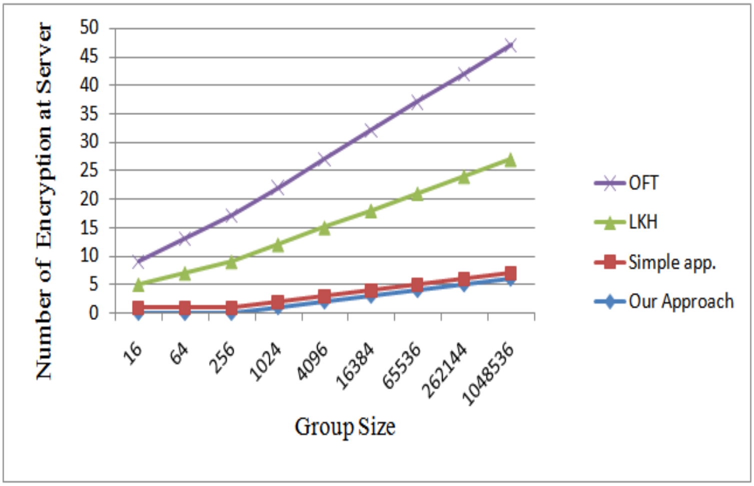


Fig. 6. (a): Number of Encryption at the Key Server during Join Process

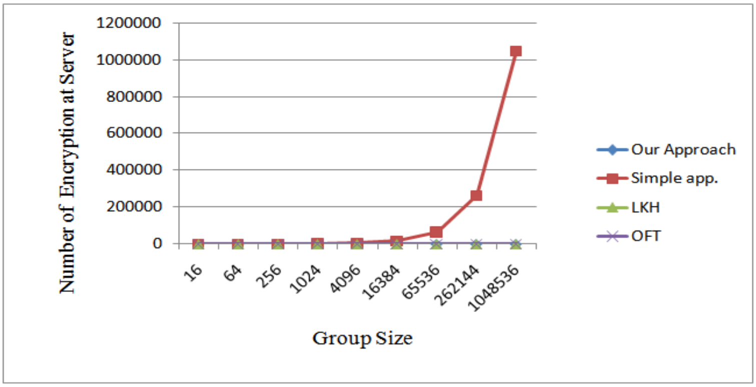


Fig. 6. (b): Number of Encryption at the Key Server during Leave Process

Figure 6 (a) and 6(b) show the number of encryption at the key server at join and leave process respectively. At a member node, the decryption overhead at join is proportional to height of the key tree for all protocols, in which simple app. has smallest overhead, but LKH and OFT have the largest overhead. Because of height of LKH and OFT is largest as compare to propose approach. So, that proposed approach minimize the number of decryption at the member node.

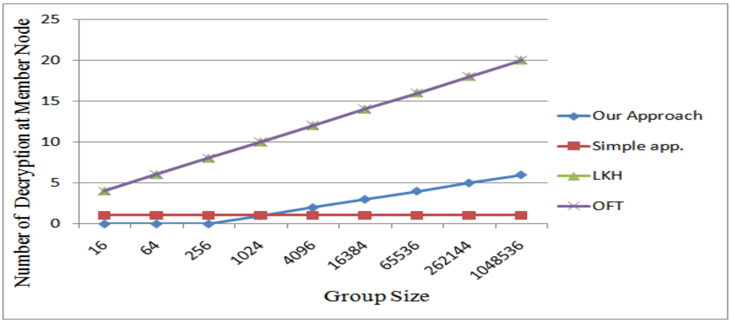


Fig. 7. (a): Number of Decryption at the Member Node during Join Process

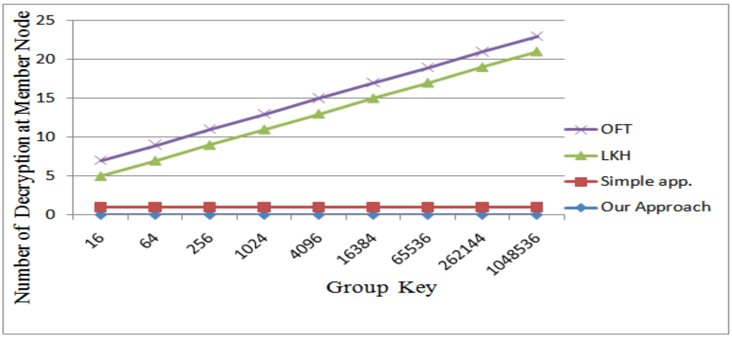


Fig. 7. (b): Number of Decryption at the Member Node during Leave Process

Figure 7 (a) and 7(b) show the number of encryption at the member node at join and leave process respectively.

3.1.1.3. Communication Overhead. Communication overhead at join and leave for multicast communication as shown in table 4. As described above, this is measured by the number of transmitted control messages. Figure 8(a) and 8(b) illustrate numerical results at join and leave, respectively. At join, Simple App., and our proposal have a small overhead. LKH has the largest overhead and OFT has half of that. On the other hand, at leave, Simple App. has an extremely large overhead and our protocol have a small overhead.

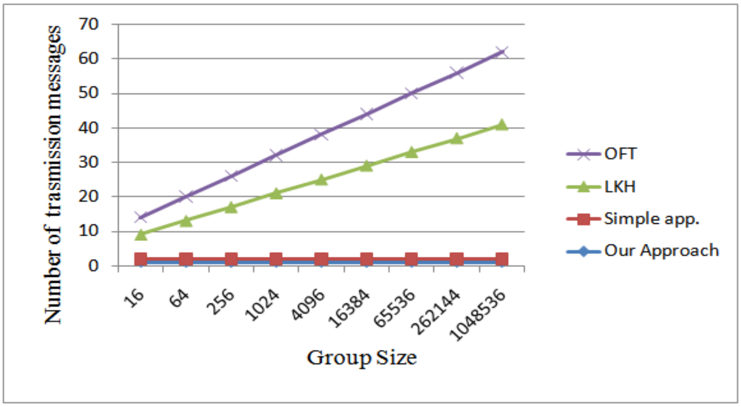


Fig. 8. (a): Communication Overhead at Server during Join Process

On the basis of comparative results our protocol the best in terms of the communication overhead. Because of this is send all key information in one multicast message to existing members at join, and to remaining members at leave. It should be noticed that the message size is different as shown later; it is bigger for LKH than for our protocol.

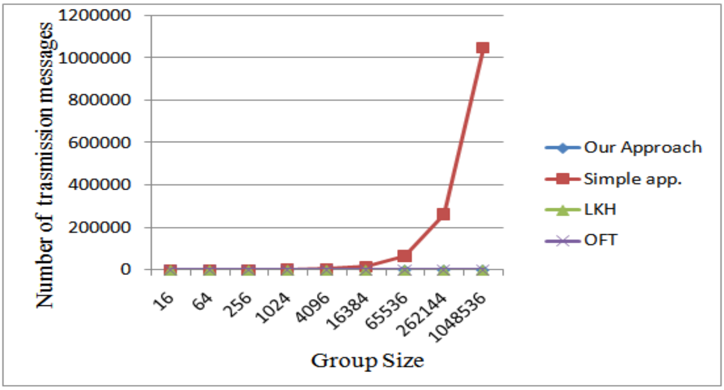


Fig. 8. (b): Communication Overhead at Server during Leave Process

4 Conclusion

We have discussed different type of security in the multicast such as multicast receiver access control, multicast source authentication, multicast fingerprinting and group key management. We have selected group key management area of multicast security and we can say group key management is the part of multicast security. We have found different type of issues such as computational overhead, communicational overhead, message size and storage overhead in the group key management and many researchers provided LKH, OFT etc. solutions for these issues but these issues are not solved. Therefore, in this dissertation, we have proposed a security improvement in group key management approach to solve the problem of distributing a symmetric key between the whole group members for secure group communication. The performance of the proposed protocol is compared with that of the Simple App., OFT and LKH protocols. The comparison is undertaken according to the computational overhead, communication overhead, storage overhead, and message size. The results show that the proposed protocol improves the group performance in terms of computation overhead, message size and communication overhead.

References

[1] Je, D.-H., Lee, J.-S., Park, Y., Seo, S.-W.: Computation-and-storage-efficient key tree management protocol for secure multicast communications. Computer Communications 33(2), 136–148 (2010)

[2] Pour, A.N., Kumekawa, K., Kato, T., Itoh, S.: A hierarchical group key management scheme for secure multicast increasing efficiency of key distribution in leave operation. Elsevier, Computer Networks 51(17), 4727–4743 (2007)

[3] Chen, X., Ma, B.N.W., Yang, C.: M-CLIQUES: Modified CLIQUES key agreement for secure multicast. Computers & Security 26(3), 238–245 (2007)

[4] Sun, Y(L.), Ray Liu, K.J.: Hierarchical Group Access Control for Secure Multicast Communications. IEEE/ACM Transactions on Networking 15(6) (December 2007)

[5] Lee, P.P.C., Lui, J.C.S., Yau, D.K.Y.: Distributed Collaborative Key Agreement and Authentication Protocols for Dynamic Peer Groups. IEEE/ACM Transactions on Networking 14(2) (April 2006)

[6] Abdellatif, R., Aslan, H.K., Elramly, S.H.: New Real Time Multicast Authentication Protocol. International Journal of Network Security 12(1), 13–20 (2011)

[7] Zheng, S., Manz, D., Alves-Foss, J.: A communication computation efficient group key algorithm for large and dynamic groups. Elsevier, Computer Networks 51(1), 69–93 (2007)

[8] Wallner, D., Harder, E., Agee, R.: Key Management for Multicast: Issues and Architecture. National Security Agency (June 1999), RFC 2627

[9] Saroit, I.A., El-Zoghdy, S.F., Matar, M.: A Scalable and Distributed Security Protocol for Multicast Communications. International Journal of Network Security 12(2), 61–74 (2011)

[10] Baugher, M., Canetti, R., Dondeti, L., Lindholm, F.: Multicast Security (MSEC) Group Key Management Architecture. RFC 4046 (April 2005)

[11] Challal, Y., Seba, H.: Group Key Management Protocols: A Novel Taxonomy. International Journal of Information Technology 2(1) (2005) Issn: 1305-2403

[12] Wong, C.K., Gouda, M., Lam, S.S.: Secure Group Communications Using Key Graphs. IEEE/ACM Transactions on Networking 8(1) (February 2000)

[13] Jabeenbegum, S., Purusothaman, T., Karthi, M., Balachandar, N., Arunkumar, N.: An Effective Key Computation Protocol for Secure Group Communication in Heterogeneous Networks. IJCSNS International Journal of Computer Science and Network Security 10(2) (February 2010)

[14] Srinivasan, R., Vaidehi, V., Rajaraman, R., Kanagaraj, S., Chidambaram Kalimuthu, R., Dharmaraj, R.: Secure Group Key Management Scheme for Multicast Networks. International Journal of Network Security 11(1), 33–38 (2010)

[15] Ng, W.H.D., Howarth, M., Sun, Z., Cruickshank, H.: Dynamic Balanced Key Tree Management for Secure Multicast Communications. IEEE Transactions on Computers 56, 590–605 (2007)

[16] Lu, H.: A Novel High-Order Tree for Secure Multicast Key Management. IEEE Transactions on Computers 54, 214–224 (2005)

[17] Wong, C.K., Gouda, M., Lam, S.S.: Secure group communications using key graphs. IEEE/ACM Transactions on Networking 8(1), 16–30 (2000)

[18] Wallner, D., Harder, E., Agee, R.: Key Management for Multicast: Issues and architectures. National Security Agency (June 1999), RFC 2627

[19] Dierks, T., Rescorla, E.: The Transport Layer Security (TLS). Protocol Version 1.1 (April 2006), RFC 2346

[20] Stinson, D.R.: Cryptography Theory and Practice”, Second edition. Chapman and Hall/CRC Press, 155–175 (2002)

[21] Rigney, C., Willens, S., Rubens, A., Simpson, W.: Remote Authentication Dial in User Service (RADIUS) (June 2000), RFC 2865

[22] Deering, S.: Host Extensions for IP Multicasting. RFC 1112 (August 1989)

[23] Tanenbaum, A.: Computer Networks, 4th edn. Prentice Hall (2009)

[24] Stallings, W.: Cryptography and Network Security Principles and Practices, 4th edn., p. 592 (November 16, 2005)

Detailed Dominant Approach Cloud Computing Integration with WSN

Niranjan Lal[1], Shamimul Qamar[2],
and Mayank Singh[3]

[1] MODY Institute of Technology and Science,
Laxmangarh, Sikar (Raj.) –India
niranjan_verma51@yahoo.com
[2] Noida Institute of Engineering and Technology,
Greater Noida (UP) - India
drsqamar@rediffmail.com
[3] THDC Institute of Hydropower Engineering & Technology Tehri (UK) -India
mayanksingh2005@gmail.com

Abstract. The maximum benefit out of the recent developments in sensor networking can be achieved via the integration of sensors with Internet. The real-time specific sensor data must be processed and the action must be taken instantaneously. This distributed architecture has numerous similarities with the wireless sensor networks (WSN) where lots of motes, which are responsible for sensing and preprocessing, are connected with wireless connection in the real-time. Since wireless sensor networks are limited in their processing power, battery life, communication speed and storage resources , cloud computing offers the opposite , which makes it fetching for endless observations, analysis and use in different sort of environment.

In this paper we proposed an architecture, which integrates the Cloud computing technology with the wireless sensor network. In this paper we also discussed some research challenges with respect to cloud computing and wireless sensor networks, and important key component of sensor cloud

Keywords: Cloud computing, Distributing computing, Wireless sensor networks , Sensor cloud, Research challenges of cloud computing and Internet.

1 Introduction

Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications. It allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access, with more efficient computing by centralizing storage, memory, processing and bandwidth. It can be securing immense amounts of data which is only accessible by authorized users. Cloud computing in broad way is shows in Figure 1.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 894–903, 2013.

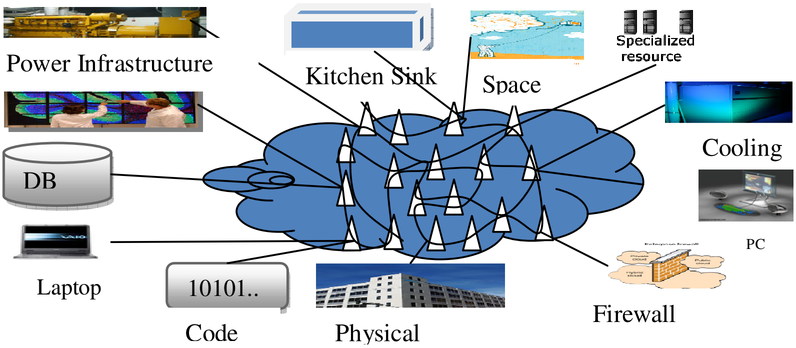


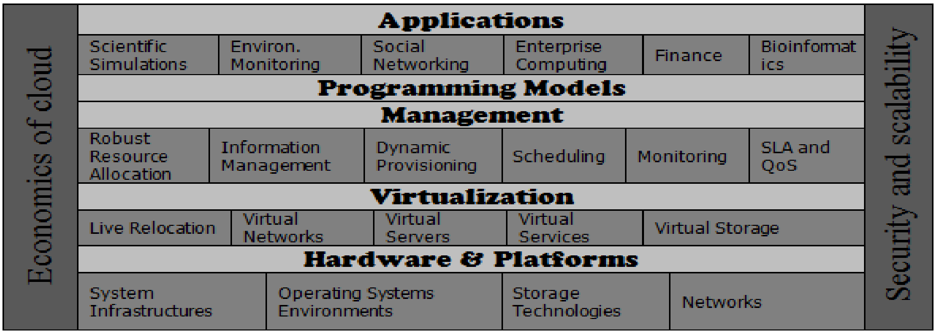
Fig. 1. Cloud Computing

Cloud computing is the technology that enables functionality of an IT infrastructure, IT platform or an IT product to be exposed as a set of services in a seamlessly scalable model so that the consumers of these services can use what they really want and pay for only those services that they use (Pay per use) [2].

Many formal definitions have been proposed in both academia and industry, the one provided by U.S. NIST (National Institute of Standards and Technology) [16] :

“Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [16]. “

Wireless sensor network consists of a large number of such sensor nodes that are able to collect and disseminate data in areas where ordinary networks are unsuitable for environmental and/or strategic reasons. Each sensor node comprises sensing, processing, transmission, mobilize, position finding system (Such as GPS) and power units [8].



Economics of cloud

Applications

Scientific Simulations	Environ. Monitoring	Social Networking	Enterprise Computing	Finance	Bioinformatics

Programming Models

Management

Robust Resource Allocation	Information Management	Dynamic Provisioning	Scheduling	Monitoring	SLA and QoS

Virtualization

Live Relocation	Virtual Networks	Virtual Servers	Virtual Services	Virtual Storage

Hardware & Platforms

System Infrastructures	Operating Systems Environments	Storage Technologies	Networks

Security and scalability

Fig. 2. Architectural Level cloud computing

The system architecture of wireless sensor network is shown in Figure 3. In other ways Wireless sensor a network is seamlessly couples the physical environment with the digital world. Sensor nodes are small, low power, low cost, and provide multiple functionalities sensing capability, processing power, memory, communication, bandwidth, battery power. Useful in many application domains.

The organization of our paper is as follows. In section 2, section 3 and section 4 we have discussed key features of our interest and Limitations of cloud computing and sensor networks. In section 5 describe need to integrate cloud computing with wireless sensor networks? In section 6 we describe research challenges, where some research works can be done in these areas. In section 7 we present a proposed architecture of cloud computing with WSN. In section 8 point out some key components of the proposed architecture and Section 9 conclude and future work.

2 Key Features of Our Interest

There are some key features that are useful for everyone to use internet. i) Immense computational and storage resources that are collocate. ii) Very high speed data processing and movement. iii) Accessibility over the Internet Service-Oriented Architecture and virtually from any platform.

3 Limitation of Cloud Computing

Cloud computing is limited – as of now: i) The immense power of the Cloud can only be fully exploited if it is seamlessly integrated into our physical lives. ii) It is providing the real world's information to the Cloud in real time and getting the Cloud to act and serve us instantly so it need to add the sensing capability to the Cloud.

4 Limitation of Sensor Networks

Sensor networks are limited too: i) It is very challenging to scale sensor networks to large sizes with proprietary vendor-specific designs, which is difficult for different Sensor networks to be interconnected. ii) We know sensor networks is operate in separate silos, so sensor data cannot be easily shared by different groups of users. iii) Sensor network is used for fixed and specific applications that cannot be easily changed once deployed, due to this it slow adoption of large-scale sensor network applications.

5 The Missing Piece

The missing piece of cloud computing is shown in the Figure 3. In which cloud computing is would be integrate with sensor networks.

5.1 A Scenario

A scenario is a description of a flow of messages in the network via cell phone shown in the Figure 4. An Insight into the Scenario has some steps shown on next page .

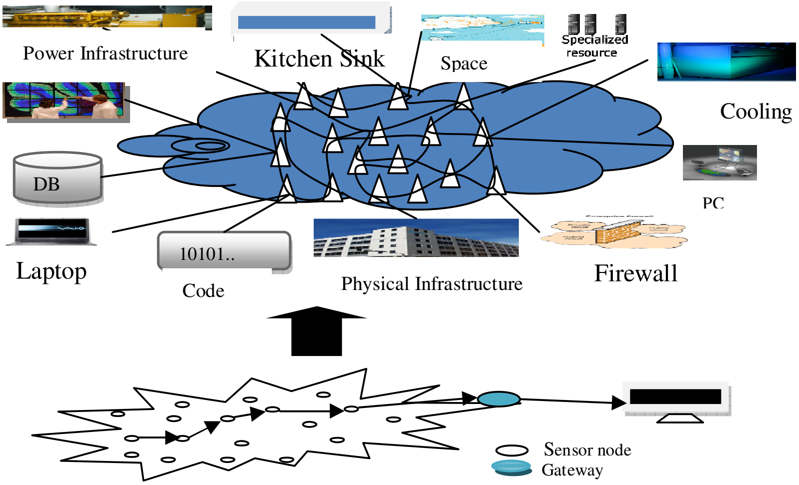


Fig. 3. The missing piece

Step 1. Cell phone records the tourist's gestures and activates applications such as camera, microphone, etc.
Step 2. The cell phone produces very swift responses in real time after: i) Processing geographical data. ii) Acquiring tourist's physiological data from wearable physiological sensors (blood sugar, precipitation, etc) and cross-comparing it with his medical records. iii) Speech recognition. iv) Image processing of restaurant's logos and accessing their internet-based profiles. v) Accessing tourist's social network profiles to find out his friends.
Step 3. Fact: the cell phone cannot perform so many tasks!

5.2 Need to Integrate Cloud with Sensors

These are the some point why need to integrate cloud with sensors : i) Acquisition of data feeds from numerous body area (blood sugar, heat, perspiration, etc) and wide area (water quality, weather monitoring, etc) sensor networks in real time. ii) Real-time processing of heterogeneous data sources in order to make critical decisions. iii) Automatic formation of workflows and invocation of services on the cloud one after another to carry out complex tasks.

5.3 The Sensors Cloud

"An infrastructure that allows truly pervasive computation using sensor as interface between physical and cyber world, the data compute cluster as the cyber backbone and the internet as the communication medium."

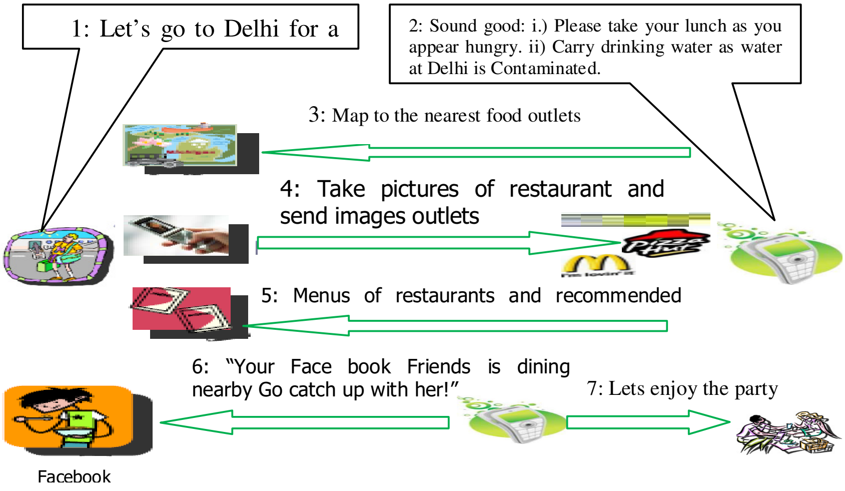

1: Let's go to Delhi for a
2: Sound good: i.) Please take your lunch as you appear hungry. ii) Carry drinking water as water at Delhi is Contaminated.
3: Map to the nearest food outlets
4: Take pictures of restaurant and send images outlets
5: Menus of restaurants and recommended
6: "Your Face book Friends is dining nearby Go catch up with her!"
7: Lets enjoy the party
Facebook

Fig. 4. A scenario

Sensor cloud integrates large-scale sensor networks with sensing applications and cloud computing Infrastructures , which collects and processes data from various sensor networks, that will enables large-scale data sharing and collaborations among users and applications on the cloud, then it will delivers cloud services via sensor-rich mobile devices, which allows cross-disciplinary applications that span organizational boundaries, enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications. It also Supports complete sensor data life cycle from data collection to the backend decision support system.

6 Research Challenges

Research challenges in which areas research works can be done: i) Complex Event Processing and Management. ii) Massive Scale and Real Time Data Processing. iii) Large Scale Computing Frameworks. iv) Harvesting Collective Intelligence

7 The Proposed Architecture of the Cloud Computing with Sensor Networks

The architecture of the cloud computing with sensor network is shown in the Figure 5 [1], this architecture Enables users to easily collect, access, process, visualize, archive, share and search large amounts of sensor data from different applications. Supports complete sensor data life cycle from data collection to the backend decision support system. Vast amount of sensor data can be processed, analyzed, and stored using computational and storage resources [3] of the cloud.

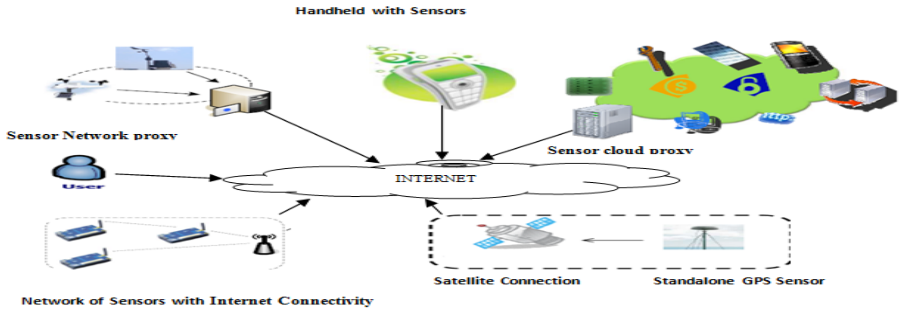


Fig. 5. Sensor Cloud Architecture

7.1 Requirements for the System

Cloud computing model is mainly based on pipes and filters [17] .The pipes and filter design (see Figure 6) is used in digital processing applications, which also used in wireless sensor networks [18].

Pipes are used to buffer data and provide uniform interconnection mechanism of filters. Filters process and transform input data and deliver it to an output port.

The general system architecture, which integrates cloud computing with wireless sensor networks, contains several basic services.

The main requirements for the system are: i) Receive and manage sensor data from heterogeneous motes. ii) Manage a set or chain of filters that perform on-line analysis on sensor data. iii) Run filters offline on a given set of sensor data. iv) Permanently run filters on a given set of sensoric data. v) Provide different, user definable views and visualizations on the sensor data and calculation results. vi) Provide an interface for changing existing filters or to develop new filters out of an existing domain specific modeling tool. vii) Provide an interface for data export, so that the stored data can easily be taken from the Cloud storage and to be used in non-cloud solutions. viii) Provide a notification service, e.g. a filter or machine learning component identifies a specific situations or has finally calculated a specific result. ix) Provide data access rules x) Provide configuration capabilities for filter chains, (web) services, notifications, and data access rules. xi) Provide a management console for the configuration of the whole system.

7.2 Proposed System Level Architecture of Cloud Computing Integration with Wireless Sensor Network(WSN)

The system level architecture of cloud computing integration with wireless sensor network in Figure 8 shows the collection of base services provided to the user as SaaS

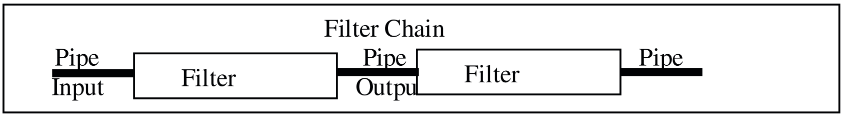


Fig. 6. Pipe, Filter, and Filter Chain

(Software as a Service), PaaS (Platform as a Service), and IaaS (Infrastructure as a Service) applications as cloud service providers. The Wireless Sensor Network Analytical Services and Cloud groups all services that are necessary to fulfill the requirements that are integrate cloud computing with wireless sensor networks. The following main services are necessary to collect and analyze sensor network data within the cloud: Necessary Services for wireless sensor network and Cloud

Global Sensor Data Management or Sensor Metadata Management or Cloud Based Management System: It is responsible for the management of the sensor data within the Cloud Computing environment. Since the cloud computing environment provide several ways for storing data, e.g.: Google AppEngine offers Bigtables [4,15] for data persistence and Microsoft Azure provides BLOBs, queues and tables, it's necessary to have a flexible data access layer, which raises the level of abstraction, so that persistence mechanism can be easily exchanged to the global sensor networks.

Provide Runtime for Filter Chains: Filters are usually configured for a specific filter chain. A filter chain is a reliable runtime environment for filters and it executes various user defined filters.

User Interface via Web Browser: A user interface provide the interaction between people (users) with a machine via Web Browser. The user interface includes hardware (physical) and software (logical) component.

Filter Chain and Management of the Filter: Filters and filter chains we need some management. This management service allows the administration of filters, so that a user is able to add new filters, delete filters and aggregate existing filters to combined filters. Since many filter chains are executed in parallel, it is necessary to offer a flexible configuration mechanism.

Visualization / Views for Data Analysis: The visualization service provides various predefined and user-defined views on the data and analysis results. The visualizations and views can be implemented with languages like data warehouses, OLAP, The spatial OLAP Visualization and Analysis Tool (SOVAT), NET Reporting and Heat map. So a powerful visualization is necessary to manage the sensor data.

Notification Service: This service is a mechanism to inform external applications and services about specific situations, where it fires an event. This could be for example, an indication that a gas pump failure is approaching

8 Key Components of the Proposed Architecture of the Cloud Computing with Sensor Networks

There are some key components of the cloud computing with sensor Networks as follows

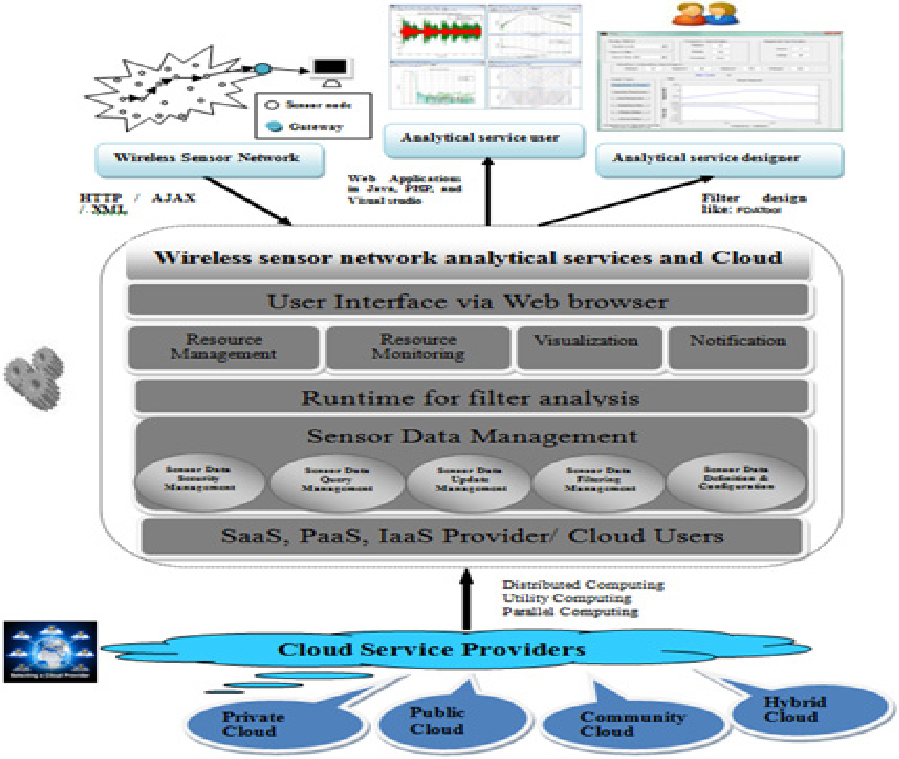

Sensor node
Gateway
Wireless Sensor Network
Analytical service user
Analytical service designer
HTTP / AJAX / XML
Web Applications in Java, PHP, and Visual studio
Filter design like: FDATool
Wireless sensor network analytical services and Cloud
User Interface via Web browser
Resource Management
Resource Monitoring
Visualization
Notification
Runtime for filter analysis
Sensor Data Management
Sensor Data Security Management
Sensor Data Query Management
Sensor Data Update Management
Sensor Data Filtering Management
Sensor Data Definition & Configuration
SaaS, PaaS, IaaS Provider/ Cloud Users
Distributed Computing
Utility Computing
Parallel Computing
Cloud Service Providers
Private Cloud
Public Cloud
Community Cloud
Hybrid Cloud

Fig. 7. Proposed system level architecture of cloud computing integration with wireless sensor networks

8.1 Sensor-Cloud Proxy

Sensor cloud proxy provide the interface between sensor resources and the cloud fabric, which Manages sensor network connectivity between the sensor resources and the cloud., that exposes sensor resources as cloud services, to manages sensor resources via indexing services. Cloud discovery services used for resource tracking. for manages sensing jobs for programmable sensor networks, manages data from sensor networks, data format conversion into standard formats (e.g. XML).

8.2 Sensor-Network Proxy

Sensor network proxy provides the connection to sensor resources that do not have direct connection to the cloud, sensor network is still managed from the Sensor-Cloud Interface via Sensor Network proxy, which collects data from the sensor network continuously or as and when requested by the cloud services to enhances the scalability of the Sensor Cloud, finally sensor cloud proxy provides various services for the underlying sensor resources, e.g. power management, security, availability, Quality of Services(QoS).

9 Conclusion and Future Work

The communication among sensor nodes using Internet is a challenging task since sensor nodes contain limited band width, memory and small size batteries. The issues of storage capacity may be overcome by widely used cloud computing technique with sensor networks. Cloud integration with WSN mechanism may provides dynamic collaboration between clouds to enable many services. We also conclude that the cell phone cannot perform so many tasks! There for cloud computing is useful with sensor networks. As we know "Sensor networks are distributed across extended terrain so they open up an entirely new scope of applications. Another critical feature of this technology is that it has a very light footprint, it can be installed using fairly non-intrusive methods, and as a result, we do not impact the environment we are trying to observe." Everyone should interest to discuss about the use of cloud computing for real-world applications, and explore opportunities for collaboration which lead to the intelligence integration into the Internet. This solution has been extended to sensor clouds, which leads to high availability and hence reliability is achieved.

Acknowledgment. We would like to acknowledge the support and input from our family, Mr. Munish Kumar (HOD IT, CDAC Noida,) –INDIA, Dr. J.V. Desai (Dean), Prof. Prema K.V. (HOD) CSE, and our colleagues from FET, MITS Laxmangarh, Sikar (Raj.) India, NIET Greater Noida Uttar Pradesh and THDC Institute of Hydropower Engineering & Technology Tehri Uttarakhand -India

References

1. Beng, L.H.: Sensor cloud: towards sensor-enabled cloud services. Intelligent systems center. Nanyang Technological University (April 13, 2009)
2. Introduction to Cloud Computing architecture White Paper on sun Microsystems, 1st edn. (June 2009)
3. Ulmer, C., Alkalai, L., Yalamanchili, S.: Wireless distributed sensor networks for in-situ exploration of mars. Work in progress for NASA Technical Report, `http://users.ece.gatech.edu/`
4. Chang, F., et al.: Bigtable: A Distributed Storage System for Structured Data. In: OSDI 2006: Seventh Symposium on Operating System Design and Implementation, Seattle, WA (2006)

5. Lal, N.: A novel survey on Cloud Computing Issues. International Journal of Computer Information Systems (IJCIS) 01(02), 18–21 (2010)
6. Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud Computing, UCB/EECS-2009-28. University of California, Berkeley (2009)
7. Zhao, F., Guibas, L.: Wireless Sensor Networks - An Information Processing Approach. Morgan Kaufmann (2004)
8. Joseph, J.: Cloud Computing: Computing: Patterns For High Availability, Scalability, and Computing Power With Windows Azure. MSDN Magazine (May 2009)
9. Kurschl, W., Mitsch, S., Schönböck, J.: Modeling Distributed Signal Processing Applications. In: Proceedings of 6th International Workshop on Body Sensor Networks, Berkeley, USA (2009)
10. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: A Survey. Computer Networks, 393–422 (March 2002)
11. Shi, J., Liu, W.: A Service-oriented Model for Wireless Sensor Networks with Internet. In: Proceedings of the Fifth International Conference on Computer and Information Technology, CIT 2005 (2005)
12. Ullal, J.: Cloud Computing Conference, President and Chief Executive Officer, Arista Networks. Abstract (2009)
13. Madden, S., Franklin, J., Hellerstein, J.M., Hong, W.: TinyDB: An Acqusitional Query Processing System for Sensor Networks. ACM Transactions on Database Systems, 47 (2005)
14. Levis, P., et al.: TinyOS: An Operating System for Wireless Sensor Networks. Ambient Intelligence (2005)
15. Severance, C.: Using Google App Engine. O'Reilly (2009)
16. Mell, P., Grance, T.: Draft nist working definition of cloud computing - v15, 21 (August 2005-2009)
17. Gamma, E., Helm, R., Johnson, R.E.: Design Patterns. Elements of Reusable Object-Oriented Software. Addison-Wesley Longman (1995)
18. Kurschl, W., Mitsch, S., Schonbock, J.: Modeling Distributed Signal Processing Applications. In: Proceedings of 6th International Workshop on Body Sensor Networks, Berkeley, USA (2009)

Secure Real Time Scheduling on Cluster with Energy Minimization

Rudra Pratap Ojha[1], Rama Shankar Yadav[2], and Sarsij Tripathi[2]

[1] Galgotis College of Engineering & Technology
Greater Noida, India
[2] Department of Computer Science & Engineering
Motilal Nehru National Institute of Technology, Allahabad, India
{rpojha,mentor.sarsij}@gmail.com, rsy@mnnit.ac.in

Abstract. Security critical real time applications running over clusters are increasing day by day. These applications some times are battery operated thus they should consume minimum energy while providing both timeliness and security. Conventional real time scheduling algorithm performs poorly when used for scheduling real time application with above said constraints. So it is required to develop a scheduling approach which satisfies the above said constraints i.e. Security, Energy in such applications. We present an approach based on Dynamic Voltage Scaling (DVS) which guarantees at least minimum security (one of the QoS parameter) for the tasks with energy

Keywords: Real time System, Scheduling, Security Services, Energy.

1 Introduction

Real-Time systems span over several domains of computer science. They are defense and space systems, networked multimedia systems, embedded automotive electronics [2] etc. Real-time systems are considered to be those types of systems which have to respond to certain stimuli within a finite and specified delay. Real-Time systems are classified into two types [1] i.e. Hard Real time and Soft Real time.

Many real time applications are using clusters [3][4] for satisfying the need of high computing power where nodes are interconnected through high speed network. These applications have to satisfy timeliness of response and security requirements [13]. Applications and users can be source of security threats to cluster [20]. The security threats to these applications are primarily related to the authentication [24][27], integrity, and confidentiality[25] of application. [5].

Among these applications there are some applications which are battery operated. Therefore these applications have to satisfy security requirements [7] with minimizing energy[34]. For minimizing energy discrete speed level of processors are used based on DVS.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 904–914, 2013.

1.1 Dynamic Voltage Scaling (DVS):

To achieve better energy saving performance *dynamic voltage scaling* (DVS) technique is being used. DVS saves on the energy consumption by dynamically changing the processor supply voltage levels a characteristic supported in many modern processors such as Intel's XScale, Transmeta's Crusoe, and AMD's Duron processors. Processor power consumption can be represented by

$$P \propto \alpha C_L V^2 f \tag{1}$$

Where α is the switching activity, C_L is the load capacitance, V is the supply voltage, and f is the system clock frequency. Due to the quadratic relationship between the voltage and power consumption, reducing voltage can significantly save the power consumption for the processor. On the other hand, however, reducing the voltage supply increases the circuit delay, and thus the processor speed (s), which is given by

$$s \propto \frac{(V - V_T)^2}{V} \tag{2}$$

where V_T is the threshold voltage will decrease. This can lead to deadline miss of task. Many DVS techniques, e.g. [4, 11, 12, 13] have been proposed to reduce the energy consumption for real-time computing system.

Now problem is how to judiciously meet the two conflicting requirements i.e. Security and Energy. We have proposed an approach which ensures minimum security requirement of task while minimizing energy consumption for real time applications.

2 Related Work

Scheduling algorithms are categorized as Offline (static) and Online (dynamic) according to the time when scheduling decisions are taken. In [8] authors have proposed an uniprocessor based algorithm whereas scheduling algorithm for multiprocessor system is given in [9][11][16].In [10] a non preemptive static scheduling algorithm is used whereas dynamic scheduling algorithm for multi-processor system is given in [21][26].These algorithms did well for the real time systems but they are not applicable to applications demanding security and energy constraints.

The work reported in [14, 15] addresses scheduling on clusters with security for non real time applications. In [17, 18] authors have proposed approaches for Cluster and Grid respectively.

R. David & S. Son has given approach for secured real time databases[21]. In [22] authors have proposed a secure concurrency control protocol in real time systems. S.H. Son, C. Chaney, has also proposed security policies in this area [23]. Tao Xie and Xiao Qin proposed a security aware scheduling policy for scheduling real time applications on clusters [5]. Abhishek Songra et. al. has proposed a security criticality based approach as Modified approach for securing Real time applications on clusters (MASA) [6].

Energy-aware computing has been realized as one of the key area for research in real time systems [31]. Energy-driven voltage scheduling algorithms have been developed to reduce system's energy consumption while satisfying the timing constraints [28, 29, 30, 31, 32, 33,]. They are applicable for system having frequency dependent component (such as processors) as resource. Hua and Qu [30] introduce greedy based approach to reduce the energy consumption of the systems by utilizing the concept of DVS.

A lot of work for scheduling with security [19] and energy aware scheduling has been reported separately. But there is no work to best of our knowledge which has tackled the issues simultaneously. Thus, it is required to develop a secure energy aware efficient algorithm. Timeliness, energy and security are often seen as conflicting goals while scheduling a real-time cluster based applications. For such systems we have to find a trade off between the energy as well as security while honoring the deadline of the applications. The algorithm should guarantees the minimum security with lesser energy consumption while meeting the timing constraints.

We have proposed an approach *EMBS* in which scheduling decisions are taken in two phases: first satisfies the minimum -security requirement whereas energy saving is with speed fitting is done in second phase. The results showed the effectiveness of our approach.

3 System Model and Proposed Approach

Cluster is a group of N nodes {N_1, N_2, N_3 ….N_n) connected through a high speed network where real time applications submit task having security requirements at different speed of processor. Real time task is accepted if the cluster can schedule the task so that they complete within their respective deadline and ensures for at least minimum security requirement (related to application) in phase 1. Improvement over energy minimization may be achieved through utilization of available slack in schedule by variation of speed (Phase 2). We consider a task set having n tasks, T = { $\mathcal{T}_1, \mathcal{T}_2$ …… $\mathcal{T}_n$ }. Each task $\mathcal{T}_i$ is described with the attribute ($\mathbf{a_i}$,$\mathbf{e_i}$,$\mathbf{d_i}$,$\mathbf{L_i}$,$\mathbf{S_i}$, $\mathbf{V_i}$) where $\mathbf{a_i}$ is the arrival time, $\mathbf{e_i}$ is the execution time, $\mathbf{d_i}$ is the deadline, $\mathbf{L_i}$ is the amount of data to be secured, $\mathbf{S_i}$ is security level requirement, $\mathbf{V_i}$ speed level of processor As snooping, alteration and spoofing are three common attacks on cluster that can be handled by security services such as Authentication, Integrity and Confidentiality. These services incurred computational overhead, which depends upon amount of data secured used for securing these attacks. The overhead model for different security services i.e. Authentication [24], Integrity and Confidentiality [25] has been taken from [5].

3.1 Energy Minimization Based Scheduling

In proposed Energy minimization based scheduling algorithm we used the DVS technique to reduce the energy requirement by dynamically changing processor supply voltage. Many modern processors such as Intel's XScale, Transmeta's Crusoe and AMD's Duron processor support DVS technique. In this work we assumed that the processors have five discrete voltage levels and the corresponding normalized speed frequencies were (0.2, 0.4, 0.6, 0.8, and 1.0). We assumed that the processor speed is proportional to the supply-voltage and the processor power consumption is a cubic function of the processor speed [35]. Therefore, Energy can be calculated as:

$$\text{Energy } E = e_i\,(S_i) * S_i^3 \qquad (3)$$

Where e_i is the total execution time of the task at the assigned speed S_i. Speed and execution time depends on each other. If speed of the processor increases, execution-time will be decreased and vice-versa.

3.2 System Architecture Used

System architecture used in this architecture consists of 'm' identical nodes connected through a high speed network, where real time task submitted by the 'R number of users is shown in Figure 1 .The schedule queue holds incoming tasks submitted by users. The task submitted by the user is dispatched to the accepted queue if it pass acceptance test (Phase 1). A task is said to be pass acceptance test if task is able to complete in its deadline with minimum security requirement within available speed of the processor. Real time scheduler performs feasibility analysis (Phase 2) of newly accepted task along with other task according waiting for service or partially executed. A task passes feasibility analysis join dispatch queue where speed variation is done to minimize energy. A task fail to satisfy feasibility test join rejected queue and accepted task is dispatched to local queue of nodes in cluster.

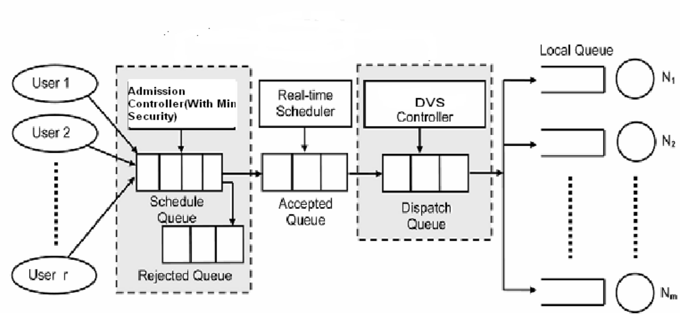


Fig. 1. System Architecture Used

Energy Minimization Based Scheduling Algorithm (EMBS):

```
1. for each task $T_i$
2. for each node $N_j$
3. Compute Utilization of nodes
4.  If (Utilization is zero)
4 (i) Select minimum speed level of processor.
  (ii).Check feasibility of task $T_i$ at present speed level
     with minimum security.
  (iii). If feasible go to step 10
          Else
          if speed level is maximum reject the task and
           go to step 11
    (iv) Increase the speed level and go to step 4(ii).
5. Else Compute Utilization of each each node $N_j$
6 (i). If $U_i$>1
  (ii) Select a node $N_J$ whose ($U_i$ > 1)
  For all task $T_n$ to $T_1$ at local queue of node $N_j$
        (i) Increase speed of task $T_n$ by one level
        (ii) Calculate Utilization
       (iii) If ($U_j$>1) and speed is maximum of $T_n$ and there
             is no task in local queue of node reject task $T_i$ goto
            Step 11 else goto step 10.
7.  Else select a node $N_j$ whose utilization is minimum.
    Initialize minimum speed level for task $T_i$.
8. Calculate Utilization after including task $T_i$ (with current speed level and minimum security) on node $N_j$
 8  (i). If ($U_j$ < 1) go to step 10
          Else
   If (U > 1) and speed level of task $T_i$ is maximum go
   to step 9.
    Else
          Increase speed level of task $T_i$ and go to step 8.
9   For all tasks $T_k$ to $T_1$ in local queue of $N_j$
    (i)   Increase speed of $T_k$ by one level at node $N_j$
    (ii)  Calculate Utilization on node $N_j$ if $U_j$<1 go to
           Step 10.
    (iii) If $U_j$ >1 and speed of $T_k$ is maximum and there
          is no task left in local queue of  node $N_j$ reject
          task $T_i$ and go to step 11.
         Else
        Increase speed of task $T_k$ by one level.
10. Assign the task to node $N_j$ with assigned speed.
11. Rejection ratio =Rejected task / No. of tasks
12. Continue if any task is there in the assigned queue.
```

Analytical Example

Let there are 4 tasks which have to be scheduled are as

T_1 = (0, 10, 180, 50, 0.4, 0.3, 0.4) T_2 = (10, 10, 250, 100, 0.3, 0.3, 0.5)

T_3 = (20, 10, 350, 50, 0.37, 0.4, 0.4) T_4 = (25, 10, 275, 100, 0.3, 0.4, 0.3)

Given data at maximum speed and minimum security required.
Number of nodes = 2, Communication Overhead=Data size in bytes/ Bandwidth

Case1: Energy consumption in case of Non DVS

Task T_1 arrived at t=0

For task T_1:- Confidentiality Overhead [0.4] = 50/33.75 = 1.481, Integrity Overhead [0.3]=50/12 = 4.167 and Authentication Overhead [0.4] = 90
Communication Overhead.(CO)= 50 KB/100Mbps = 4ms
Min. Security overhead (MSO)=1.481+4.167+90= 95.648
Finish time for task T_1 = 95.648 + 10 + 4 = 109.648
So completion Time (ET) of T_1 is e_1=109.648<150

Similarly for task T_2:-
MSO =2.667+ 8.33 + 90 =100.997
CO= 100 KB/100Mbps = 8ms
So ET for T_2 is e_2 = 100.997 +10+8 =118.997 < 250

For task T_3:-
MSO = 1.481+ 5.139 + 90 = 96.62
CO= 50 KB/100Mbps = 4ms
So ET of T_3 is e_3 = 96.62+10+4=110.62 < 350

For task T_4:-
MSO=2.667+10.277+90 =102.944
CO = 100 KB/100Mbps = 8ms
So ET for T_4 is e_4= 102.944 +10+8= 120.944 < 275

In this case tasks are executed at the maximum speed so energy consumption is
Energy E = $e_i(S_i) * S_i^3$

Let speed (maximum speed) = V
So = [109.648V^3+ 118.997 V^3 +110.62 V^3 +120.944 V^3] = 460.209 V^3

Case2: Using the concept of DVS (Dynamic Voltage Scaling)
Let Discrete speed levels are 0.2V, 0.4V, 0.6V, 0.8V, 1.0V.

For task $T_{1:}$at 0.2V e_1 = 5*109.647 = 548.24>180
at 0.4V e_1 = 274.12 >180
at 0.6V e_1 = 182.7467 >180
at 0.8V e_1 = 137.06 < 180
and at t=0 Utilization of nodes are zero so task T_1 dispatched to node N_1 with speed of 0.8V.
For task T_2:-
at 0.2 V e_2 = 594.985> 250
at 0.4V e_2 =297.4925 >250
at 0.6V e_2 = 198.328< 250
at t =10 node is free so task T_2 dispatch to node N_2 with speed 0.6V

For task T_3:-

When task T_3 arrived at t =20
Now find the utilization of nodes
For node N_1 U_1= (137.06-20)/180= 0.6503
For node N_2 = (198.328-10)/250 = 0.7533
Here U_1<U_2, so select node N_1

For task T_3 at 0.2V e_3 = 553.1>350
At 0.4v e_3 = 276.55 <350
So total U_1= 0.6503+276.55/350= 1.4404>1
Again increase sped of task T_3 at 0.6 V e_3 = 184.367<350
Total U_1= 0.6503 + 0.527 =1.1773 >1
Again increase the speed of T_3 at 0.8V e_3= 138.275
Total U_1 = 0.6503+ 0.3951 = 1.0454 >1
Again increase the speed of T_3 at speed 1.0V e_3=110.62
Total U_1= 0.6503 + 0.3160 = 0.9663 < 1
So task T_3 dispatch on node N1 with speed of 1.0 V.

For task T_4:-
Now T_4 arrived at t=25
Again calculate utilization:-
Utilization of N_1 U_1 =0.6225+ 0.3018 = 0.9243<1
Utilization of N_2 U_2= (198.328 -15)/250 = 0.73312 <1
Here U_2< U_1
So select node N_2
And for task T_4:-
at 0.2V e_4 = 604.72 > 275
at 0.4V e_4 = 302.36 >275
at 0.6 V e_4= 201.5733 < 275
So total U_2 = 0.733 + 0.733 =1.466>1

Again increase the speed of task T_4 at node N_2
At 0.8 V e_4 = 151.180
Total Utilization U_2 = 0.733+ 0.5497 = 1.2827>1
Again increase the speed
At 1.0 V e_4 = 120.944
Total Utilization U2 = 0.733 +0.4398 = 1.1728>1
Now again increase the the speed of task T_2 at node N_2
For task T_2 at 0.8V e_2 = 148.74625 <250
Total U_2 = 0.594985 + 0.4398= 1.034785 >1
Again increase the speed of task T_2 at node N_2
For task T_2 at 1.0V e_2 = 118.997
Total U_2 = 0.475988 +0.4398 = 0.915788<1
So T_4 schedule on node N_2.

Now calculate the consumed energy in case of DVS
E_1 = [137.06 *(0.8 V)^3 +118.997 V^3+110.62*V^3 + 120.944 V^3] = 420.73572 V^3
Save in energy = E – E_1 = 39.47328 V^3
% saving in energy = 8.577 %

4 Simulation and Result

The performance of proposed algorithm has been measured through simulation studies. To reveal the performance improvements gained by our proposed algorithm, we compare the Energy minimization Based Scheduling Algorithm with Non DVS.

Table 1. Simulation Parameters

Parameter	Value (Fixed) - (Varied)
Required Security Service	(Mixed)-(Confidentiality,Integrity,Authentication)
Bandwidth of Network	Fixed 100Mbps
Data size	Varied(10,50,100….500KB)
Processor Speed	Varied(0.2,0.4,0.6,0.8,1.0)

4.1 Simulator and Simulation Parameters

Studied the behavior of the EMBS algorithm under various load conditions by varying different simulation parameters. Like number of tasks, utilization of the task set, mean arrival rate.

Effect of Utilization on Energy. As from the figure 2 we can observe that almost 27% of energy is reduced in lower utilization task set 0.1 to 0.3 as compare to non DVS approach. However, at medium level utilization from 0.3 to 0.6 energy saving is 21 % and at the highest utilization energy saving is nearly 3 to 5%, this is due to the fact that at higher utilization DVS technique has lesser scope to execute task at lower speed. Hence at higher utilization DVS and non DVS have approximately same energy consumption.

Effect of Security Services over Energy. In the figure 3 we can observe that the increase in the minimum security level increases the energy. At minimum security level overhead of the security services becomes lesser due to less overhead.

Execution time of task will be minimum and large amount of slack available for task execution. If slack is available then task can execute at lower speed in case of DVS and energy is saved. When security level increases overhead of the security services and execution time of task increase and slack decreases. Due to decrease in slack, task will execute at higher speed so energy consumption increases. As security level increases at maximum value1.0 overhead and execution time will be maximum, so task will be executed at maximum speed in both the cases.

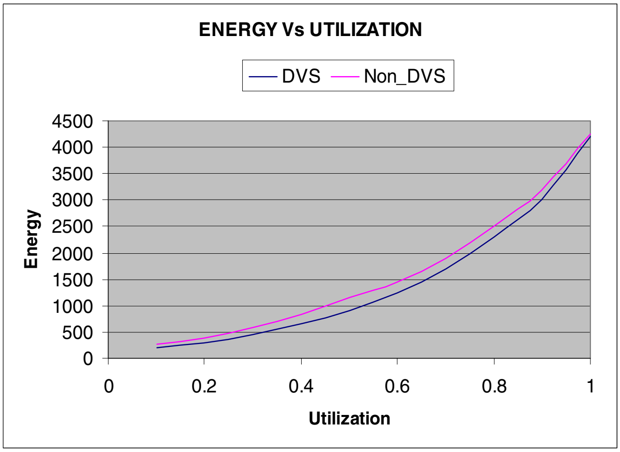


Fig. 2. Energy vs. Utilization

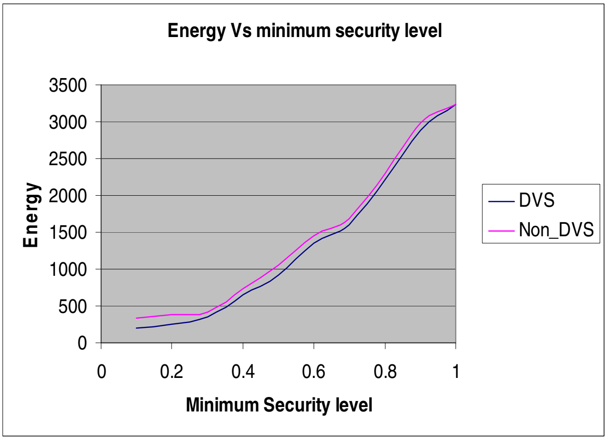


Fig. 3. Energy vs. Minimum security level

5 Conclusion

While emphasizing on security and energy we compromise with the guaranteed completion of the task and which is more important. Thus, our scheduling strategy considers the timing requirements first; the rest of the requirements come under quality. In our work we have seen that energy consumption decreased by using, our proposed approach and ensures minimum security requirement to all release of each task in such a way that receives higher guarantee ratio.

In future we would work on how to optimize security as well as the energy, where energy is minimized and the security would be maximized

References

1. Liu, J.W.S.: Real Time Systems
2. Stewart, D.B.: Courtesy of Embedded System Design
3. Pourzandi, M., Haddad, I., Levert, C., Zakrzewski, M.: A New Architecture for Secure Carrier-Class Clusters. IEEE International Conference on Cluster Computing, 494–497, 23-26 (September 2002)
4. Dessouly, A.A., Ammar, R., Ayman, E.: Scheduling Real Time Parallel Structures on Cluster Computing with Possible Processor Failure. In: IEEE 9th International Symposium on Computers and Communications, 28 June-1 July, vol. 1, pp. 62–67 (2004)
5. Xie, T., Qin, X.: Sheduling Security Aware Real Time Applications on Clusters. IEEE transactions on computers 55(7), 864–879 (2006)
6. Yadav, R.S., Songra, A., Tripathi, S.: Modified approach for securing Real time applications on clusters. International Journal of Security 1(1) (July 2007)
7. Stalling, W.: Cryptography and Network Securities (2003)
8. Parnas, D.L., Xu, J.: Scheduling Processes with Release Times, Deadlines, Precedence and Exclusion Relations. IEEE Transactions on Software Engineering 16(3), 360–369 (1990)
9. Gagne, T., Shepard, M.: A Pre-Run-Time Scheduling Algorithm for Hard Real Time Systems. IEEE Transactions on Software Engineering 17(7), 669–677 (1991)
10. Kavi, Li, W., Krishna: A Non Premptive Scheduling Algorithm for Soft Real Time Systems. Computers & Electrical Engineering 33(1), 12–29 (2007)
11. Dertouzos, M.L., Mok, A.K.: Multi-Processor Online Scheduling of Hard Real- Time Tasks. IEEE Transactions on Software Engineering 15(12), 1497–1506 (1989)
12. Hong, A.J., Tan, X., Towsley, D.: Performance analysis of minimum laxity and earliest deadline scheduling in a real time system. IEEE Transactions on Computers 38(12), 1736–1744 (1989)
13. Martel, C.U., Jeffay, K.: On Non-Preemptive Scheduling of periodic and Sporadic Tasks. In: Proceedings of the 12th IEEE Real-Time Systems Symposium, pp. 129–139. IEEE Computer Society Press, San Antonio (1991)
14. Sterling, T., Savarese, D.: A parallel workstation for scientific computation. In: Proceedings of the 24th International Conference on Parallel Processing, August 14-18. Architecture, vol. I, Urbana-Champain, Illinois (1995)
15. Shiloh, O.T., Barak, A.: Scalable cluster computing with MOSIX for LINUX. In: Proceedings of 5th Annual Linux Expo., pp. 95–100 (May 1999)
16. Genesis, M.H., Goscinski, A.M., Silock, J.: The operating system managing parallelism and providing single system image on cluster. In: Kosch, H., Böszörményi, L., Hellwagner, H. (eds.) Euro-Par 2003. LNCS, vol. 2790, Springer, Heidelberg (2003)
17. Lee, M., Kim, E.J., Yum, K.H.: An overview of security issues in cluster interconnects. In: Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops, May 16-19, vol. 2, p. 9 (2006)
18. Foster, I., Kesselman, C., Tsudik, G., Tuecke, S.: A security architecture for computational grids. In: Proceedings of the 5th ACM Conference on Computer and Communications Security CCS 1998 (1998)
19. Foster, I., Karonis, N.: Managing Security in High Performance Distributed Computations. Journal of Cluster Computing 1(1), 95–107 (1998)
20. Ferrari, A., et al.: A flexible security system for Metacomputing Environments (1999), `http://www.cs.virginia.edu/papers/hpcn99.pdf`

21. David, R., Son, S., Mukkamala, R.: Supporting Security Requirements in Multilevel Real Time Database. In: IEEE Symposium on Security and Privacy, May 8-10, pp. 199–210 (1995)
22. Mukkamala, R., Son, S.: A Secure Concurrency Control Protocol for Real-Time Database. In: IFIP Workshop on Database Security (1995)
23. Son, S.H., Chaney, C., Thomlinson, N.: Partial Security policy to Support Timeliness in Secure Real Time Databases. In: IEEE Symposium on Security and Privacy, May 3-6, pp. 136–147 (1998)
24. Bosselaers, A., Govaerts, R., Vandewalle, J.: Fast Hashing on the Pentium. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 298–312. Springer, Heidelberg (1996)
25. Elkeelany, O., Matalgah, M., Sheikh, K.: Performance Analysis of IPSEC Protocol: Encryption & Authentication. In: IEEE International Conference on Communication, vol. 2, pp. 1164–1168 (2002)
26. Zhang, X., Qu, Y., Xiao, L.: Improving Distributed Workload Performance by Sharing both CPU and Memory Resources. In: International Conference on Distributed Computing Systems (2000)
27. Deepkumara, J., Heys, H.M., Venkatesan, R.: Performance Comparison of Message Authentication Code for Internet protocol Security (2003), http://www.engr.mun.ca/~howard/PAPERS/necec_2003b.pdf
28. Mossé, D., Aydin, H., Childers, B., Melhem, R.: Compiler-Assisted Dynamic Power-Aware Scheduling for Real-Time Applications. In: Workshop on Compiler and OS for Low Power (2000)
29. Qiu, Q., Wu, Q., Pedram, M.: Dynamic Power Management in a Mobile Multimedia System with Guaranteed Quality-of-Service. In: ACM/IEEE Design Automation Conference, pp. 834–839 (2001)
30. Hua, S., Qu, G.: Energy-Efficient Dual-Voltage Soft Real-Time System with (m, k)-Firm Deadline Guarantee. In: CASES 2004, Washington, DC, USA, September 22–25, pp. 116–123 (2004)
31. Niu, L., Quan, G.: Energy-Aware Scheduling for Real-Time Systems With (m; k)-Guarantee., Dept. Comput. Sci. Eng., Univ. South Carolina. Tech. Rep. (2005)
32. Qu, G., Potkonjak, M.: Power Minimization Using System-Level partitioning of Applications with Quality of Service Requirements. In: IEEE/ACM International Conference on Computer-Aided Design, pp. 343–346 (1999)
33. Quan, G., Hu, X.: Energy Efficient Fixed-Priority Scheduling for Real-Time Systems on Variable Voltage Processors. In: 38th IEEE/ACM Design Automation Conference (2001)
34. Niu, L., Quan, G.: System-wide dynamic power management for multimedia portable devices. Accepted by IEEE International Symposium on Multimedia, ISM 2006 (2006)
35. Niu, L., Quan, G.: Energy Minimization for Real-Time Systems With (m; k)-Guarantee. IEEE Transaction on Very Large Scale Integration (VLSI) System 14(7) (July 2006)

Wireless Sensor Node Placement Due to Power Loss Effects from Surrounding Vegetation

B.S. Paul[1] and S. Rimer[2]

[1]Department of Electrical and Electronic Engineering Technology
[2]Department of Electrical and Electronic Engineering Science,
Faculty of Engineering, University of Johannesburg, South Africa
babusena_paul@yahoo.com,suvendic@uj.ac.za

Abstract. Wireless communication in an agricultural environment is weakened by surrounding vegetation. The scattering effect on the wireless signal by the foliage surrounding plants means that sensor nodes within the application area have to be placed so that the received signal strength ensures reliable communication. We propose modeling the scattering effect of surrounding foliage with a Gaussian distribution to determine the optimum placement of sensor nodes within the application area. An algorithm to place sensor nodes at optimum positions to ensure reliable communication is presented and analyzed.

Keywords: Wireless sensor networks, signal propagation, path loss, foliage, vegetation, topology.

1 Introduction

Current precision agriculture systems rely on sensors to obtain data about a specific environment. These sensor nodes are standalone devices without access to a non-renewable energy source and are located either within or close to the phenomena they are observing. The nodes communicate with one or more central control point(s), generally called a sink or base station. To collect data from sensors, a precision agriculture application densely deploys multiple wireless sensor nodes within the application area to create a multi-hop wireless sensor network (WSN) that can report real-time operational data to a central sink. The WSN bridges the virtual world of information technology and the real physical world, and is used for information gathering in smart environments [1].

The architecture of a wireless sensor node (as shown in Figure 1), is that of a small electronic device, comprising one or more transducers (for monitoring physical phenomenon), a processing unit to convert the electrical signal received from the transducer into an intelligible message format and to perform simple computations, a communication unit for transmitting and receiving messages and a non-renewable power source to provide energy to the above units [1],[2],[3][4].

Sensors are placed within the application area to ensure adequate coverage of the area. These sensor nodes are designed for unattended operation and are generally

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 915–927, 2013.

stationary after deployment. Because of the need to conserve battery lifetime, WSNs have low data rates and the data traffic is discontinuous. In WSNs the flow of data is predominantly unidirectional, from nodes to sink [5]. WSNs operate with a low duty cycle and with low data rates. Communication is initiated when data-specific information about the immediate environment around a node is requested or a specific event that the sensor has been set-up to monitor is triggered.

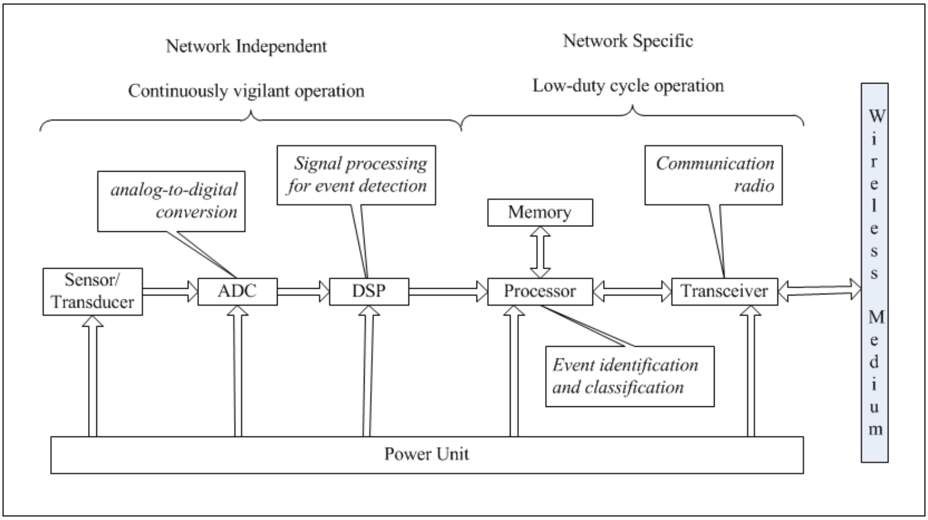


Fig. 1. Wireless sensor node architecture

Precision agriculture applies technological concepts from various sciences, including, agronomy, computer-, communication- and environmental engineering, to optimally manage spatial and temporal variability in soil and crop ecosystems in order to increase long-term quality and yield of farm products while reducing the negative effects on the surrounding flora and fauna [6, 7, 8, 9]. Wireless sensors are used as part of a precision agriculture system to provide localized, real-time data about the current temperature, humidity and soil moisture content of a specific area. Current WSN precision agricultural applications use sensors as yield monitors to support precision harvesting, and for variable rate fertilization and salinity mapping to support precision plant care [10, 11].

To increase the use of WSNs amongst both large and small-scale farmers, the topology design and deployment of sensor nodes should become easily configurable so that a non-technical person could easily deploy a WSN within an agricultural application area. One of the stumbling blocks preventing rapid adoption of WSN technologies in agriculture is that placement of nodes is dependent on experimentation and as signal strength fades, additional nodes are installed.

Precision agricultural applications require the placement of wireless sensor nodes at or near the flora being monitored. The appearance of the foliage medium in the path of the communication link has significant effects on the quality of the received signal, because discrete scatterers such as the randomly distributed leaves, twigs, branches and tree trunks can cause attenuation, scattering, diffraction, and absorption of the radiated propagating waves [12].

Thus, one of the problems associated with using a WSN application in an agricultural system is that as the life-cycle of plants progress from a seedling to a young flowering plant and eventual maturation, the surrounding vegetation around the sensor increases. This means that while fewer sensors may have been required at the seedling stage to transmit data, as the size and number of leaves of the plant increases, the wireless signal is increasingly scattered, requiring more wireless sensor nodes to be deployed in the application area to reliably transmit data to the central sink.

Various attenuation models have focused on trees, and do not consider vegetation density. Examples of current models include Weissberger's modified exponential decay model, ITU Recommendation (ITU-R) and the COST 235 model [13]. Models are required that take into consideration the different types of foliage prevalent in agriculture so that a relatively non-technical person could determine the optimum number and deployment location of nodes within an agricultural area.

In this paper, we develop a model that will optimize the topology and coverage of a WSN application area depending on the scattering of a wireless signal by a mature plant. A theory is presented and a model is developed to predict the effects of foliage on a line-of-sight propagating field.

2 Related Work

In many applications of WSN for precision agriculture, the numbers of sensor nodes are increased to ensure communication reliability as the crop matures. For example, Beckwith et al [14] deployed sensor nodes 20 to 25 meters apart in a dense 65 node, multi-hop WSN over 2 acres, to measure temperature variations over a management block of a wine vineyard. A communication reliability rate of 77% of received messages was achieved.

In the European Lofar Agro project, Baggio [15] created a 150 node WSN to monitor phytophthora, a fungal disease, in a potato field. He noted that the radio range performance of the nodes decreased substantially when the potato crop was flowering. To ensure wireless network connectivity, 30 additional relaying nodes were deployed. These relaying nodes were installed at a height of 75 cm to enhance communication, while the sensing nodes were installed at a height of 20, 40 and 60 cm. In related research, Thelen, Goense and Langendoen [16] determined that the reduced range is mainly caused by the foliage of the potato plants. The maximum distance for reliable communication is much shorter than the plane earth propagation equation indicates and that dry, fully developed crop canopy limits the distance that wireless signals can cover to around 11 meters when placed near the soil surface.

The propagated radio signals are modified by surrounding vegetation, especially due to the presence of water inside the leaves and stalks, causing delay, deviation (diffraction), or absorption (attenuation) of signal strength [17, 18]. The scattering of the radio signal by surrounding vegetation has been used to determine the growth stage and yield level of a crop. Vegetation scatter models using microwave radar signals to identify moisture in plants and grains have been developed to quantify relations between radiometric observations and vegetation parameters, like leaf area index (LAI), biomass, plant water content, etc. [17].

For example, Fung [18] developed a vegetation scatter model for interpreting scattering from a plane vegetation layer. He demonstrated that layer effects increases with a decrease in volume ratio, depth of layer, plant moisture, and, in general, on the incidence angle of the surrounding foliage. Fung determined that a successful vegetation scatter model could not be established without an adequate permittivity model which properly describes the variations of the permittivity as a function of moisture, frequency and leaf density.

Koay et al. [19] describes a theoretical model developed for paddy fields based on the radiative transfer theory applied to a dense discrete random medium with consideration given to the coherent effects and near-field effects of closely packed scatterers. The Fung and Koay papers are focused on microwave remote sensing using spaceborne radars and sensors to monitor growth and predict yield with a reasonable accuracy. However, their work can be useful in the WSN application field as there has been a large amount of research done on various scattering models and the effects of soil, moisture and leaf orientation on scattering of electromagnetic waves.

Meng, Lee and Ng, in a review of radio wave attenuation in forest environments, argue that the main external factors causing propagation loss variation is antenna height-gain, the plain terrain effect, depolarization, and the humidity effect [12].

Ndzi et al. [20] evaluated various vegetation attenuation models for frequencies in the range 0.4-7.2 GHz in mango and oil palm plantations. Their observations indicate that greater attenuation is obtained for measurement at canopy height, where there are more branches, twigs and leaves, compared to measurements at trunk heights. The authors suggest placing the nodes above the crop canopy to maximize range. However as the sensors may need to measure soil moisture, humidity and temperature etc., the placement of nodes above the crop canopy may not always be feasible.

Riquelme [21] describes the deployment of a WSN to monitor the water content, temperature and salinity of soil at a cabbage farm located in a semi-arid region of Spain. The topology of the network was not fixed and nodes were deployed arbitrarily and adapted to changing needs. Wireless coverage of the system was assured by fitting a long-range radio module to the sensor to allow direct communication with the base station 5.5km away.

Liu, Meng, and Wang [22] evaluate radio propagation performance of sensor nodes in a field of wheat at the seedling stage, booting stage and jointing stage. The minimum antenna height for each sensor node to ensure reliable communication during all three growth stages of wheat was determined. The authors found that the radio range width increases with increasing antenna height and that for any antenna height, the radio range decreases as the crop grows.

Thaskani and Rama Murthy [7] propose a WSN topology in which each stationary sensor node is placed at the corner of each grid, and a mobile base station collects data from sensor nodes and processes them. Two models for data collection are considered. In model 1, the base station moves in a horizontal direction forward and backward across the field. The mobile base station collects data from sensor nodes which are near to it at fixed instance of times. In model 2, a mobile base station is placed on mid road of the farm and sensor nodes forward data horizontally towards base station (left to right) using multiple hops. The authors do not consider the power scattering effect of surrounding foliage in determining their topology design.

3 Current Foliage Power Attenuation Models

In The main empirical foliage loss models for the horizontal path on a wireless signal's Line-of-Sight (LoS) propagation path is discussed in this section [12, 23]. Although, these models focus on foliage loss caused by trees and forests for mainly cellular type communication at microwave frequencies there is applicability of the approaches in the WSN field.

1. Weissberger's modified exponential decay model covers a frequency range of 230 MHz to 95 GHz:

$$\begin{aligned} L(dB) &= 1.33F^{0.284}I_f^{0.588}, 14 < I_f < 400m \\ L(dB) &= 0.45F^{0.284}I_f, 0 < I_f < 14m \end{aligned} \tag{1}$$

Where:

L(dB) = Vegetation loss in dB
I_f = Depth of foliage in meters
F = Frequency in GHz.

2. The Early ITU Vegetation Model (ITU-R) was mainly developed from measurements carried out at UHF, and was proposed for cases where either the transmitter or the receiver is near to a small (d < 400 m) grove of trees so that the majority of the signal propagates through the trees.

$$L(dB) = 0.2F^{0.3}I_f^{0.6} \tag{2}$$

Where:

L(dB) = Vegetation loss in dB
I_f=Depth of foliage in meters along line of sight.
F= Frequency in MHz.

3. The COST235 model is based on measurements made in millimeter wave frequencies (9.6 GHz to 57.6 GHz) through a small (d < 200 m) grove of trees. Measurements were performed over two seasons, when the trees are in-leaf and when they are out-of-leaf.

$$\begin{aligned} L(dB) &= 26.6F^{-0.2}I_f^{0.5}, out-of-leaf; \\ L(dB) &= 15.6F^{-0.009}I_f^{0.26}, in-leaf \end{aligned} \tag{3}$$

Where:

L(dB) = Vegetation loss in dB
I_f=Depth of trees in meters.
F= Frequency in MHz.

4. The Fitted ITU-R (FITU-R) model yields the smallest RMS error in vegetation loss for both the in-leaf and the out-of-leaf generic cases with reference to the experimentally measured results.

$$L_{FITU-R}(dB) = 0.37F^{0.18}I_f^{0.59}, out-of-leaf$$
$$L_{FITU-R}(dB) = 0.39F^{0.39}I_f^{0.25}, in-leaf \tag{4}$$

Where:

$L_{FITU\text{-}R}$(dB) = Vegetation loss in dB
I_f=Depth of trees in meters.
F= Frequency in MHz.

4 Theoretical Background

The transmission of signals from the source node to the destination node takes place either by line-of-sight or by signals reaching the destination after being scattered by the vegetation. In dense vegetation the probability of having a line-of-sight communication between the nodes is practically impossible. In such scenarios the signal reaches the next node only after being scattered by the leaves and branches. In this paper we model the leaves and branches as point scatterers. These scatterers are distributed in the application area (the area over which the vegetation is present) by some predefined distribution depending on the type of vegetation. Signals reaching the destination node after multiple scattering has been neglected as the power in such signals are very low compared to the signal received after single scattering [26].

Consider, P_t as the power radiated isotropically by the transmitting node. The flux density crossing the surface of a sphere with radius R meters from the transmitting node is given by,

$$F = \frac{P_t}{4\pi R^2}\ {}^{W}\!/_{m^2} \tag{1}$$

For a transmitter with output power P_t watts driving a lossless antenna with gain G_t, the flux density in the direction of antenna boresight at a distance R meter is,

$$P_r = \frac{P_t G_t}{4\pi R^2}\ {}^{W}\!/_{m^2} \tag{2}$$

If A is the effective aperture area of the antenna at the receiver, the received power is given by,

$$P_r = \frac{P_t G_t A}{4\pi R^2}\ Watts \tag{3}$$

The gain and area of an antenna are related by [24]:

$$G = \frac{4\pi A}{\lambda^2} \tag{4}$$

Where G and A are the gain and the effective aperture area of the antenna and λ is the wavelength of operation.

Combining the above equations the received power can be written as:

$$P_r = \frac{P_t G_t G_r}{\left(4\pi R/\lambda\right)^2} watts \tag{5}$$

where, G_r is the gain of the antenna at the receiver.

In the present case we consider, the gain of the transmitting and the receiving antennas to be unity. The frequency of operation is 2.4 GHz.

4.1 Model

In most plantations, for e.g. paddy, maize etc., the plants are planted at equal distances. For a full grown plant there are more leaves and branches near the central stem of the plant than away from it. The surrounding foliage can be modeled as Gaussianly distributed point scatterers around the central stem of the plant.

When statistically modeling the application area, first its dimension is set. The signal from the transmitter reaches the receiver only after being scattered by the leaves and the branches which are in range of the transmitter. The leaves and the branches are modeled as point scatterers that isotropically scatter the signals incident on it. In the present model equidistant points are initially set over the application area representing the position where the plants are located. Then Gaussianly distributed point scatterers are placed around each plant, thus modeling the surrounding foliage. In previous work [26], the scatterers are generated randomly following a uniform distribution and placed over the application area. Results from randomly generated scatterers indicate that there number of scatterers per square meter is a factor in the calculation to determine if an extra node is required to ensure connectivity and reliable wireless communication in the network. These two models will assist in identifying variables that are important in the development of a foliage attenuation model for WSNs.

The signal received at the destination node may have gone through single or multiple scattering. Multiple scattering has been neglected as the power contributed is less than the power received after single scattering. Multiple scattering takes place when there are other scatterers on the line joining the transmitter and the scatterer under consideration or the receiver and the scatterer under consideration. Thus scatterers that have only line-of-sight communication with the transmitter and the receiver take part in the scattering process. The signal received at the receiver from a particular scatterer is governed by the equation (5). As only line-of-sight communication between firstly, the transmitter and the scatterer; and secondly between the scatterer and the receiver has been considered the application of the free space model is justified.

5 Algorithm Design

Since sensors may be spread in an arbitrary manner; an important design consideration in a WSN is to ensure sensing coverage and network connectivity. In general, sensing coverage represents how well an area is monitored by sensors. The quality of a WSN can be reflected by the levels of coverage and connectivity that it offers [25]. The key to solving the coverage problem lies in the way the sensors are deployed in the area of concern. Densely covering an area with sensor nodes may not be financially viable. Instead the WSN design should take cognizance of the application requirements and put in place a WSN topology that will ensure sensing coverage and network connectivity.

We propose an algorithm which can satisfy both coverage and connectivity requirements in a WSN deployed in different types of vegetation. In earlier work, we have developed an algorithm using a uniform distribution to determine the effect of surrounding foliage on the optimum placement of the next node in a WSN. In this paper, we describe an algorithm using a Gaussian distribution to approximate the scattering effect of surrounding foliage on the signal strength. Nodes are placed at points within the applications area where the received power is greater than or equal to the experimentally determined optimum power level of **-75** dBm [25]. The algorithm ensures reliable communication between two nodes in the presence of scattering of the radio wave due to surrounding foliage.

In this algorithm, the plants are distributed in a grid a predefined distance apart. A Gaussian scattering of foliage is placed around each plant for a specific radius around each plant. Figure 2 shows the design of the algorithm to optimally place sensor nodes within an application area when a Gaussian distribution is used to approximate foliage around each plant in an application area. In previous experiments, the scatterers were randomly distributed around the sensor nodes [26].

Experimental studies were carried out in the field using an Xbee S1 XB24-AWB-001 RF transceiver that operates in the ISM 2.4 GHz frequency band, with a receiver sensitivity of -92 dBm and the transmit power is 1 mW. The Xbee modules where loaded with the function set XBEE 802.15.4 version 10E6. Measurements were taken to determine the effects of different types of foliage on the EM signal. Readings were taken of error-free received messages versus number of messages with errors depending on Received Signal Strength Indicator (RSSI). It was found that a RSSI of **-75** dBm provided 100% correct received messages per 50 transmitted messages [25].

The initial node is placed within the application area and all scatterers within range of this node are calculated. To guarantee reliable communication between adjacent nodes, a maximum communication range of 25m is used. The next node's position is calculated to be located at a position where the received power due to scattering will not be less than the experimentally determined cutoff value of -75dBm for reliable communications

If there are insufficient scatterers to place a node within the specified 25m range of another node, an additional sensor node to ensure network coverage of the application area is placed at the maximum free space range of 25m from the transmitting node. Obviously, this extra node will only be able to communicate with the next sensor node placed according to a received power of -75 dBm.

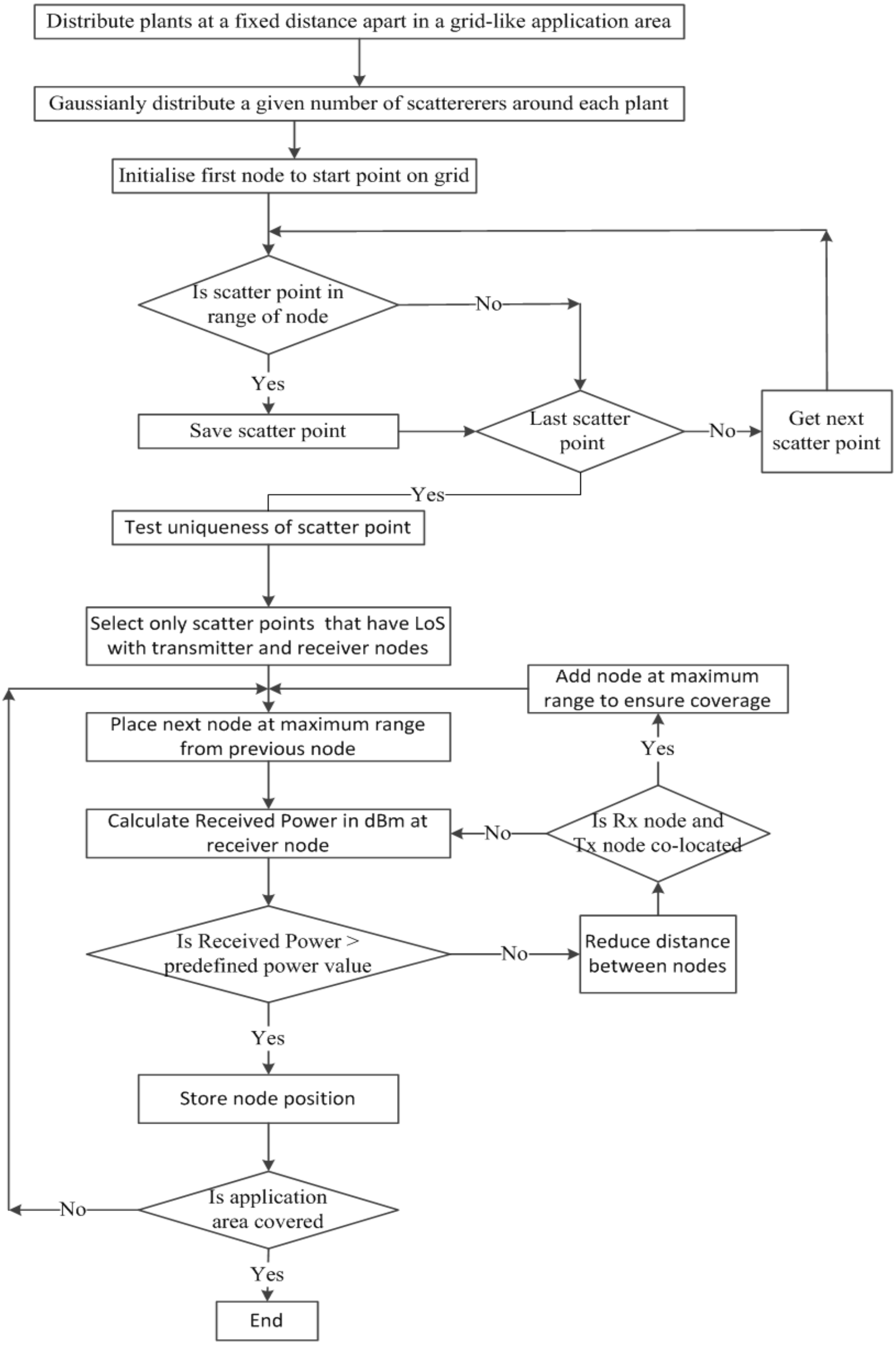

Distribute plants at a fixed distance apart in a grid-like application area
Gaussianly distribute a given number of scattererers around each plant
Initialise first node to start point on grid
Is scatter point in range of node
No
Yes
Save scatter point
Last scatter point
No
Get next scatter point
Yes
Test uniqueness of scatter point
Select only scatter points that have LoS with transmitter and receiver nodes
Add node at maximum range to ensure coverage
Place next node at maximum range from previous node
Yes
Calculate Received Power in dBm at receiver node
No
Is Rx node and Tx node co-located
Is Received Power > predefined power value
No
Reduce distance between nodes
Yes
Store node position
No
Is application area covered
Yes
End

Fig. 2. Algorithm for node placement in presence of scatterers Gaussian distributed around each plant

6 Results and Discussion

A simulation using Matlab was run to determine the optimum node position within a 100m by 100m application area. Two experiments were conducted. One to return a result depicting a uniform placement of sensor nodes around the application area, similar to the case if the sensor nodes were placed in a free space environment. The second experiment was directed to show the placement of sensor nodes and extra nodes when the power loss due to scattering by the surrounding foliage is taken into consideration.

Experiment 1: Plants placed 15m apart with 10 scatterers Gaussianly distributed around each plant

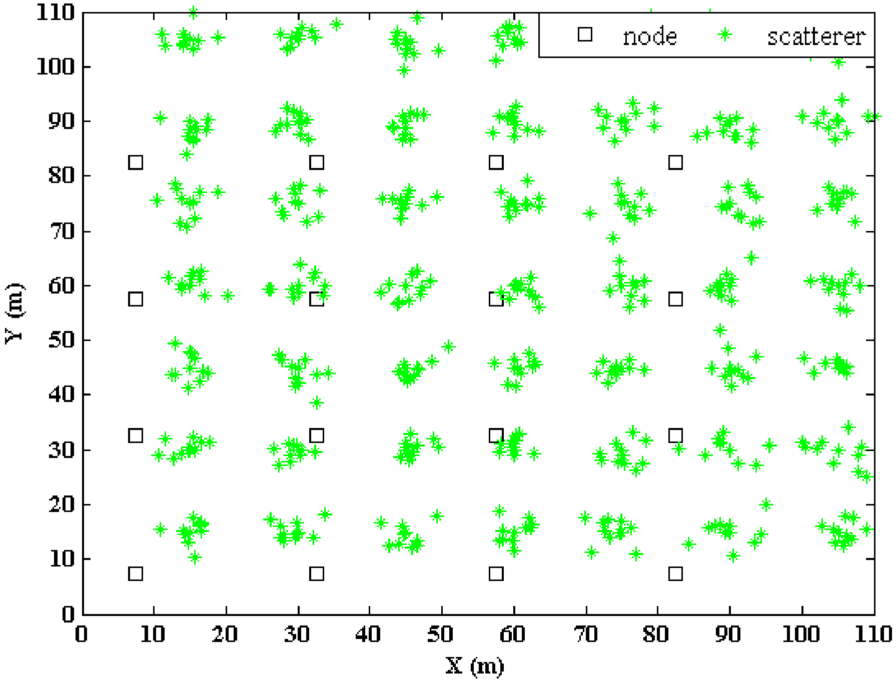


Fig. 3. Plants placed 15m apart with 10 scatterers placed around each plant

In Figure 3, ten scatterers (green asterisk) are Gaussianly distributed around each plant. The initial sensor node (black square) position was placed approximately 5 m (10/2 =5) from the plant position. The next node position to ensure coverage was chosen to be at the maximum range of the initial node position along the y-axis.

The received power level at this position due to all scatterers within range of the initial node is calculated. If the received power is greater than the experimentally determined value of -75 dBm then the position of the next sensor node is determined. If the received power is less than -75 dBm, then the node was gradually shifted closer to the initial transmitting node at a distance of 1m per calculation; and the received power recalculated. This process continues iteratively to ensure complete coverage of the application area.

Experiment 2: Plants placed 25m apart with 10 scatterers Gaussianly distributed around each plant

In Figure 4 the plants were placed 25m apart. The maximum sensor node range to ensure reliable communication is set at 25m. The initial sensor node (black square) position was placed at the same distance from the plant position as in experiment 1. The next node position to ensure coverage was chosen to be at the maximum range of the initial node position along the y-axis.

The received power level at the next node position (located to be at the maximum range of the initial node position along the y-axis position) due to all scatterers within range of the initial node is calculated. If no sensor node (within range of the transmitter node), can receive at the acceptable power level due the scattering effect of the surrounding vegetation, an extra node is placed at the maximum range distance from the transmitter node. This extra node will only be able to communicate with the next sensor node placed according to the algorithm specifications. The scattering effect of surrounding foliage results in a different placement of sensor nodes compared to sensor placement under free space conditions.

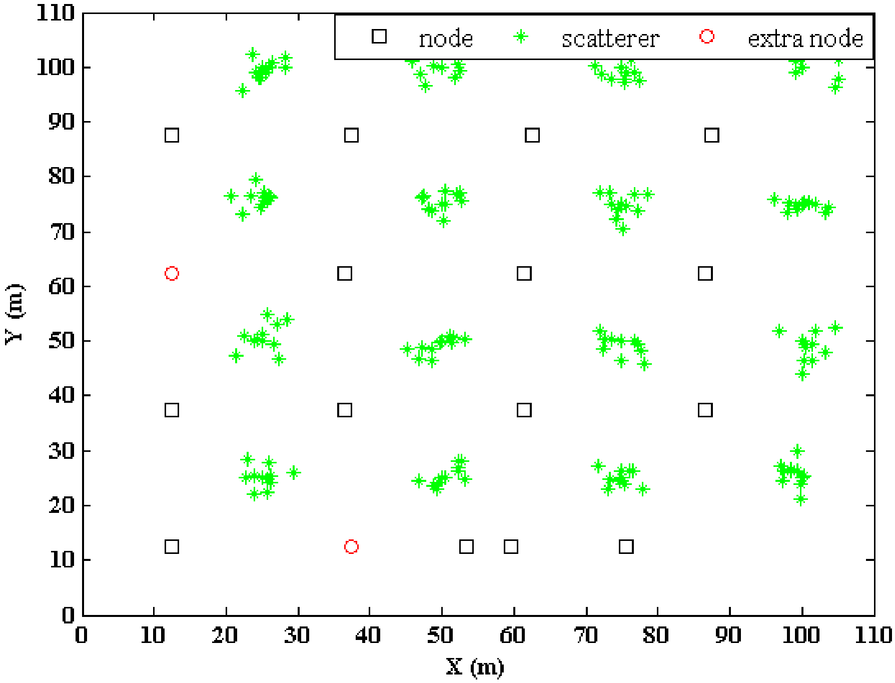


Fig. 4. Plants placed 25m apart with 10 scatterers placed around each plant

7 Conclusion

To guarantee reliable connectivity and adequate coverage of a wireless sensor network application area, sensor nodes have to be placed at positions where the received power is acceptable to ensure reliable communication between nodes. Wireless sensor nodes cannot be placed in similar positions in an agricultural environment as that calculated for a free space environment. The transmitted signal strength is weakened by surrounding vegetation.

The scattering effect on the wireless signal by the foliage surrounding plants means that sensor nodes within the application area have to be placed so that the received signal strength ensures reliable communication. We have proposed modeling the scattering effect of surrounding foliage with a Gaussian distribution to determine to optimum placement of sensor nodes within the application area. We have shown that when nodes are placed at a distance where the effect of the surrounding foliage is minimized than the location of nodes is similar to that of node placement in a free space setting.

When sensor nodes are placed so that the received signal strength is affected by the surrounding foliage, the location of the nodes differs from the free space model. Extra nodes are required to ensure coverage and connectivity within the application area.

Acknowledgments. This research was supported by funding received from the National Research Foundation THUTHUKA Programme (grant number: 76203).

References

1. Krishnamachari, B.: Networking Wireless Sensors. Cambridge University Press (2005)
2. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: A survey. Computer Networks Journal 38(4), 393–422 (2002)
3. Pottie, G., Kaiser, W.J.: Wireless integrated network sensors. Communications of the ACM 43(5), 51–58 (2000)
4. Karl, H., Willig, A.: Protocols and Architectures for Wireless Sensor Networks, 1st edn. Wiley (2005)
5. Rentala, P., Musunuri, R., Gandham, S., Saxena, U.: Survey of Sensor Networks. Technical Report UTDCS-33-02, University of Texas at Dallas (2002)
6. Patil, P., Vidya, H., Patil, S., Kulkarni, U.: Wireless Sensor Network for Precision Agriculture. In: International Conference on Computational Intelligence and Communication Networks, CICN (2011)
7. Thaskani, S., Rama Murthy, G.: Application of topology under control wireless sensor networks in precision agriculture. Report No: IIIT/TR/2010/55, Centre for Communications, International Institute of Information Technology, Hyderabad (2010)
8. Zhang, N., Wang, M., Wang, N.: Precision Agriculture – a Worldwide Overview. Computers and Electronics in Agriculture 36, 113–132 (2002)
9. Tagarakis, A., Liakos, V., Perlepes, L., Fountas, S., Gemtos, T.: Wireless Sensor Network for Precision Agriculture. In: IEEE Panhellenic Conference on Informatics (2011)

10. Dargie, W., Poellabauer, C.: Fundamentals of Wireless Sensor Networks: Theory and Practice. Wiley (2010)
11. Romer, K., Mattern, F.: The design space of wireless sensor networks. IEEE Wireless Communications 11(6), 54–61 (2004)
12. Meng, Y.S., Lee, Y.H., Ng, B.C.: Study of propagation loss prediction in forest environment. Progress In Electromagnetics Research B 17, 117–133 (2009)
13. Rama Rao, T., Balachander, D., Nanda Kiran, A., Oscar, S.: RF Propagation Measurements in Forest & Plantation Environments for Wireless Sensor Networks. In: International Conference on Recent Trends In Information Technology, ICRTIT (2012)
14. Beckwith, R., Teibel, D., Bowen, P.: Report from the field: Results from an agricultural wireless sensor. In: Proceedings of the 29th Annual IEEE International Conference on Local Computer, Washington, DC, USA (2004)
15. Baggio, A.: Wireless sensor networks in precision agriculture. In: ACM Workshop Real-World Wireless Sensor (2005)
16. Thelen, A.J., Goense, D., Langendoen, K.: Radio wave propagation in potato fields. In: First workshop on Wireless Network Measurements (co-located with WiOpt 2005), Riva del Garda, Italy (2005)
17. Giacomin, J., Vasconcelos, F.: Wireless sensor network as a measurement tool in precision agriculture. In: XVIII IMEKO World Congress, Rio de Janeiro, Brazil (2007)
18. Fung, A.: Scattering from a Vegetation Layer. IEEE Transactions on Geoscience Electronics 1(17), 1–6 (1979)
19. Koay, J.Y., Tan, C.P., Lim, K.S., Abu Bakar, B.S., Ewe, H.T., Chuah, H.T., Kong, J.A.: Paddy Fields as Electrically Dense Media: Theoretical Modeling and Measurement Comparisons. IEEE Transactions on Geoscience and Remote Sensing 45(9), 2837–2849 (2007)
20. Ndzi, D.L., Kamarudin, L.M., Mohammad, E.A.A., Zakaria, A., Ahmad, R., Fareq, M.M.A., Shaka, A.Y.M., Jafaar, M.N.: Vegetation attenuation measurements and modeling in plantations for wireless sensor network planning. Progress In Electromagnetics Research B 36, 283–301 (2012)
21. Riquelme, J.L., Soto, F., Suardíaz, J., Sánche, P.: Wireless Sensor Networks for precision horticulture in Southern Spain. Computers and Electronics in Agriculture 68(1), 25–35 (2009)
22. Liu, H., Meng, Z., Wang, M.: A Wireless Sensor Network for Cropland Environmental Monitoring. In: International Conference on Networks Security. Wireless Communications and Trusted Computing (2009)
23. Joshi, S.: Outdoor propagation models: A literature review. International Journal on Computer Science and Engineering 4, 281–291 (2012)
24. Stutzman, W.L., Thiele, G.A.: Antenna Theory and Design. John Wiley and Sons, New York (1981)
25. Ngandu, G., Nomatungulula, C., Rimer, S., Paul, B.S., Ouahada, K., Twala, B.: Evaluating effect of foliage on link reliability of wireless signal. In: IEEE International Conference on Industrial Technology (IEEE ICIT 2013), Cape Town, South Africa (2013) (accepted)
26. Paul, B.S., Rimer, S.: A foliage scatter model to determine topology of wireless sensor network. In: 1st International Conference on RADAR, Communication and Computing (ICRCC 2012), Tiruvannamalai, Tamil Nadu, India (2012) (accepted)

Increasing the Reliability of Fuzzy Angle Oriented Cluster Using Peer-to-Peer

Remani Naga Venkata Jagan Mohan[1,*], Vegi Srinivas[2], and Kurra Rajasekhara Rao[3]

[1] CSE, Swarnandra College of Engg&Tech., Narsapuram-534275, India
[2] CSE, Dadi Institute of Engg. & Tech., Anakapalli-531002, India
[3] Dept. of CSE, K.L. University, Vaddeswaram, Guntur-522502, A.P., India
{mohanrnvj,srini.vegi}@gmail.com, krr_it@yahoo.co.in

Abstract. The vast volume of data is collected and it needs to be analyzed rapidly for Quality of Data and Quality of Service (QoS), both are used for verification of sharing the information not only for web applications, but also used for many user applications over a network. In this paper, we proposed MapReduce (i.e., Parallelized and Distributed) process used for improving the performance of peer to peer communication on angle oriented clusters in Big Data. To study of this paper, the data set classified into two types namely, Clock wise and Anti-clock wise rotations using Fuzzy cluster classification. The data is extracted by using the angle oriented DCT (Discrete Cosine Transform) that invokes certain normalization techniques. Also, matching the data is compared with the technique of similarity based approach using Tanimatto distance. A high recognition rate is observed using Nelson model for this approach, and it is proved by giving an example.

Keywords: Angle Oriented, Fuzzy Cluster, MapReduce, Peer-to-Peer, Quality of Service.

1 Introduction

Reliability can be classified into two major concepts; the first one is simply the continuity of correct service delivery. On other hand, the probability of failure-free operation of a computer program for a specified period of time in a specified environment. To increasing the reliability in every software process, we can follow the two aspects to measure namely, Quality of Data and the other one is Quality of Service. Quality of Data is important for recognition of the system. Data Quality aspects are accuracy, completeness, update status, relevance consistency, reliability and accessibility. The purpose of data quality is fitness to serve in a particular perspective. It is important for operational, transactional and reliability processes. It is affected by inserting, storing the data and managing the data. The quality of data assurance is the process of verifying the reliability and effectiveness of data. To maintain the quality that needs data, going through periodical and scrubbing it. Quality of Service is a set of characteristics that are related to the performance of the

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 928–938, 2013.

connection. For every application software supports highly adaptable and be capable of tolerating a wide range of reconfigurations and extensions whereas guarantee to conduct the meeting their Quality of Service (QoS). Earlier works on these lines were proposed by various authors, like *Improving QoS for Peer-to-Peer Applications through Adaptation* is suggested by Daniel Hughes et al., 2003[3].

In our present study of this approach, quality of service is relies on complexity of software for providing high quality services to their internal and/or external clients. While the process is going on to meet the requirements such as high availability, dynamic resource allocations, and ease of management in the manufacture environment. As in the new technologies such as Hadoop and MapReduce continue their perceptions into mainstream market, more applications will be developed and moved into production. Quality of Service will certainly develop a critical consideration to move to the Big Data.

As mentioned in the above concept, is applicable on the angle oriented recognition system and is important for law-enforcement security and video-surveillance systems. In this system, we use to compare the image database and input image by peer-to-peer communication i.e., communicate directly with one-to-one or one-to-many. It is a common way to share resources on a peer-to-peer is by modifying the file sharing controls through the operating system. The angle orientation is broadly classified into two categories i.e., Clock wise and Anti-Clock wise and is suggested by Jagan Mohan et al., 2011-12 [4, 5, and 6] discussed about how to increase the reliability of angle oriented face recognition using DCT [17]. It is used to recognize the feature images of the faces even though they are in angle oriented. If the angle of input image is not equal to 90^0, rotate the image into 90^0 and then apply normalization technique such as geometric and illumination technique. Recognition of an image, by using rotational axis is easy to achieve or recognize the face. If input image rotates from horizontal axis to vertical axis the face rotates anti-clock wise and the face appears in which it is the same as the database pose, then the object is recognized. Similarly, if input image rotates from vertical axis to horizontal axis the face rotates clock wise and the face appears in which it is the same as the database pose, then the object is recognized. Therefore, if input image is angle oriented, the pose is changed or angle is altered using rotational axis and then only comparison is done.

By careful observation and comprehensive reference of above concepts, the organization of this paper as follows; In Section II, Classification of Fuzzy angle oriented cluster approaches are introduced. Section III, deals with Distance Classifier Method. Linear Regression technique is introduced in Section IV. Angle oriented with MapReduce described in section V. Cluster Reliability classifier method is discussed in Section VI. Finally, the experimental results on angle oriented cluster database are provided in section VII, and we present our Conclusion and Future Perspectives in Section VIII.

2 Nomenclature of Fuzzy Cluster

Earlier works, on Fuzzy Classification were proposed by various authors stated Fuzzy Classifications. Kuok et. al., designated fuzzy association rules such as, Mining Fuzzy Association Rule is the discovery of association rule using fuzzy set concepts such that the quantitative attribute can be handled, 2007 [10]. Jen Chianga.I et al., have given a Fuzzy classification trees for data analysis, 2002 [8][9], proposed Fuzzy classification trees are a model that integrates the fuzzy classifiers with decision trees that can work well in classifying the data with noise. Instead of defining a single class for any given instance, fuzzy classification predicts the degree of possibility for every class.

In our present discussion, I = $\{0^0, 30^0, 45^0, 60^0, 90^0, 180^0, 360^0\}$ represents all the attributes appearing that are angles in Input Training Database is C = $\{0^0,30^0,60^0,90^0\}$ contains all the possible items of a database C. Jagan Mohan et. al., discussed about the classification of clustering database images using decision tree given an Efficient K-Means Cluster Reliability on Ternary Face Recognition using Angle Oriented Approach, 2012 [12]. Whole database images denoted by C which is the root node i.e., $[0^0\text{-}90^0]$, divided into two groups namely, Cluster C_1 and C_2 representing internal nodes i.e., Clock wise and Anti Clock wise. It tells that, the similar image object groups being rotated in the clock wise belong to the internal node, cluster C_1. On the other hand, C_2 which has a group of similar image objects rotates in Anti clock wise direction. Again the Cluster C_1 is Re-grouped into three terminal nodes C_{11} $[0^0\text{-}30^0]$, C_{12} $[31^0\text{-}60^0]$ and C_{13} $[61^0\text{-}90^0]$, called nested clusters. On the other hand, the cluster C_2 is re-grouped into three terminal nodes C_{21} $[0^0\text{-}30^0]$, C_{22} $[31^0\text{-}60^0]$, C_{23} $[61^0\text{-}90^0]$, called nested clusters of C_2. The θ is the angle of rotation in each cluster.

Fuzzy sets and their corresponding membership functions have to be defined by domain experts each of the fuzzy sets can be beheld as [0, 1] valued attribute, called fuzzy attribute.

Note: A membership function f (x) is characterized by the following mapping: f: x→ [0, 1], x ε X where x is a real number describing an object or its attribute. Let X= $[0^0, 90^0]$. Define f(x) = Sin x.

A fuzzy rule as follows

$$F = \sum_{\theta=0^\circ}^{90^\circ} \sin\theta \tag{1}$$

In this regard, we propose the variable θ is broadly classified into three fuzzy relations. The fuzzy relation can be classified as follows:

$$\sum_{\theta=0^\circ}^{30^\circ} Sin\theta \ , \sum_{\theta=31^\circ}^{60^\circ} Sin\theta \ , \sum_{\theta=61^\circ}^{90^\circ} Sin\theta \tag{2}$$

Finally, the fuzzy relation can be represented by

$$\mathrm{F} = \sum_{\theta=0^\circ}^{30^\circ} Sin\theta + \sum_{\theta=31^\circ}^{60^\circ} Sin\theta + \sum_{\theta=61^\circ}^{90^\circ} Sin\theta \tag{3}$$

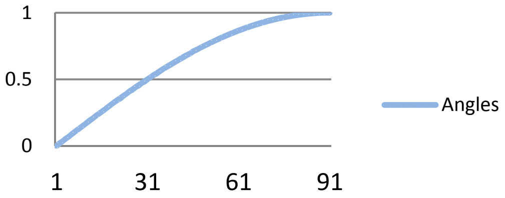


Fig. 1. Fuzzy Cluster Graph

The Θ values vary between 0^0-90^0 as shown in the above graph. By this we recognize fuzzy cluster values always in between 0 and 1.

3 Distance Classifier Method

There are few transformation techniques like DCT, DWT, KLT, SVD and Gabber methods are used for calculate the extraction of image feature vectors have certain limitations like poor discriminatory power and ability to handle large computational load. While studying of this section, we discussed, *Tanimatto distance classifier method for image matching* as in [12, 15].

The classification system analyzes the numerical properties of various image features and organizes the data into categories. This classification includes a wide range of theoretic-decision approaches for the identification of images for each image debits one or more features and each of these features belongs to one of several distinct classes. In practice, the minimum distance classifier works well, provided the distance between means is large when compare to the randomness of each class with respect to its mean. The minimum distance classifier is used to categorize into an unknown image data classes and therefore, the minimum distance between the image data and the classes in multi feature space exists. The distance is defined as an index of similarity and in consequence the minimum distance is identical to the maximum similarity. And now, this paper describes Tanimatto distance is as follows.

3.1 Tanimatto Distance Measure

The Tanimatto distance classifier is a comprehensive of Jaccard Coefficient and can be used for document data but reduce the Jaccard coefficient in the case of binary attributes. This is represented by T.

$$\text{Tanimatto Distance} = \frac{\boldsymbol{X.Y}}{||\boldsymbol{X}||^2 + ||\boldsymbol{Y}||^2 - \boldsymbol{X.Y}} \tag{4}$$

If a "similarity ratio" is given over bitmaps, where each bit of a fixed-size array represents the presence or absence of a character being modeled [2]. The definition of the ratio is the number of common bits divided by the number of bits set in either sample. The same calculation is expressed in terms of vector product and magnitude. This representation relies on the fact that, for a bit vector (where the value of each dimension is either 0 or 1) then

$$X.Y = \sum_i (X_i \Lambda Y_i) \ and \ |\mathrm{X}|2 = \sum_i (X_i) \tag{5}$$

4 Regression

In many applications, it is necessary to estimate or predict the values of a numeric attribute that has a continuous range. The numeric attribute to be estimated may depend on a multiple attributes called as multivariate regression [14]. Regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and casual inference with regression. These problems are usually solved by assuming some form of an equation that relates the input variables to dependent variable. In this paper, we proposed linear regression technique used in reduce function.

4.1 Linear Regression

In linear regression, the equation takes the form of a line and hyper plane in higher dimensions. If y is the dependent variable and x_1, x_2, x_3... x_n are the input variables, then the relation between them is of the form

$$\mathrm{Y} = w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{6}$$

The training data contains multiple records giving the values of y for specific values of the input variables. The problem thus reduces to computing the values of w_1, w_2, w_3, …, w_n that best fits the data. Once a regression model has been fit to a group of data, examination of the residuals allows the modeler to investigate the validity of the assumption that a linear relationship exists.

In the multivariate regression, the time complexity can be calculated for the number of parameters and the sample size is O (kp^2), where k is the sample size, and p is the number of parameters to be estimated. Also it deals with the popular regression methods like; robust regression, decision tree regression, and support vector regression.

5 Map Reduce Query Processing

The notion of angle oriented cluster based recognition query processing, using nearest neighbor classifier with the help of Tanimatto distance and MapReduce programming model is proposed. MapReduce has been working on query processing for optimization technique [1]. MapReduce is a functional programing model that implements for processing and generating large data sets as in[7][11][13]. It is well-developed technology for distributing and parallelized environments and also working on large cluster data sets. Initially, the user takes on input image pairs i.e., Clock wise clusters or Anti-Clock wise clusters. For each cluster is segmented into three nested clusters, to classifies in fuzzy cluster classifications as we already discussed in the above section. In each cluster produces a set of intermediate key value pairs (images) with the help of map functions written by users.

Now, first we calculate the mean value for each image i.e., intermediate key value with the help of image normalization process and feature extraction. Different authors are, Mathew Turk and Alex Pentland, 1991 [16] expanded the idea of face recognition. Jagan Mohan R.N.V. and Subbarao R, 2011 [4] given the idea to increase the Reliability of Angle Oriented Face Recognition using DCT. Each cluster consisting of intermediate key values associated with the same intermediate key and passes them to the Reduce function with the help of MapReduce library. In the same way remaining clusters are also having the same process.

Finally, the user of the reduce function is accepts an intermediate key and a set of values for that key for each cluster. We compare the mean value with the target image of database with the help of face matching using similarity based distance measures are calculated i.e., Tanimatto distance with the help of linear regression technique.

Note: Def. V: DXI→ [-1, 0]

$$(x, y) = xy = \min(x, y) - \max(x, y) \tag{7}$$

$$\forall \ \ (x, y) \ \in D\,X\,I \tag{8}$$

The user request to produce the output value whichever is minimum is required database image. The remaining values are denied.

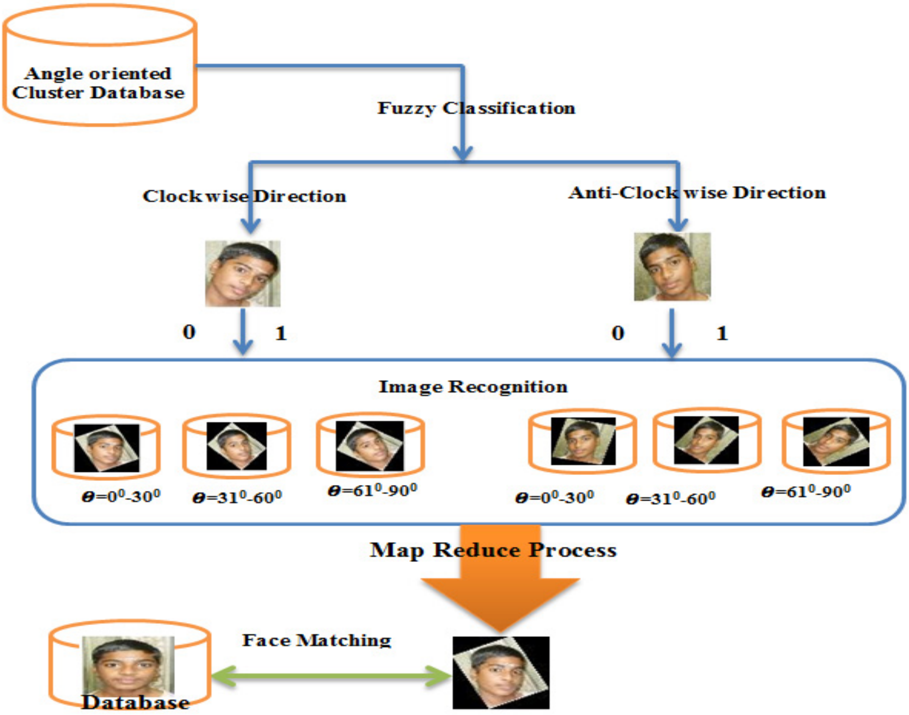


Fig. 2. MapReduce with Angle Orientation

6 Cluster Reliability

The Cluster Reliability for an angle oriented recognition system is more efficient, as shown in the above figure. This process is analyzed through experimental results.

6.1 Nelson Model

To find the cluster reliability we used distance classifier method i.e., a software reliability model is called Nelson Model [12].In this model, the software reliability is assessed by,

$$R = 1 - \frac{n_f}{n} \qquad (9)$$

Where *n* is total no.of runs and *f* is no. of failure runs.

7 Experimental Results

7.1 Anti-clock Wise Rotations

The experimental results are observed for both clock wise and anti-clockwise rotations for various clusters. The method for Tanimatto Distance is used for face

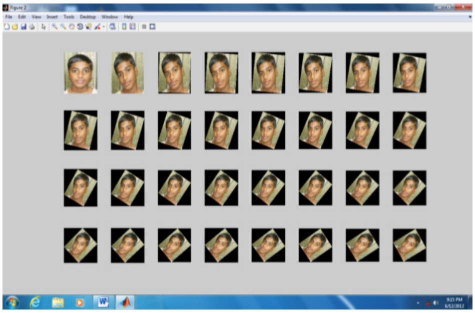


Fig. 3. Anti-Clock Wise Rotation Images

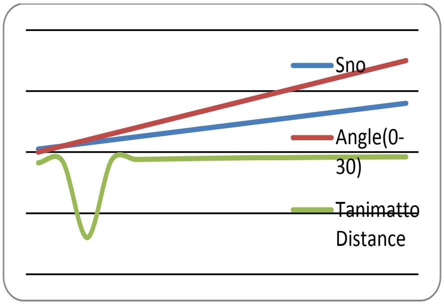


Fig. 4. C11 (0^0-30^0)

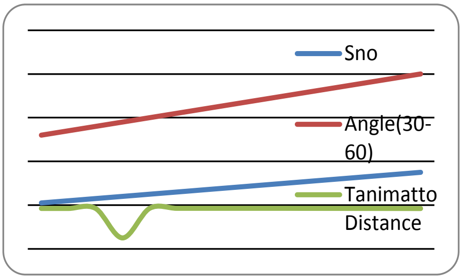


Fig. 5. C12 (30^0-60^0)

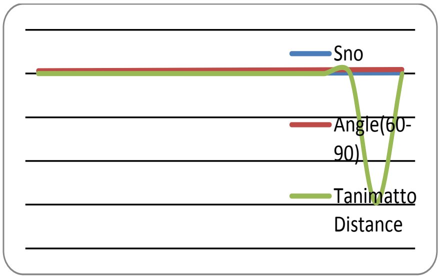


Fig. 6. C13 (60^0-90^0)

matching. The Nelson model of software reliability is used for efficiency of the face recognition in all types of clusters. The mean recognition values of the above method for all the three clusters c_{11}, c_{12} and c_{13} are calculated and the corresponding graphs for each of the three clusters are depicted in figures 4, 5 and 6 respectively.

7.2 Clock-Wise Rotations

As in the Clock wise rotation, the experimental results are calculated in anti-clock wise direction also. The same method along with Nelson model for reliability are adopted for face matching, and the mean recognition values using the algorithm given in section 1 for the three clusters c_{21}, c_{22} and c_{23} are calculated and their corresponding results are given in figures 8, 9 and 10 respectively.

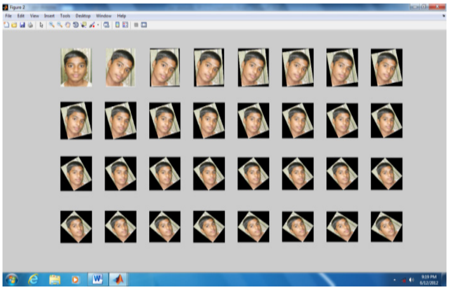


Fig. 7. Clock Wise Rotation Images

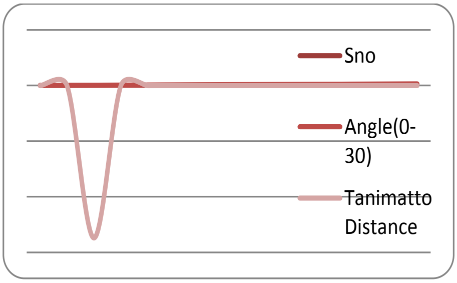


Fig. 8. C21 (0^0-30^0)

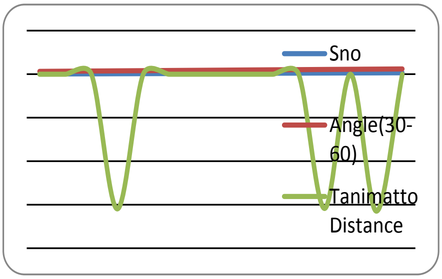


Fig. 9. C22 (30^0-60^0)

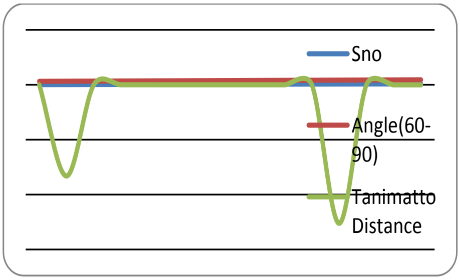


Fig. 10. C23 (60^0-90^0)

7.3 Comparison between DCT and KLT

The Tanimatto distance method is used in DCT and KLT techniques are experimented under standard execution environment by considering the created data by the students of Sri Vasista Educational Society. The phenomenal growth of DCT reliability is observed when compared with the KLT. The graph clearly shows that the reliability performance of DCT is constantly increasing with respect to KLT, while the number of records is increased.

Table 1. Performance Records in DCT and KLT

S. No.	No. of records	Performance in DCT	Performance in KLT
1	1000	91.46	65.42
2	2000	92.01	64.23
3	3000	93.25	52.15
4	4000	94.12	54.12
5	5000	94.62	53.10
6	6000	96.5	52.63
7	7000	97.23	51.71
8	8000	97.56	50.46
9	9000	98.46	50.04
10	10000	98.89	54.13

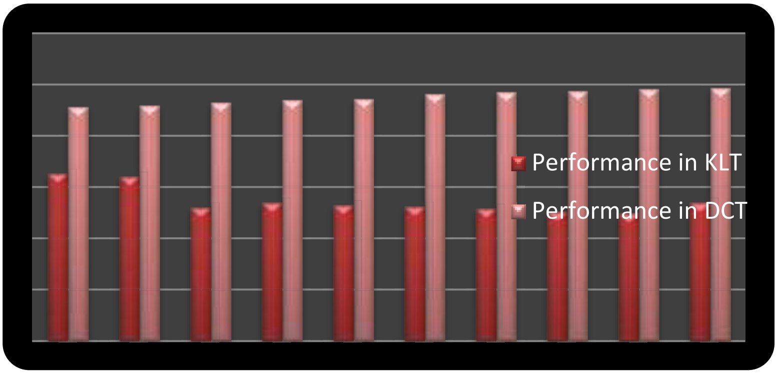


Fig. 11. Bar Chart for Performance in DCT and KLT

8 Conclusion and Future Percepective

In this paper, the Angle oriented approaches on Face Recognition is proposed and it is also used for MapReduce programming model. This model has been successfully used at Google for many different purposes. This model is easy to use, even for programmers without experience with parallel and distributed systems, since it hides the details of parallelization, fault tolerance, locality optimization, and load balanceing. MapReduce is used for a large variety of problems are easily solvable. MapReduce can implement in large clusters of datasets including thousands of clusters. The concept of face recognition is studied through Fuzzy cluster classification on the basis of rotation of images both clock-wise and anti-clock wise directions. Then, each cluster is segmented with respect to angles for face recognition using distance classifier method, Tanimatto Distance. Nelson model of reliability approach is considered for comparing cluster reliability. It is proved through the experimental results that the Tanimatto method has high probability recognition rate. This approach has several applications in Cloud Computing, Spatial Mining and new technologies like Biometric Systems etc.

References

1. Atnafu, S., Brunie, L., Kosch, H.: Similarity-based operators and query optimization for multimedia database systems. In: Proc. of Int. Database Eng. and App. Symposium (IDEAS), pp. 346–355. IEEE computer Society, Grenoble (July 2001)
2. Atnafu, S., Brunie, L., Kosch, H.: Similarity-based operators in image database systems. In: Wang, X.S., Yu, G., Lu, H. (eds.) WAIM 2001. LNCS, vol. 2118, p. 14. Springer, Heidelberg (2001)
3. Hughes, D., Warren, I., Coulson, G.: AGnuS: The Altruistic Gnutella Server. In: proceedings of the Third International Conference on Peer-to-Peer Computing, Linköping Sweden (2003)
4. Jagan Mohan, R.N.V., Subbarao. R.: Increasing the Reliability of Angle Oriented Face Recognition using DCT. In: Published In the Proceedings of International Congress on Productivity, Quality, Reliability, Optimization and Modeling, Allied Publishers (2011)
5. Jagan Mohan, R.N.V., Subba Rao, R., Raja Sekhara Rao, K.: Similarity of Inference Face Matching on Angle Oriented Face Recognition. Journal of Information Engineering and Applications 1(1) (2011) (print) ISSN2224-5758, ISSN 2224-896X
6. Jagan Mohan, R.N.V., Subba Rao, R., Raja Sekhara Rao, K.: Similarity of Inference Face Matching on Angle Oriented Face Recognition. Intelligent Systems from Published in Journal of Computer Engineering and 3(2) (2012) (Paper), `http://www.iiste.org` ISSN 2222-1719, ISSN 2222-2863
7. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Cluster. In: OSDI 2004: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA (December 2004)
8. Jen Chianga, I., Yung, J., Hsu, J.: have given a Fuzzy classification trees for data analysis. Fuzzy Sets and Systems 130, 87–99 (2002), `http://www.elsevier.com`
9. Makkithaya, K., Subba Reddy, N.V., Dinesh Acharya, U.: Intrusion Detection System using Modified C-Fuzzy Detection Tree Classifier. Published in the International Journal of Computer Science and Network Security 8(11) (November 2008)
10. Kuok, C.M., Fu, A., Wong, M.H.: Mining Fuzzy Association Rules in Databases. SIGMOD Record 27 (1998)
11. Ramakrishna, V.M., Nepal, S.: Similarity Query Processing in Image Databases. Technical Report, University Melbourne (2000)
12. Subbarao, R., Raja Sekhara Rao, K., Jagan Mohan, R.N.V.: Efficient K-Means Cluster Reliability on Ternary Face Recognition using Angle Oriented Approach. Published In the Proceedings of International Conference on Advances in Communication, Navigation and Signal Processing Technically Co-Sponsored by IEEE, Hyderabad Section, March 17-18, Dept of ECE, Andhra University College of Engineering (2012)
13. Babu, S.: Towards Automatic Optimization of MapReduce Programs. In: The1st ACM Symposium on Cloud Computing (2010)

14. Schmerr, L.W., Prabhu, G.M., Forouraghi, B.: Fuzzy Multiobjective Optimization with Multivariate Regression Trees. In: IEEE World Congress on Computation Intelligence, vol. 2, pp. 1400–1405 (1994)
15. Tanimoto, T.T., Rogers, D.J.: A Computer Program for Classifying Plants. Science 132, 1115–1118 (1960)
16. Turk, M., Pentland, A.: Eigen faces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
17. Ziad Hafed, M., Levine, M.: Face Recognition using Discrete Cosine Transform. International Journal of Computer Vision 43(3), 167–188 (2001)

Extended Ant Colony Optimization Algorithm (EACO) for Efficient Design of Networks and Improved Reliability

Mohd Ashraf and Rajesh Mishra

School of Information and Communication Technology
Gautam Buddha University, India
ashraf.saifee@gmail.com, rmishra@gbu.ac.in

Abstract. The problem of efficient network design is nothing but the NP hard problem which consisting of possible links subset selection or network topology to lower network cost subjected to the reliability constraint. Thus, in this paper we are presenting the new improved method of ant colony optimization in order to overcome such network design problem. This new algorithm is based on existing ant colony optimization algorithm. This new algorithm is having aim to optimize network reliability with least cost. This new proposed algorithm we called as extended ant colony optimizations (EACO) in which two new methods are presented, those two methods are used to optimize the search process for neighborhood and re-initialization process. Here we presented the practical approach with different network topologies in order to show the efficiency of proposed algorithm. The results of proposed method are compared with previous existing algorithms such as tabu search algorithm (TSA), genetic algorithm (GA), and ACO. From the simulation results, the proposed approach is better reliability as compared to existing algorithms.

Keywords: ACO, Network Reliability, Optimization, Heuristic, Topology.

1 Introduction

The set of links (or arcs) and nodes (or switches) a communication set of connections can be illustrate where all nodes are linked by links. The first one is backbone network and the next one is local access network (LAN). The classic communication network structure is collected of two levels the backbone network is dedicated for delivery information from source to destination (end to end) using its switch nodes. The LAN network is naturally regional system which access hosts or local servers allows users. This manuscript is focused only on distributed network. The least cost devices availability in market results in tremendous improvement in communication networks. The design of network topology is nothing but the network planning is responsible for building the feasible topology by considering the network constraints to satisfy. The distributed network is having great support to improve its reliability as compared to centralized networks. The network reliability is majorly depends on

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 939–950, 2013.

network nodes, links and network topology designed. The reliability of fully connected network is more as compared to ring network.

There are many research presents over the network design problem subjected to the network reliability. The problem of network reliability is defined as two terminal network reliability problem as well as overall reliability problem. The efficient network topology design problem in which links are selected those are either minimizes the cost or maximizes the reliability and formulated as combinatorial problem. Such kind of problem is called as NP hard problem which is tough task to solve. There were many researchers presented methods to overcome such problems. Jan et al. [3] developed an algorithm using the decomposition approach based on brand and bound to minimize link cost of communication network subjected to reliability constraint. Aggarwal et al. [4] employed greedy heuristic approach to maximize reliability given a cost constraint for networks with different reliability of links and nodes. Pierre et al. [5] also used simulated annealing to find the optimal design for packet switching networks where delay and capacity were considered, but reliability was not. For the network design, Kumar et al. [6] developed a genetic algorithm (GA) considering diameter, average distance and communication network reliability then applied it to four test problems of up to nine nodes. Deeter and Smith [7] presented a GA approach for minimum cost network design problem with alternative link reliabilities and all-terminal network reliability constraint. Furthermore, Glover et al. [8] used tabu search algorithm (TSA) to choose topologies of network when considering cost and capacity but not reliability. Other work of TSA, Beltran and Skorin-Kapov [9] used TSA to design reliable networks for searching the least cost spanning two-tree where the two-tree objective was a coarse surrogate for reliability. Next presented algorithm was ACO [10, 11] which was applied over combinatorial optimization problem successfully such as vehicle routing problem, travelling salesman problem (TSP) etc. However still to the date it's not applied network topology design problem with an objective of reliability optimization and least cost.

Hence in this paper we are discussing the new approach of ACO called as extended ACO in order to address the network topology design problem by satisfying the constraint of maximum reliability and minimum cost. Following section 2 will discuss the problem statement, problem formulation and will discuss how to calculate reliability. Section 3 will discuss the existing and section 4 discussing proposed ACO algorithms. Section 5 discussed the computational result

2 Problem Statement and Formulation

Before the formulation of problem, we will have to assume following notations to be used.

For two-terminal reliability and all-terminal reliability network design problems, there are a set of N nodes with specified topologies, which can be originate from real networks or interpreted as Euclidean distance between coordinate on a plane. It only

L	Set of Possible link
l_{ij}	Option of each link
d_{ij}	Distant between node I and node j
N	Set of given link
n	Number of node
$p(l_k)$	Reliability of link option
$c(l_k)$	Unit cost of link option
x	Architecture of network
C(x)	Total cost of network
R(x)	Reliability of network
R_0	Minimum network reliability

Fig. 1. Notations

represents some costs of connection between two nodes anyway reveal that the distance is not an offered space, of connection type. The network nodes are implicit completely dependable or assumed not to fail under any conditions. There are a set of L links which connected all nodes in N. In this problem fully connected network. It is also unspecified that there is only one link per a location. Then, all links is failed separately and restore link is necessary if any link fail. The search space of runner solutions is related to the number of the possible links, which can be found by:

$$| L |= \frac{| N || N - 1 |}{2} \tag{1}$$

A link can maybe have extra than two states. Thus a runner solution, x, is related to the state of the possible links. Therefore, this problem concern with selection a state or connected level of links l_{ij}= k where k is the level that those links connected nodes ni and nj. The mathematical formulation for the problem when minimize cost subjected to a smallest amount network dependability constraint is:

$$\textbf{Minimize } C(x) = \sum_{i-1}^{n-1} \sum_{j=i+1}^{n} c_{ij}.l_{ij}.d_{ij} \tag{2}$$

Subject to R(x)≥ R0

The cost of a specific architecture, x, is given by C(x) and the reliability of x is given by R(x). The problem is to find x which minimize the connection cost subjected to R(x) > = R0 by considering the following assumptions:

(1) All the nodes in network location is given.
(2) The cost Cij and the operation probability P_{ij} of each link (i, j) are fixed.
(3) Every link is bi-directional.
(4) No redundant link is allowed in the network.

2.1 Reliability Calculation

The problem of a network is one of research areas related to the economic network calculating or estimating the reliability of design. There are two main approach for finding the reliability which are the exact calculation through analytic methods it is tricky to find the correct dependability since these methods generally drop competence when network approaches a fully related state. And the estimation calculation through for all-terminal network reliability problem, There are also the upper and lower bound expressions for network reliability still, they drop the useful. Furthermore many bound procedures and superior efficiency surrogates in all-terminal design process. Which is related in this examine, simulation depend on the statement that all links have the equal using a backtrack procedure determination the arrangement reliability was calculated.

Due to the computational obedient size, a algorithm [12] is used to correctly calculate the global reliability R(x). The outline of the backtracking algorithm is given as follows:

```
Step # 1: [Initialize] Label all the node of communication network from 1 to N and
          all the link from 1 to M, where N is the number of node and M is the
          number of link between the nodes.
Step # 2: Represent the network with incidence matrix I.
Step #3: Generate all possible combination of link $^{M}C_{N-1}$ with N-1 link out of M
         link.
Step # 4: Repeat step #5 to step# 7 until the link combination list is empty.
Step # 5: Create the sub graph by taking the entire node (N) and add the N-1 link
           from reading link combination list.
Step # 6: Generate an adjacency matrix A corresponding to sub-graph.
Step# 7:  Call Graph_connectivity (g (V, E), s), apply to the sub graph.
          For all node v do
                /* Check the status of all node in Sub graph*/
           If (STATUS[v]=3 )
               then Store the sub graph as Spanning tree and Go to step# 4.
          Else
                Reject this combination of link (Sub graph). Go to step# 4.
          End of step # 4 loop
 Step# 8: Display all spanning tree generated in step #7 (a).
 Step # 9: R(x) = c/t.
```

2.1.1 Procedure Graph_Connectivity (g (V, E), s)

The graph_connectivity is a sub-algorithm which is used to check the connectivity of a sub graph. This algorithm visit all the node of sub graph and store the visited status in data structure array, STATUS[ ]. If the graph_connectivity ( ) visited all the node of the graph, Graph is connected otherwise graph is disconnected.

The general idea behind this algorithm beginning at a start node A is as follows. First we examine the starting node A. Then we examine each node v along a path P which begins at A; that is, we process a neighbor of A, then a neighbor of a neighbor of A, and so on. After coming to a “dead end” that is, to the end of path P, we backtrack on P until we can continue along another path P and so on. A field STATUS is used to tell us the current status of a node.

During the execution of our algorithms, each node v of G will be in one of three states, called the status of v, as follows:

```
STATUS [vi ] ←1: (Ready state) The initial state of the node N.
STATUS [vi] ←2: (Waiting state) The node is in stack, waiting to be processed.
STATUS [vi] ←3: (Processed state.) The Node v has been processed.
Step #1: Initialize all the node to ready state  STATUS[v]←1
Step #2: call Push(STACK, s)  ;
            /* push() used to insert the vertex  on the top of the stack where  STACK [1,-
              - n] be an array implementation of stack, s is the starting  vertex/*
              Set STATUS[s] ←2 ;
Step # 3: While stack is not empty
Step # 4:  call Pop(STACK, v )
              /* Remove the  top node of stack an become  visited node N */
                set STATUS [v] ←3;   /* visited  node*/
Step #5: For each neighbor of processed node v
             If (STATUS [next node]= 1)    /*ready State of the node*/
             Then
                call PUSH(STACK, v) ;
               /* insert the adjacent node of N to the top of the stack*/
             Set STATUS[v] ←2 ;          /*waiting State*/
             If (STATUS [next node]= 2)
                Call  Pop( STACK, top) ;     /*delete the current node from the stack*/
                Set STATUS [v] ←3;
                call PUSH(STACK, v)
               /*Insert the Adjacent node of v which have  STATUS[v] ←1 */
             If (STATUS[next node]= 3)
                  ignore the vertex.
             END for
             END while
Step # 6: END graph connectivity
```

When the need of a network’s reliability simulation has been arisen, two issues become important. One is the biased estimator. Others are the difference of inference. Every referenced technique is a balanced estimator where the variance of the method describes above is:

$$Var(R(x)) = \frac{R(x)(1-R(x))}{t} \tag{3}$$

To get more accurate consistency estimation, t should have larger value.

3 Conventional ACO

3.1 Ant Colony Basic Principle

The study was further continued with positive feedback distributed computation and the use of a constructive greedy heuristic is the uniqueness of an artificial ant colony. Positive feedback accounts for rapid discovery of good solutions, distributed computation avoids premature convergence and the greedy heuristic helps to find acceptable solutions in the early stages of the search process. The authors apply this approach to the classical TSP, asymmetric TSP, quadratic assignment problem (QAP) and job-shop scheduling for demonstrate the AS approach. The AS shows very good results in each applied area. Just recently, Dorigo and Gambardella have worked on extended versions of the AS paradigm. ACO is one of the extensions and has been applied to the symmetric and asymmetric TSP with excellent results. Other combinatorial optimization problems such as the vehicle routing problem has been successfully applied to the Ant System ACO is an algorithm which was inspired by the behavior of real ants. Ethnologists have deliberated how blind animals such as ants capable of finding the shortest path from food sources to the nest without using visual cues. They are also able to adapt changes in the Environment.

3.2 ACO for Network Design Problem

In this section we will discuss, how to apply ACO over network design problem. For an application of ACO algorithm to design a network topology, it is convenient to represent the network by a graph G = (N, E) where N is the set of nodes and E is the set of links. Ant's support uses the indirect form of communication mediated by pheromone they drop on the links of the graph G while building solutions. For example, a two-terminal network has four nodes and five links as showing in figure 2 below. It is possible to choose four levels for each link as shown in this network can be modeled as the routes between nest and food source for ACO as shown in.

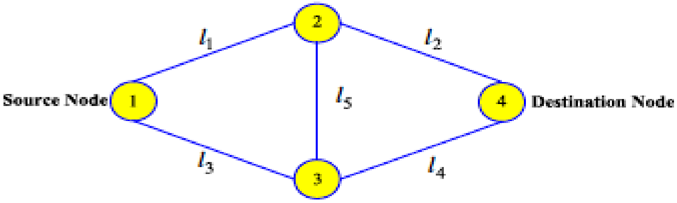


Fig. 2. Two terminals Network

This model reveals that the topology of network can be constructed from the selected connection level of each link, which seems to be the ant's route between nest and food source. In general, the procedure of ACO algorithm can be described as follows: m ants are initially positioned at the nest. Each ant will choose a possible route as a solution. In fact, each ant builds a feasible solution (called a tour) by repeatedly applying a stochastic greedy search called the state transition rule. Once all ants have terminated their tours, the following steps are performed. The amount of pheromone is modified by applying the global updating rule.

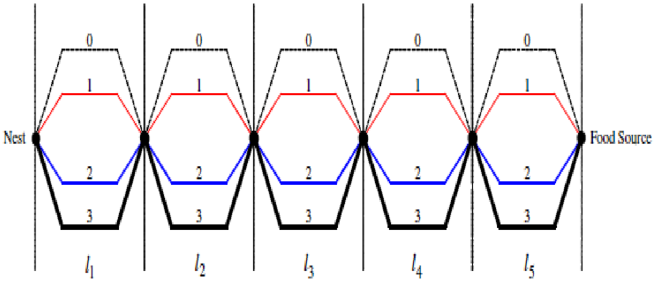


Fig. 3. Network mode as the routes between nest and food source

The pheromone updating rules are designed so that they tend to give more pheromone to edges, which should be visited by ants. A flowchart of a conventional ACO algorithm is shown.

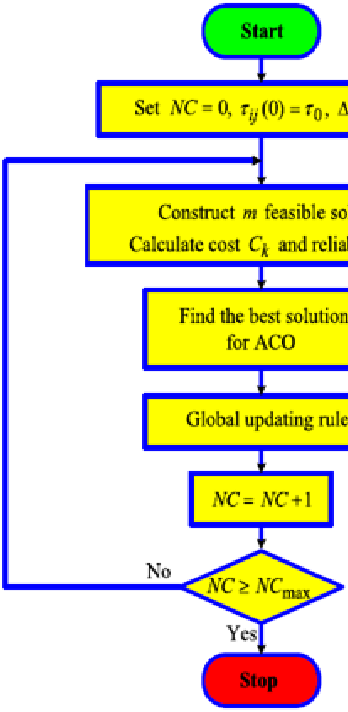

Start
Set $NC = 0$, $\tau_{ij}(0) = \tau_0$, $\Delta\tau_{ij} = 0$
Construct m feasible solutions
Calculate cost C_k and reliability R_k
Find the best solution for ACO
Global updating rule
$NC = NC + 1$
$NC \geq NC_{max}$
No
Yes
Stop

Fig. 4. Existing ACO Algorithm

3.3 Conventional Ant Colony Optimization Algorithm

Below is the description of Ant colony Optimization Algorithm

Step# A1: Initialization

```
                Set NC=0  // NC: Cycle counter
                For every combination( i, j)
                  Set an initial value $\tau_{ij}(0) = \tau_0$ and $\Delta\tau_{ij}=0$
                 End
  Step #A2: Construct feasible solution
                 For k-=1 to m  // m: number of ants
                   For i=1 to n   // n: number of links
                     Choose a level of connection with transition probability by given
                    End
                    Calculate cost $C_k$  //  $C_k$ : cost for each ant
                    Calculate reliability $R_k$  // $R_k$  : reliability for each ant
                  End
                  Update the best solution
  Step #A3: Global Updating rule
                  For every combination(i,j)
                     For k=1 to m
                        Find $\Delta\tau^{k}_{i.j}$
                     End
                        Update $\Delta\ \tau_{ij}$
                   End
                Update the trail value
                Update the transition probability
  Step # A4: Next Search
                 Set NC=NC+1
                    For every combination (i,j)
                      $\Delta\tau_{ij}=0$
                    End
  Step# A5: Termination
                 If (NC< $NC_{max}$)
                    Then
                        Go to Step A2
                        Print the best feasible Solution
                        Stop
                     End
                  End
```

4 Proposed Extended Ant Colony Optimization Algorithm

One of the strong limitations of existing ACO is that all ants are taking same place. The algorithm may be intent in a local optimal point if this situation occurs. To improve the stagnation problem of conventional ACO algorithms, two improvement procedures are applied in order to improve the ant colony optimization method for better solutions or escape from this solution in order to guarantee ants diversity. This approach is called extended ant colony optimization (EACO). The additional procedures are a specific improvement algorithm (called neighbourhood search) and re-initialization. The neighbourhood search algorithm is shown in Step E3 in EACO's algorithm, and it is proceeded to change in turn each connection level of chosen link by another connection level. For each link, connection levels are indexed in ascending order in accordance with their reliability.

4.1 Extended Ant Colony Optimization Algorithm

```
Step#B1: Initialization
         Set NC=0  // NC: Cycle counter
         For every combination( i, j)
            Set an initial value $\tau ij$ (0) =$\tau 0$  and $\Delta\tau ij$=0
         End
Step#B2: Construct feasible solution
          For k-=1 to m  // m: number of ants
            For i=1 to n   // n: number of links
                Choose a level of connection with transition probability by given
            End
                 Calculate cost $C_k$  //  $C_k$ : cost for each ant
                 Calculate reliability $R_k$  // $R_k$  : reliability for each ant
          End
             Update the best solution
  Step #B3: Apply the neighborhood search
             For k=1 to m
                 For i=1 to (2*n)
                     If i= odd
                         Change the chosen level of link i with level p by level p+1
                     Else
                         Change the chosen level of link i with level p-1
                  End
                          Calculate reliability $R_k$
                      If   ( $R_k \geq R_0$)
                          Except for exchanging
                          Record the obtained solution
                      Else
                           Do not except for exchanging
                 End
                           Calculate the cost $C_k$
                 End
             Update the best solution
  Step #B4: Next Search
            Set NC=NC+1
                For every combination (i,j)
                    $\Delta\tau ij$=0
                End
  Step #B5: Re-initialization
            If the best solution has not been improved for a long time
             Then
```

```
                Set an initial value τij(0)= τ0
            End
Step #B6: Global updating rule
            For every combination(i,j)
               For k=1 to m
                   $Find \vartriangle \tau_{i.j}^{k}$
               End
                  Update Δ $\tau_{ij}$
            End
                 Update the trail value
                 Update the transition probability
Step #B7: Termination
            If (NC< NCmax)
               Then
                   Go to Step#B2
                   Print the best feasible Solution
                   Stop
               End
            End
```

4.2 Flow Chart for Extended Ant Colony Optimization

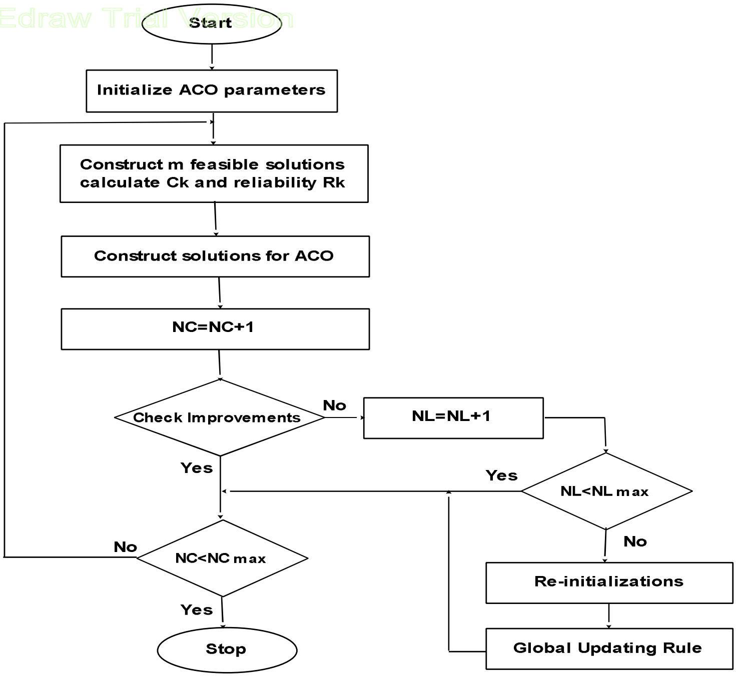

Start
Initialize ACO parameters
Construct m feasible solutions calculate Ck and reliability Rk
Construct solutions for ACO
NC=NC+1
Check Improvements
No
NL=NL+1
Yes
Yes
NL<NL max
No
NC<NC max
No
Re-initializations
Yes
Stop
Global Updating Rule

Fig. 5. Flow chart for proposed extended Ant colony optimization

5 Computation Result

The effectiveness of the proposed EACO algorithm has been evaluated with different network topology designs and compares its performance with existing ACO, TS, and GA approaches. Each studied system was run 30 times with differential random initial solutions. In order to evaluate the performance of each technique following graphs showings the cost required and reliability computation for each approach in case of solving Schaffer function and sphere function:

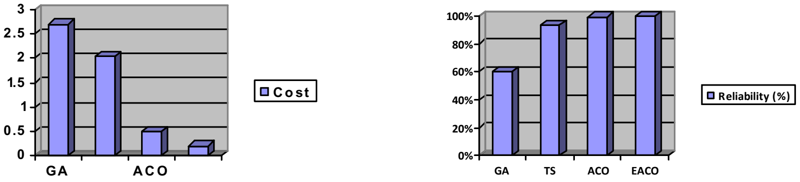


Fig. 6. The results obtained by GA, TS, ACO and IACO for solving Schaffer function

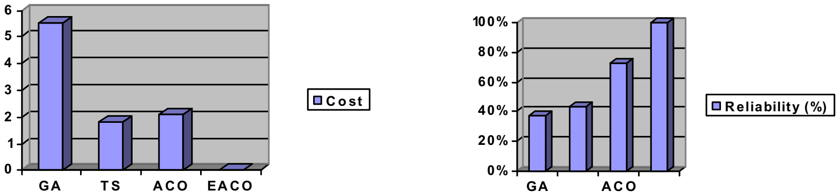


Fig. 7. The results obtained by GA, TSA, ACO and IACO for solving Sphere function

6 Conclusions

Thus in this paper we presented the EACO for optimal reliability for network design with least cost. The proposed EACO method has been useful to solve the topology network design problem considering both economics and reliability. The shows superior features such as high-quality solution, stable convergence characteristic and good computation efficiency EACO algorithm. Convergence characteristic and computation efficiency compared with GA, TSA and ACO methods Above results which we perform with sample test function showing that proposed algorithm having better improvements in reliability and having least cost. For future work we will apply this algorithm over graph theory based networks and evaluate its results with existing cases.

References

1. Aho, A.V., Hopcroft, J.E., Ullman, J.D.: Design and analysis of Computer Algorithms. Addison – Wesley, Massachusetts (1974)
2. Gerla, M., Frank, H., Eckl, J.: A cut saturation algorithm for topological design of packet witched communication networks. Proc. NTC, 1074–1085 (1974)
3. Rong-Hong, J., Fung-Jen, H., Sheng-Tzong, C.: Topology optimization of a communication network subject to a reliability constraint. IEEE Trans. Reliab. 42(1), 63–70 (1993)
4. Aggarwal, K.K., Chopra, Y.C., Bajwa, J.: Reliability evaluation by network decomposition. IEEE Trans. Reliab. R-31, 355–358 (1982)
5. Pierre, S., Hyppolite, M.A., Bourjolly, J.M., Dioume, O.: Topology design of computer communication network using simulated annealing. Eng Appl. Artif. Intell. 8, 61–69 (1995)
6. Kumar, A., Pathak, R.M., Guptaand, Y.P., Parsaei, H.: A genetic algorithm for distributed system topology design. Comput. Ind. Eng. 28, 659–670 (1995)
7. Deeter Darren, L., Smith Alice, E.: Economic design of reliable networks. IEEE Trans. Econ. Reliab. Eng. (July 1998)
8. Glover, F., Lee, M., Ryan, J.: Least-cost network topology design for a new service: an application of a tabu search. Ann. Operat. Res. 33, 351–362 (1991)
9. Beltran, H.F., Skorin-Kapov, D.: On minimum cost isolated failure immune networks. Telecommun. Syst. 3, 183–200 (1994)
10. Colorni, A., Dorigo, M., Maniezzo, V.: Distributed optimization by ant colonies. In: Proceedings of the European conference on artificial life, pp. 134–142 (1991)
11. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperative agents. IEEE Trans.Syst. Man. Cybernet B: Cybernet 26(1), 29–41 (1996)
12. Ashraf, M., Rajesh, M.: Computing global reliability for wireless sensorNetworks subject to spanning tree enumeration approach. In: International Journal of Computer Science and Engineering (IJCSE), vol. 1(1), pp. 1–14 (Auguest 2012)

Evaluation of Understandability of Object-Oriented Design

Devpriya Soni

Department of Computer Science and Engineering
Jaypee Institute of Information Technology, Noida, sec- 128, India
devpriyasoni@gmail.com

Abstract. Quality of software design directly affects the understandability of the software developed. As the size and complexity of the software increases it drastically affects quality attributes, especially understandability. The direct measurement of quality is difficult because there is no single model that can be applied in all situations. Models proposed by various researchers are not comprehensive. Quantitative measurement of an operational system's understandability is desirable both as an instantaneous measure and as a predictor of understandability over time. This work proposes the method of measuring understandability using Logical Scoring of Preferences (LSP) method. I have also evaluated one design through this model.

Keywords: Software Quality, Quantitative Measurement, LSP.

1 Introduction

The demand of the quality software is increasing at rapid pace due to the society's increasing dependence on software. Measuring quality in the early stage of software development is the key to develop high-quality software. Wrong interpretations can lead to misunderstandings and to faulty development results. It is difficult to manage and improve the process without understanding and the ability to properly express the process in use Therefore, the readability and understandability of the software has a lot of influence on the factors that directly or indirectly affect software quality. Complex design may lead to poor testability, which in turn leads to ineffective testing that may result to severe penalties and consequences. It is well understood fact that flaws of design structure have a strong negative impact on quality attributes. But, structuring a high-quality design continues to be an inadequately defined process [1]. Therefore, software design should be built in such a way so as to make them easily understandable, testable, alterable, and preferably stable. This work focuses on the understandability assessment during the design phase to produce quality software.

Our methodology for the quantitative evaluation of software's understandability in the design phase is based on the core evaluation models and procedures are grounded in the LSP model and continuous preference logic as mathematical background [2]. Kumar and Soni [4] have proposed a hierarchical model of quality attributes. This is used to evaluate quality of human resource system design which was proposed by Kumar and Gandhi [3].

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 951–961, 2013.

2 Previously Proposed Quality Models for Object- Oriented Software Products

One of the earliest software product quality model was suggested by McCall et.al.[5]. They defined software product qualities as a hierarchy of factors, criteria and metrics. The McCall's quality factors are correctness, reliability, efficiency, integrity, usability and maintainability. Boehm [6] described a set of quality characteristics. International bodies ISO/IEC came up with ISO9126 model for ensuring quality in software products. The ISO9126 [7, 17] model defines six quality attributes namely functionality, reliability, efficiency, usability, maintainability and portability. They are further subdivided into 26 sub-attributes (criteria) and nearly 100 sub-criteria or metrics. All these models were developed for structured methodology of software product development.

Even though there are many object-oriented analysis and design methodologies, languages, database management systems and tools, relatively less work has been done in the area of object-oriented design quality assurance [7, 8]. However, many metrics were developed to measure size and complexity of an object-oriented software system. One of the most popular set of metrics (commonly know as CK Metrics suite) was proposed by Chidamber and Kemerer [9]. The same suite was later refined and presented with empirical validation by Chidamber and Kemerer [10]. Basili et.al. [11, 12] also performed the empirical validation of CK metrics suite.

A framework for building product based quality models has been developed by Dromey [13,14]. The framework is a methodology for development of quality models in a bottom-up fashion, providing an approach that will ensure that the lower-level details are well specified and computable [12]. Bansiya et.al. [15] extended this methodology to develop the hierarchical Quality-Model for Object-Oriented Design (QMOOD) assessment. In the Quality Model for Object-Oriented Design (QMOOD), Bansiya et.al [15] identified the initial set of design quality attributes as: functionality, effectiveness (efficiency), understandability (maintainability), extendibility (portability), reusability and flexibility.

Further, Keller and Cockburn [16] organized a workshop, in which one group agreed upon following list of perspectives, with each having substantial influence on the quality of design artifacts: maintainability, documentation, extensibility, cost, reliability, ease of use, internationalization, usability, market goals, performance, team structure.

The second group discussed design properties that are of interest for project participants (developers) and gave following attributes: clarity, simplicity, scalability, modifiability, extendibility, reusability, effectiveness, reliability, robustness, security, and cost.

The metrics proposed by Bansiya et.al. [15] are quite general in nature and they have not provided the methodology to measure these metrics. Keller and Cockburn [16] have observed that there was no consensus on the quality attributes. However, they prescribed attributes and metrics that are very broad in nature and are not in conformance with ISO/IEC 9126 standards.

The author in her previous work [4] has given a generic model which assesses quality of design in early stage of software product development life cycle. This hierarchical model is based on five factors, their sub-factors and metrics and shown in Figure 2.1. These five-factors for design quality assessment are: functionality (modifiability), effectiveness (efficiency), understandability (usability), reusability, and maintainability (flexibility).

3 Steps for the Evaluation of Design Quality

Steps required for the evaluation of design quality are:

1. ***Consider a hierarchical model for quality characteristics and attributes (i.e. A1 …. An):*** here, evaluators should define and specify the quality characteristics and attributes, grouping them into a model. For each quantifiable attribute Ai, we can associate a variable Xi, which can take a real value: the measured value.
2. ***Defining criterion function for each attribute and applying attribute measurement:*** In this process, the evaluators should define the basis for elementary evaluation criteria and perform the measurement sub-process. An elementary evaluation criterion specifies how to measure quantifiable attributes. The result is an elementary preference, which can be interpreted as the degree or percentage of satisfied requirement. For each variable Xi , i = 1, ...,n it is necessary to establish an acceptable range of values and define a function, called the elementary criterion. This function is a mapping of the measured value in the empirical domain [18] into the new numerical domain. Then the final outcome is mapped in a preference called the elementary quality preference, EQi. We can assume the elementary quality preference EQi as the percentage of requirement satisfied by the value of Xi . In this sense, EQi = 0% denotes a totally unsatisfactory situation, while EQi = 100% represents a fully satisfactory situation [2]. Ultimately, for each quantifiable attribute, the measurement activity should be carried out.
3. ***Evaluating elementary preferences:*** In this task, the evaluators should prepare and enact the evaluation process to obtain an indicator of partial preference for design. For n attributes, the mapping produces n elementary quality preferences.
4. ***Analyzing and assessing partial quality preferences*:** In this final step, the evaluators analyze and assess the elementary, partial and total quantitative results regarding the established goals.

3.1 Establishing Elementary Criteria for Understandability

The significance of understandability is very obvious that can be perceived as 'If we can't learn something, we won't understand it. If we can't understand something, we can't use it - at least not well enough to avoid creating a money pit. We can't maintain a system that we don't understand - at least not easily. And we can't make changes to our system if we can't understand how the system as a whole will work once the changes are made' [22]. Understandability of software documents is thus important as 'the better we know what the thing is supposed to do, the better we can test for it'.

A good software design with manageable complexity usually provides proper data abstraction; it reduces coupling while increasing cohesion that make them easily understandable. As advocated by researchers and practitioners that understandability aspect of software is highly desirable and significant for developing quality

1. Functionality
1.1 Design Size
1.1.1 Number of Classes (NOC)
1.2 Hierarchies
1.2.1 Number of Hierarchies (NOH)
1.3 Cohesion
1.3.1 Cohesion Among Methods of Class (CAM)
1.4 Polymorphism
1.4.1 Number of Polymorphic Methods (NOP)
1.5 Messaging
1.5.1 Class Interface Size (CIS)
2. Effectiveness
2.1 Abstraction
2.1.1 Number of Ancestors (NOA)
2.1.2 Number of Hierarchies (NOH)
2.1.3 Maximum Depth of Inheritance (MDIT)
2.2 Encapsulation
2.2.1 Data Access Ratio (DAR)
2.3 Composition
2.3.1 Number of aggregation relationships (NAR)
2.3.2 Number of aggregation hierarchies (NAH)
2.4 Inheritance
2.4.1 Functional Abstraction (FA)
2.5 Polymorphism
2.5.1 Number of Polymorphic Methods (NOP)
3. Understandability
3.1 Encapsulation
3.1.1 Data Access Ratio (DAR)
3.2 Cohesion
3.2.1 Cohesion Among Methods of Class (CAM)
3.3 Inheritance
3.3.1 Functional Abstraction (FA)
3.4 Polymorphism
3.4.1 Number of Polymorphic Methods (NOP)
5.6.1 Number of aggregation relationships (NAR)
5.3.2 Number of aggregation hierarchies (NAH)

4. Reusability

4.1 Design Size

4.1.1 *Number of Classes (NOC)*

4.2 Coupling

4.2.1 *Direct Class Coupling (DCC)*

4.3 Cohesion

4.3.1 *Cohesion Among Methods of Class (CAM)*

4.4 Messaging

4.4.1 *Class Interface Size (CIS)*

5. Maintainability

5.1 Design Size

5.1.1 *Number of Classes (NOC)*

5.2 Hierarchies

5.2.1 *Number of Hierarchies (NOH)*

5.3 Abstraction

5.3.1 *Number of Ancestors (NOA)*

5.4 Encapsulation

5.4.1 *Data Access Ratio (DAR)*

5.5 Coupling

5.5.1 *Direct Class Coupling (DCC)*

5.5.2 *Number of Methods (NOM)*

5.6 Composition

5.6.1 *Number of aggregation relationships (NAR)*

5.3.2 *Number of aggregation hierarchies (NAH)*

5.7 Polymorphism

5.7.1 *Number of Polymorphic Methods (NOP)*

5.8 Documentation

5.8.1 *Extent of Documentation (EOD)*

Fig. 2.1. Hierarchical design quality assessment model

software. Through the findings of literature survey there are various aspects of software that either directly or indirectly influences quality of software design including understandability factor [19], [20].

Therefore, out of the five factors of the hierarchical model [4] I have focused on the understandability aspect in this work. Understandability is further decomposed into four sub factors namely: encapsulation, cohesion, inheritance and polymorphism. However, I have measured only three sub-factors in this work and they are: encapsulation, cohesion and polymorphism.

For each attribute Ai we can associate a variable Xi which can take a real value by means of the elementary criterion function. The final result represents a mapping of the function value into the elementary quality preference, EQi. The value of EQi is a real value that ‘fortunately’ belongs to the unit interval. As stated by Dujmovic et al. in [2]:

“*the elementary preference is interpreted as a continuous logic variable. The value 0 denotes that Xi does not satisfy the requirements and the value 1 denotes a perfect satisfaction of requirements. The values between 0 and 1 denote a partial satisfaction of requirements. Consequently, all preferences are frequently interpreted as a percentage of satisfied requirements, and defined in the range [0, 100%]*”.

Further, the preference can be categorized in three rating levels namely: satisfactory (from 60 to 100%), marginal (from 40 to 60%), and unsatisfactory (from 0 to 40%). For instance, a marginal score for an attribute could indicate that a correction action to improve the attribute quality should be taken into account by the manager or developer. Figure 3.1, shows two elementary criteria for attributes of understandability. There are two major categories to classify elementary criteria, that is, absolute and relative criteria. Moreover, regarding the absolute elementary criteria, these are further decomposed in continuous and discrete variables.

The preference scale for the *Data Access Ratio (DAR)* metric is a multi-level discrete absolute criterion defined as a subset, where 0 implies ratio is less then 5%; 80% or more implies satisfactory (100%) ratio.

The resulting value of this discrete multivariable absolute criterion could be between 0 (completely unsatisfactory) and Xmax (completely satisfactory). If the measured value of X is above Xmax, the corresponding elementary preference X will be equal to Xmax. Similar criteria were followed for other metrics as well.

3.2 Computing Partial Preference for Maintainability

In this process, the evaluators should define and prepare the evaluation process to obtain a quality indicator for each competitive system. Applying a stepwise aggregation mechanism, the elementary quality preferences can be accordingly structured to allow the computing of partial preferences. Thereby global preferences can be obtained through repeating the aggregation process at the end. The global quality preference represents the global degree of satisfaction of all involved requirements. Here I am computing partial preferences for understandability. In this study, we use a logical scoring of preferences model called LSP model. A broad treatment of LSP relationships and continuous Logic Preference (CLP) operators could be found in [2, 21], as well as the mathematical background.

The strength of LSP resides in the power to model different logical relationships to reflect the stakeholders' needs, namely:

- Simultaneity, when is perceived that two or more input preferences must be present simultaneously
- Replaceability, when is perceived that two or more attributes can be replaced (there exist alternatives, i.e., a low quality of an input preference can always be compensated by a high quality of some other input).
- Neutrality, when is perceived that two or more input preferences can be grouped independently (neither conjunctive nor disjunctive relationship)
- Symmetric relationships, when is perceived that two or more input preferences affect evaluation in the same logical way (tough may be with different weights)
- Asymmetric relationships, when mandatory attributes are combined with desirable or optional ones; and when sufficient attributes are combined with desirable or optional ones.

Figure 3.2, depicts the aggregation structure for understandability characteristic. The stepwise aggregation process follows the hierarchical structure of the hierarchical model from bottom to top. The major CLP operators are the arithmetic means (A) that models the neutrality relationship; the pure conjunction (C), and quasi-conjunction operators that model the simultaneity one; and the pure disjunction(D), and quasi-disjunction operators that model the replaceability one. With regard to levels of simultaneity, we may utilize the week (C-), medium (CA), and strong (C+) quasi-conjunction functions. In this sense, operators of quasi-conjunction are *flexible and* logic connectives. Also, we can tune these operators to intermediate values. For instance, C-- is positioned between A and C- operators; and C-+ is between CA and C operators, and so on. The above operators (except A) mean that, given a low quality of an input preference can never be well compensated by a high quality of some other input to output a high quality preference. For example at the end of the aggregation process we have the sub-characteristic coded 3.1 (called Encapsulation in the hierarchical Model, with a relative importance or weight of 0.3), and 3.2 sub-characteristic (Cohesion, 0.4 weighted), and 3.4 sub-characteristic (polymorphism, 0.3 weighted).

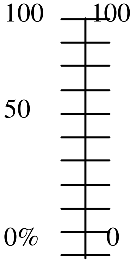
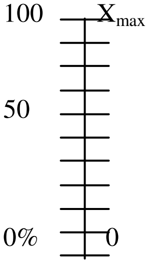
Data Access Ratio A = Availability 0= Not Available 0.8=Partially Available 1=Totally Available		**Cohesion Among Methods of Class(CAM)** 0= Not Available 1= Cohesion between 5 or more classes	

Fig. 3.1. Sample elementary criteria defined as preference scales taken from the hierarchical model

All these sub-characteristic preferences are input to the C-- logical function, which produce the partial global preference coded as 3, (called Understandability).

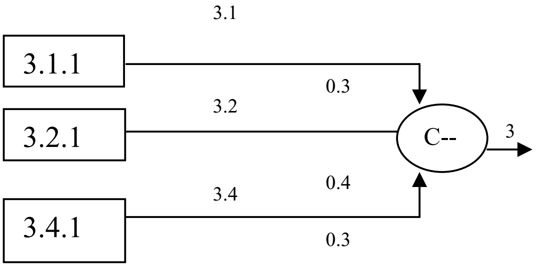


Fig. 3.2. Structure of Partial Logic Aggregation for Understandability

Similarly, we can also utilize the quasi-disjunction operators in a range of strong (D+), medium (DA), and week (D-) *or* polarization, and also their intermediate values. For instance, D-- is positioned between A and D- operators; and D-+ is between DA and D- operators; and D+- is between D+ and DA operators; and finally, D++ is between D+ and D operators. D operator represents the pure disjunction.

4 Assessing Understandability of the Design Selected

Figure 4.1 shows the design of human resource management information system, which is developed to take care of the important function of the Human Resource Development. The system keeps record of the employees both regular and ad-hoc along with their qualification details, the designation at the time of joining the organization, the present designation and number of promotions any employee has been given since he joined the organization. It keeps the detailed record of employee family members, medical facilities along with his telephone number, job responsibilities of each and every employee and the reporting officer/person of each employee is also maintained and several other information as shown in Fig 4.1.

In the evaluation process, I decided the elementary criterion for each metric, as shown in fig 3.1. I then confronted partial preferences as shown the section 3.2 and fig 3.2.

The partial outcomes for each subfactor and the total outcome for understandability is shown in Table 1.

This shows that the design of the human resource information system is falling into a satisfactory level because it has 85.79% of the quality preference.

Table 1. Detailed result of partial quality preferences after computing the aggregated criteria function of the design

Characteristics and Sub-characteristics	Values
3. Understandability	
3.1 Encapsulation	
3.1.1 Data Access Ratio (DAR)	.8
3.2 Cohesion	
3.2.1 Cohesion Among Methods of Class (CAM)	.8
3.4 Polymorphism	
3.4.1 Number of Polymorphic Methods (NOP)	1
Partial Quality Preference	**85.79**

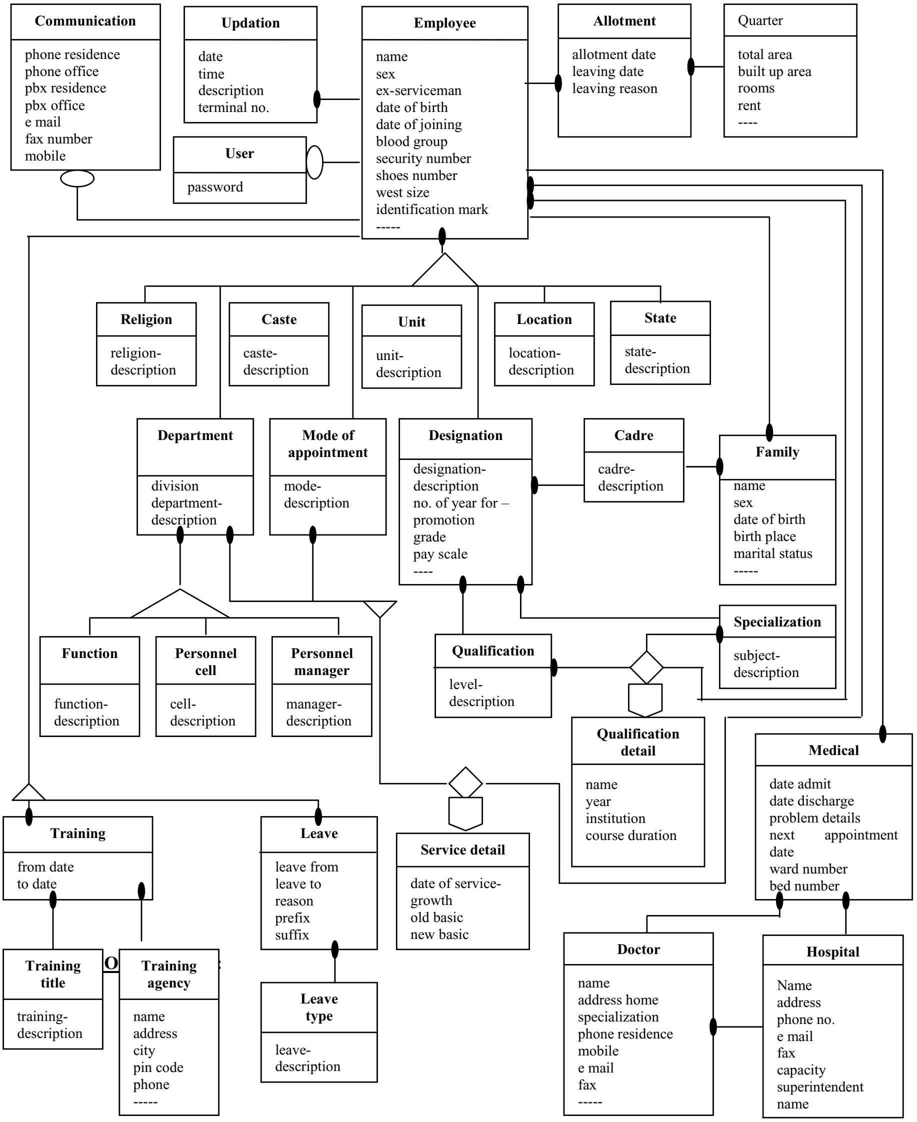


Fig. 4.1. Class Diagram for Human Resource Information System

5 Conclusion

In this work we have proposed a methodology, for the quantitative evaluation of software's understandability in the design phase. The core evaluation model and procedures are grounded in the LSP model and continuous preference logic. The attributes and metrics of understandability are measured from the hierarchical model proposed by Kumar and Gandhi [3]. The weights assigned for preferences are arbitrary and can be changed according to the requirement. I have found that the understandability of design [3] came out to be 85.79 which means that the system will be easy to understand.

The method is suitable for comparing alternative designs of a system for understandability aspect. This will help choose a design that is most suited for understanding especially when the software has been deployed.

References

1. Valdaliso, C., Eljabiri, O., Deek, F.P.: Factors Influencing Design Quality and Assurance in Software Development: An EmpiricalStudy. In: Electronic Proceedings of the First International Workshop on Model-based Requirements Engineering (MBRE 2001), San Diego, California, pp. 78–82 (2001)
2. Dujmovic, J.J.: A Method for Evaluation and Selection of Complex Hardware and Software Systems. In: Proceedings of the 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise CS, CMG 1996, vol. 1, pp. 368–378 (1996)
3. Kumar, M., Gandhi, S.K.: Object-Oriented Modeling Design Approach to General Human Resource System. Journal of MACT 2, 34–35 (2003-2004)
4. Kumar, M., Soni, D.: Observations on Object-Oriented Design Assessment and Evolving New Model. In: Proc of the National Conference on Software Engineering, pp. 161–164 (2007)
5. McCall, J.A., Richards, R.K., Walters, G.F.: Factors in Software Quality, National Tech. Information Service, Springfield, Va, vols. 1,2, and 3 (1977), AD/A-049-014/015/055
6. Boehm, B.W.: Characteristics of Software Quality. TRW Inc. (1978)
7. Software Product Evaluation - Quality Characteristics and Guidelines for Their Use, ISO/IEC Standard ISO-9126 (1991), `http://www.cse.dcu.ie/essiscope/sm2/9126ref.html`
8. Olague, H.M., Etzkorn, L.H., Messimer, S.L., Delugach, H.S.: An Empirical Validation of Object-Oriented Class Complexity Metrics and their Ability to Predict Error-prone Classes in Highly Iterative, or Agile Software: a Case Study. Journal of Software Maintenance 20(3), 171–197 (2008)
9. Chidamber, S.R., Kemerer, C.F.: Towards a Metric Suit for Object-Oriented Design. In: Proc. of Sixth Object-Oriented Programming Systems, Languages and Applications, pp. 197–211 (1991)
10. Chidamber, S.R., Kemerer, C.F.: A Metrics Suite For Object-Oriented Design. IEEE Trans. Software Eng. 20(6), 476–493 (1994)
11. Basili, V., Briand, L., Melo, W.: A Validation of Object-Oriented Design Metrics as Quality Indicators. IEEE Transactions of Software Engineering 22(10), 751–761 (1996)
12. Elish, M.O., Elish, K.O.: Application of TreeNet in Predicting Object-Oriented Software Maintainability: a Comparative Study. In: Proc. of European Conference on Software Maintenance and Reengineering (CSMR 2009), March 24-27, pp. 69–78 (2009)
13. Dormey, G.R.: A Model for Software Product Quality. IEEE Trans. Software Eng. 21(2), 146–162 (1995)
14. Dormey, G.R.: Cornering the Chimera. IEEE Software 13(1), 33–43 (1996)
15. Bansiya, J., Davis, C.G.: A Hierarchical Model for Object-Oriented Design Quality Assessment. IEEE Transactions on Software Engineering 28(1), 4–17 (2002)

16. Keller, R.K., Cockburn, A.: Object-Oriented Design Quality. In: OOPSLA, Workshop#12, Atlanta, Georgia, pp. 63–67 (1997)
17. Antonellis, P., Antoniou, D., Kanellopoulos, Y., Makris, C., Theodoridis, E., Tjortjis, C., Tsirakis, N.: A Data Mining Methodology for Evaluating Maintainability According to ISO/IEC-9126 Software Engineering Product Quality Standard. In: Proc. 11th IEEE Conference on Software Maintenance and Reengineering (CSMR 2007), March 21-23, pp. 35–42 (2007)
18. Fenton, N.E., Pfleeger, S.L.: Software Metrics: a Rigorous and Practical Approach, 2nd edn. PWS Publishing Company (1997)
19. Gao, J., Ming-Chih, S.: A Component Testability Model for Verification and Measurement. In: Proceedings of the 29th Annual International Computer Software and Applications Conference, pp. 211–218. IEEE Computer Society (2005)
20. Jimenez, G., Taj, S., Weaver, J.: Design for Testability. In: The Proceedings of the 9th Annual NCIIA Conference (2005)
21. Olsina, L.S.: Web-site Quality Evaluation Method: a case Study on Museums. In: ICSE 1999 – 2nd Workshop on Software Engineering over the Internet (1999)
22. Jacob, B., Niklas, L., Waldermarsson, P.: Relative Indicators for Success in software development. In: Department of Communication Systems. Lund University (2001)

Safety Analysis of Automatic Door Operation for Metro Train: A Case Study

Ajeet Kumar Pandey[*], Srinivas Panchangam, and Jessy George Smith

Cognizant Technology Solution, Hyderabad, India
{ajeet.kumar3,srinivas.panchangam,jessy.smith}@cognizant.com

Abstract. Transportation industries are growing not only in volume but in technology as well. To keep pace with changing business paradigms, automotive manufactures needs to use latest information technology and tools to make the transportation system economically viable, safe and reliable. Safety is the most important concern for today's railway system. Various subsystems of modern rail are safety critical and could result in loss of life, significant property damage or damage to the environment, if failure occurs. This paper presents the systematic approach to counter the risk in such system by analyzing the failure mode and its effect. The automatic door operation subsystem which forms one of the major safety critical systems in metro train is discussed along with a case study by analyzing various failure modes and its effect. Analysis processes as well as the significance of different metrics are also elaborated.

Keywords: reliability, safety, failure mode and effect analysis (FMEA), risk, risk priority number (RPN), automatic door operation (ADO).

1 Introduction

The transportation industry today has to be on the move, constantly, in more ways than any other industry. It has to deal with the increasing demands of customers and suppliers, while simultaneously trying to optimize the entire business operation at minimum cost. To keep pace with changing business paradigms, transporters need to use information technology, not merely as an enabler of operations but as a strategic driver and critical business tool. Railway transportation is more energy efficient and economical than the road transportation. The railways have always been ecologically safe with much less atmospheric pollution, compared to aircrafts and motor vehicles. The Railways have performed the twin tasks of providing adequate transport for industrial sustenance and growth and ensuring cheap, reliable and safe transportation for the population. Modern rails are now using many safety instrumented system (SIS) for handling safety critical functionalities. SIS implements the required safety functions necessary to achieve or maintain a safe state for some equipment.

Safety critical systems are those systems whose failure could result in loss of life, significant property damage or damage to the environment. There are well known

[*] Corresponding author.

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 962–974, 2013.

examples in application areas as such as railways, aircraft flight control, weapons and nuclear systems. Many modern information systems are becoming safety-critical in a general sense because financial loss and even loss of life can result from their failure. There are plenty of definitions of the term safety-critical systems but the intuitive notion actually works quite well. The concern both intuitively and formally is with the consequences of failure. If the failure of a system could lead to consequences that are determined to be unacceptable, then the system is safety critical. In essence, a system is safety-critical when we depend on it for our well being. Safety is the most important challenge for railway companies worldwide. A great deal of attention and effort has been paid to making railway operations safe. Despite these best efforts, accidents still occur, shaking people's faith in safety. And each time an accident occurs, further safety measures are taken. Today's railway safety is based on the many bitter experiences of the past. Railways are deeply rooted in society and people's consciousness worldwide and they are also strongly influenced by the each nation's social, cultural and geographical climate.

Several systems of locomotive/railways have gained importance in terms of safety measures. For example, through the last decade's door systems have developed tremendously. Safety and reliability are the key points in this development. Accurate controlling and checking of this safety related component are vital for reliable operation, making the door control unit the 'brain' of the door system. Door control units control door opening / closing so that passengers can safely get in and out of trains. Doors come in all sizes and shapes, different power systems, controls and door types. This paper presents a case study of Automatic Door Operations (ADO).

Rest of the paper is organized as follows: Section 2 presents the backgrounds behind the work. Sections 3 and 4 discuss literature surveys on safety and FMEA. Section 5 gives the brief idea about the automatic door operation (ADO), and contains a case study on ADO FMEA, whereas conclusions are presented in Section 6.

2 Backgrounds

2.1 Safety Critical System: Quality, Reliability and Safety

Safety-critical systems are those systems whose failure could result in loss of life, significant property damage, or damage to the environment. It is the system where human safety is dependent upon the correct operation of the system. Safety must be considered as whole system, including hardware , software, and other E/E/PE systems. Quality, reliability and safety issues must be considered with high importance in safety critical system.

Although the terms quality and reliability are often used interchangeably, there is a difference between these two disciplines. Reliability is the probability of the system meeting adequate performance for specified period of time under specified use condition. Reliability is concerned with the performance of a product over its entire lifetime; quality control is concerned with the performance of a product at one point in time, usually during the manufacturing process. Moreover, a close relationship also exists among the terms quality, reliability and safety especially in the context of software controlled product. Quality is the degree to which the systems meet its laid down specification. Reliability is a dynamic measure and varies with time. Qualitative measure is not sufficient for making engineering decision and therefore a quantitative, reliability measure is required.

Safety and reliability are often equated in the software context, but the conflicts between these two are growing to separate them [1]. Safety is the probability that the conditions that can lead to a mishap do not occur, whether or not the intended function is performed [6]. According to MIL-STD 882B, safety is defined as "freedom from those conditions that can cause death, injury, occupational illness, or damage to or loss of equipment or property. A system can be defined in two ways: what it is supposed to do and what it is not supposed to do. Reliability focuses on what the system is supposed to do while safety focuses on what system is not supposed to do. In general, reliability requirements are concerned with making a system failure free, whereas safety requirements are concerned with making it mishap free. Reliability focuses on every failure; whereas in safety only the dangerous failures are considered. Safety may decrease reliability and availability e.g. diagnostics and shutdown mechanisms.

2.2 Safety Critical System in Rail Transportation

The complexity of transport system is growing incredibly fast, thus Safety Critical Systems in the transport domain are becoming increasingly complex, not only in scale, but also the underlying technology. The railway industry is a leader in the development of safety critical systems. Modern rail transport systems contain a diverse combination of computers controlling non-critical functions such as entertainment systems and cabin lights, as well as safety critical systems such as track/train transmission, speed controller, and level crossing controller.

Now a day's transportation systems are using electronics for controlling various subsystems that was earlier controlled mechanically or manually. While these new electronic control and monitoring systems offers many benefits; in order to assure safe operations, regulations mandate that such systems comply with industry standards for hardware and software development and are thoroughly tested and documented. In rail transportation, as more electronic systems come into play, it becomes necessary to do whatever is possible to assure correct operation of these advanced systems.

3 Literature Survey

System safety is a sub discipline of system engineering that applies scientific, management, and engineering principle to ensure adequate safety, throughout the system life cycle, within the constraints of operational effectiveness, time and cost [1]. The objective of system safety is to identify, eliminate or control, and document system hazards in order to prevent any unsafe situations. Safety analysis is regarded as an initial investment by many researchers and industry professionals to save the future losses that would result from the potential mishaps. As a result of this, various hazard analysis techniques [2] for system safety have been developed such as Preliminary Hazard Analysis (PHA), Fault Tree Analysis (FTA), Event Tree Analysis (ETA), Failure Mode and Effect Analysis (FMEA), Markov Analysis, Common Cause Failure Analysis, and HAZOP Analysis.

Each of these techniques has some advantage and disadvantage in certain circumstances. Identification of hazards may utilize more than one technique as one particular hazard analysis may not be able to identify all the hazards within a system. Many researchers have tried to combine the advantages of FMEA and FTA for the safety analysis of the systems. FMEA can be developed as a preparatory activity to fault tree construction [3]. Combining bottom-up FMEA with the top-down FTA, is much effective in understanding underlying combination of circumstances that enable a failure mode to occur, as well as the likelihood of the identified failure mode [4].

On reviewing literature, it is found that FMEA and FTA is the most widely used safety analysis techchiqus. FMEA is a design analysis method that explores the effects of possible software failure modes on the system. There are two types of FMEA for embedded control systems: system software FMEA and detailed software FMEA [5]. System software FMEA can be used to evaluate the effectiveness of the software architecture without all the work required for detailed software FMEA. The detailed software FMEA validates that the software has been constructed to achieve the specified safety requirements. Detailed software FMEA is similar to component level hardware FMEA. However, the analysis is lengthy and labor intensive and also the results are not available until late in the development process. In fact the detailed software FMEA is often cost effective only for systems with limited hardware integrity.

4 Failure Mode and Effect Analysis

Failure Modes and Effects Analysis may have the various activities such as describe product or process, define functions, identify potential failure modes, describe effects of failures, determine causes, direction methods or current controls, calculate risks, take action and assess results. The FMEA process evaluates the overall impact of each and every component failure mode. The primary FMEA goal is to determine the effect on system reliability from component failure; however the technique can be extended to determine the effect on safety. Input data for the analysis include detailed hardware / function design information. Design data may be in the form of design concept, the operational concept, and major components planned for use in the system and major system functions. Table 1 lists the inputs, processes and outputs for conducting the FMEA.

Sources for this information include design specifications, sketches, drawings, schematics, functional block diagram (FBD) or reliability block diagram (RBD). Input data also includes known failure modes for components and failure rates for the failure modes. FMEA output information includes identification of failure modes in the system under analysis, evaluation of the failure effects, identification of hazards, and identification of system critical item in the form of a critical item list.

Table 1. FMEA Input, Process & Output

Input	FMEA Process	Output
Design Knowledge	Evaluate design	Failure modes
Failure Knowledge	Identify potential failure modes	Consequences
Failure Mode Type	Evaluate effects of failure modes	RPN
Failure Rate	Document Process	Reliability Prediction
Design Knowledge	Evaluate design	Critical Item List (CIL)

The FMEA process begins by identifying "failure modes", i.e. the ways a product, service or process could fail. A project team examines every element of a service, starting from the inputs and working through to the output delivered to the customer. At each step, the team asks "what could go wrong here?" Additionally they find out the probability of such failure (occurrence), the damage it will inflict (severity), should it actually fail and the likelihood of finding out (detectability) such failures before final delivery. These three parameters are ranked on 1-10 scale and product of these three is termed as Risk Priority Number (RPN). RPN can be used as safety indicator to prioritize the control actions.

4.1 FMEA Process

FMEA can provide an analytical basis, when dealing with potential failure modes and their associated causes. When considering possible failures in a design – like safety, cost, performance, quality and reliability – an engineer can get a lot of information about how to alter the development/manufacturing process, in order to avoid these failures. A typical FMEA process is shown in Figure 1. In general, FMEA process involves the following steps:

i. Define the system to be analyzed.

ii. Identify specific design requirements that are to be verified by the FMEA.

iii. Define ground rules and assumptions on which the analysis is based.

iv. Obtain or construct functional and reliability block diagrams.

v. Identify failure modes, effects, severity, and other pertinent information on the worksheet.

vi. Evaluate the severity of failures effect in accordance with the prescribed severity categories.

4.2 FMEA Worksheet

It is recommended to perform the FMEA using a form or worksheet to provide analysis structure, consistency, and documentation. The specific format of the analysis worksheet is not critical. Typically matrix, columnar or text-type forms are utilized to help maintain focus and structure in the analysis.

An FMEA that supports system safety and hazard analysis should contain the information, as a minimum are: Failure Mode, System Effect of failure mode, System-level hazards resulting from failure, Mishap effect of hazards, Failure mode and / or hazard causal factors, How the failure mode can be detected, Recommendations (such as safety requirements / guidelines that can be applied), and the risk presented by the identified hazard. The format of the FMEA worksheet may be determined by the customer, the system safety group, the safety manager, or the reliability / safety analyst performing the analysis. In the present study, a generalized FMEA worksheet has been used as shown in Table 2.

Table 2. FMEA Worksheet

Sl. No	Component	Failure Modes	Effects	Mitigation
1.	Motor	Open Circuit / Short Circuit	Door Stuck	Hardware failure needs to check mechanical parts.
2.	…	…	…	…
3.	…	…	…	…

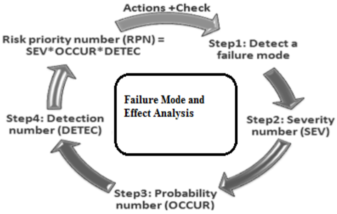

Actions +Check
Risk priority number (RPN) = SEV*OCCUR*DETEC
Step1: Detect a failure mode
Failure Mode and Effect Analysis
Step2: Severity number (SEV)
Step3: Probability number (OCCUR)
Step4: Detection number (DETEC)

Fig. 1. FMEA Process

4.3 Assessing Risk Priority Number (RPN)

The Risk Priority Number (RPN) is calculated for analyzing the risk associated with potential problems identified during a Failure Mode and Effects Analysis (FMEA). After identifying the potential failure modes; the RPNs are derived using past experience and engineering judgment to rate each potential factor according to three rating scales: *Severity, Occurrence, and Detectability.*

Severity (S) – Severity is a numerical subjective estimate of how severe the customer or end user will perceive the EFFECT of a failure.
Occurrence (O) – Occurrence or sometimes termed likelihood is a numerical subjective estimate of the likelihood that the cause, if it occurs, will produce the failure mode and its particular effect.

$$Criticality = Severity\ (S) * Occurrence\ (O)$$

Detection (D) – Detection is sometimes termed effectiveness. It is a numerical subjective estimate of the effectiveness of the controls to prevent or detect the cause or failure mode before the failure reaches the customer. The assumption is that the cause has occurred.

Assessing Risk – After the ratings have been assigned, the RPN for each issue is calculated by multiplying Severity, Occurrence, and Detection as:

$$RPN = Severity\ (S) * Occurrence\ (O) * Detection\ (D)$$

Rating scales usually range from 1 to 5 or from 1 to 10, with the higher number representing the higher seriousness or risk. The specific rating descriptions and criteria are defined by the organization or the analysis team to fit the products or processes that are being analyzed. Table 3 shows a generic five point scale for severity.

Table 3. Severity Scale

Rating	Description	Criteria
1	Very Low / None	Minor Nuisance
2	Low / Minor	Product operable at reduced performance
3	Moderate / Significant	Gradual performance degradation
4	High	Loss of Function
5	Very High / Catastrophic	Safety related Catastrophic failures

Larger RPN values normally indicate more critical failure modes but not always. For example, consider the three situations of Case-1where the RPNs are identical, but clearly the second situation would warrant the most attention. In general, any failure mode that has an effect resulting in a severity 9 or 10 would have top priority. Severity is given the most weight when assessing risk. Next, the Severity and Occurrence (S x O) combination would be considered; since this is effect represents the Criticality. Consider Case-2, situation #1 is most critical even though it has the lowest RPN value, than #2, and then #3. Here, the failure modes with the lowest RPN values are actually the most critical. One should be very careful when establishing the "threshold values" for RPNs when assessing risk.

Case-1

S	O	D	RPN
2	10	10	200
10	10	2	200
10	2	10	200

Case-2

S	O	D	RPN
10	2	2	40
3	10	2	60
2	5	10	100

5 Case Study- Automatic Door Operations: ADO System

A sliding door is a type of door which opens horizontally by sliding. A sliding door operator is a device that operates a sliding door for pedestrian use. It opens the door automatically, waits, and then closes it. Automatic sliding door is an intelligent application of advanced microcomputer and mechanical design, to meet the requirements for variety of construction, all sectors of the required automatic doors, with advantages of safe, reliable~~,~~ and long lifetime. It is being widely used in hotels, restaurants, railway stations, office buildings, supermarkets, major shopping malls and other places.

Sliding door operator will open/reopens the door as per the specifications. However, most operators use sensors to prevent the door from coming into contact with a user in the first place. The simplest sensor is a light beam across the opening. An obstacle in the path of the closing door breaks the beam, indicating its presence. Infrared and radar safety sensors are also used commonly. These are additional security methods used for the cases where an object cannot be detected by safety beam. The BLDC motor signals when an object is sensed and then the processor opens the door leafs.

Once the automatic door system is introduced, safety would automatically become foolproof for passengers, sources say, adding that the railways incur substantial expenditure for manufacturing automatic door coaches with the help of new technology.

5.1 Automatic Door Control Unit

Automatic door control units is a vital unit of sliding door and responsible for opening/closing of door safely for passengers to get in and out of trains. It consists of various modules such as Power Supply, CPU, FPGA, Vital Input Section, Vital Output Section, and general Input / Output unit. Structure of the door control unit is shown in Figure 2.

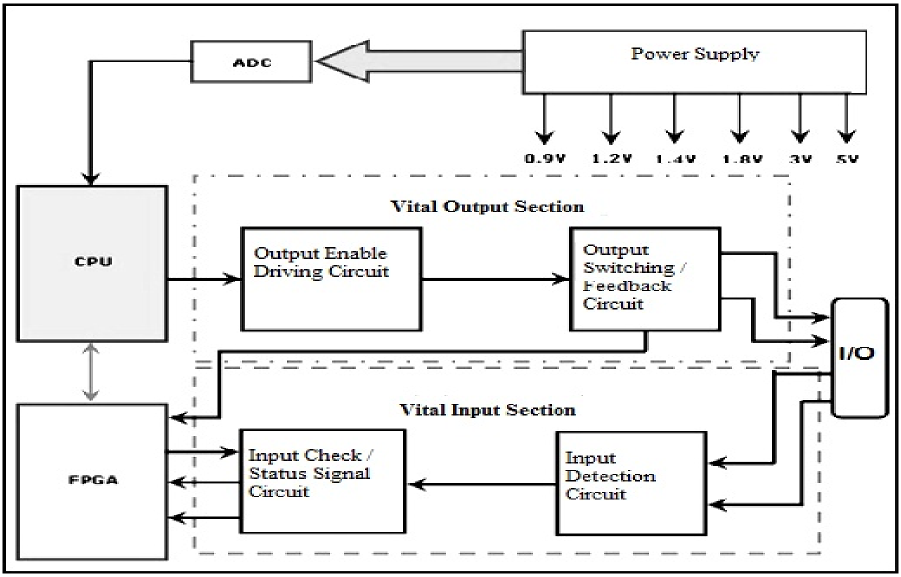


Fig. 2. Architecture of Door Control Unit

Table 4. System Level FMEA

#	Component	Failure Modes	Effects	Mitigation
1	Door Open / Close	Door does not open / close when required	Door permanently will be in open / closed mode. Passengers cannot board or alight from the metro cab.	HW may be damage, need to check the mechanical failures.
		Open / Close when not required	Door open / close when train is in running mode.	HW may be damage, need to check the mechanical failures.
		Door stuck in open / close position	Door always is in open or closed state. Passengers cannot board or alight from the metro cab.	HW may be damage, need to check the mechanical failures.
2	Motor Failure	Motor not working	Door will be in open or closed state. Passengers cannot board or alight from the metro cab.	HW may be damage; Motor conditions needs to be checked.
		Motor halts in between	Door will jam. Passengers cannot board or alight from the metro cab.	HW may be damage; Motor conditions needs to be checked.
		Motor speed is High	Rapid closure of Door. Passenger cannot board or alight.	HW may be damage; Motor interfacing unit needs to be verified.
		Motor speed is low	Slow closure of Door. Door will remain open in train running mode.	HW may be damage; Motor interfacing unit needs to be verified.
3	Power Supply	High Voltage	Door will remain open / close. Passengers cannot board or alight from the metro cab.	HW component may get damaged due to high voltage. Power supply check should take care whenever high voltage present it should cut-off the power supply.
		Low Voltage	Door will not open / close. Passengers cannot board or alight from the metro cab.	Power supply check should take care whenever low voltage present it should cut-off the power supply.
		Variable Voltage	Door will not Open / close. Passengers cannot board or alight from the metro cab.	HW component may get damaged due to unstable voltage. Power supply check should take care whenever unstable voltage present it should cut-off the power supply.
4	Sensor	Fast Detection	Door open / close before the input signal. Door will open immediately before required time.	SW should take care of the delay.
		Slow Detection	Door open / close response will be late. Door will open immediately late after required time.	SW should take care of the delay.
		No Detection	Door will not open / close.	HW may be damage need to check power source.
5	Timing	High Delay	Door will not open / close. Door will not function properly.	HW maybe damage. SW should take care of the delay.
		Low Delay	Door will not open / close. Door will not function properly.	HW maybe damage. SW should take care of the delay.

Input received from I/O section consists of two different logics (True / Compliment Logic), the system designed as redundant enough to capture and utilizes both true and false logic. Vital input section consists of Input Detection Circuit and Input Check Status Signal. Input Detection Circuit consists of an internal logic which helps to identify two different types logic. Input received from I/O circuit whether it is true or compliment is driven to the Signal Status check unit. Input Status check sends a status check signal to the FPGA. FPGA sends an acknowledgment to the status check circuitry about the health of the signal.CPU consists of logic which consists of the entire program which drives the complete unit. The required output from CPU given to the Output Enable Driving circuit, this circuit will drives the motor which helps to open / close the door automatically. Output switching units returns a feedback signals to the FPGA that helps to maintain the health of the unit.

An FMEA for the Door Control Unit is carried for a metro train in typical Indian Rail. For this a brain storming session with the various industry experts was performed. FMEA of Door Control Unit of different level i.e. system level FMEA, sub-system level FMEA and component level FMEA are listed in the Table 4, Table 5 and Table 6 respectively.

Table 5. Sub-System Level FMEA

#	Component	Failure Modes	Effects	Mitigation
1	CPU	Logic Execution Failure	System comes to unconfigured / safe state	Dual Channel SW commands CPU to Reboot
		Timing Error – Slow / Fast	System reacts variably	Watch Dog Timer takes care
		RAM Corrupted	System comes to unconfigured / safe state	Dual Channel SW commands CPU to Reboot
		Wrongly loaded	System may not work in proper state	User Manual verification
		ALU Error	System comes to unconfigured / safe state	Dual Channel SW commands CPU to Reboot
		Stack overflow	System reacts variably	Watch Dog Timer takes care
		Port Stuck high / low	System comes to unconfigured / safe state	Dual Channel SW commands CPU to Reboot
2	FPGA	Logic Execution Failure	Wrong status to CPU & consecutive checks detects	CPU commands to Reboot
		Timing Error – Slow / Fast	System reacts variably	Watch Dog Timer takes care
		Port Stuck high / low	Wrong status to CPU & consecutive checks detects	CPU commands to Reboot
3	Power Supply	High / Low Voltage	HW may damage	HW may damage and SW may call for Reboot
		Open	No effect	No Power
		Short	System reboots	System reboots
		Spikes	Zener Diode may damage and make the circuit safe	System reboots
4	Interface	Connected Wrongly	No effect	Dual Channel HW will take care
		Loose Connection	Wrong Status	Cannot be detected SW must take the feedback
		Breakage	No Status	Cannot be detected SW must take the feedback

Table 6. Component Level FMEA

Risk Priority Number (RPN) = Severity (S) * Occurrence (O) * Detection (D)

#	Component	Failure Modes	Effects	S	O	D	RPN	Defense / Mitigation	Safe
	Input Inductors	Open	Vital Input will always report as de-energized.	2	2	5	20	No need as de-energized state is safe state.	Yes
		Short	No immediate Safety effect. No error is reported.	1	2	5	10	No Defense necessary.	Yes
	Current Limiting Resistors	Open	Vital Input will always report as de-energized.	2	2	5	20	No need as de-energized state is safe state.	Yes
		Short	Energized: It fails & always shows as de-energized. De-energized: the input will show as de-energized.	8	2	10	160	The Vital Software will read the Vital Input hardware to get the status latch value containing the input states every 2ms. The Vital software will cross check the Vital Input state from Dual channel architecture for every execution cycle, to make sure they both are in agreement.	Yes
	Transorb / Zener Diode	Open	No immediate Safety effect. No error is reported.	1	2	5	10	No Defense necessary.	Yes
		Short	Vital Input Shorted so will always report as de-energized.	2	2	5	20	No need as de-energized state is safe state.	Yes
	Metal Oxide Semiconductor Field Effect Transistors (MOSFET)	Open	Loss of Vital Input status signal.	5	5	5	125	The Vital Software will turn off the input using the Vital Input test control. The results will be checked on every 3 sample sets, and require 2 in a row to fail before the test is deemed to have failed. If the test fails, the software will call shutdown.	Yes

Table 6. (*Continued*)

#	Component	Failure Modes	Effects	S	O	D	RPN	Defense / Mitigation	Safe
		Short Gate & Drain	The input reads de-energized. No failure will be reported. The check Opto-coupler will be shorted during the check and may be destroyed. The check will not work. The status check will not be able to de-energize the input. The check will always fail.	5	5	5	125	The Vital Software will read the Vital Input hardware to get the status latch value containing the input states every 2ms. The Vital software will cross check the Vital Input state from Dual channel architecture for every execution cycle, to make sure they both are in agreement. The Vital Software will turn off the input using the Vital Input test control. The results will be checked on every 3 sample sets, and require 2 in a row to fail before the test is deemed to have failed. If the test fails, the software will call shutdown.	Yes
		Short Source & Drain							
		Short Gate & Source							
	Input Check Opto Coupler	Open output transistor	MOSFET cannot be turned ON hence loss of the Vital Input check.	8	5	5	200	The Vital Software will read the Vital Input hardware to get the status latch value containing the input states every 2ms. The Vital software will cross check the Vital Input state from Dual channel architecture for every execution cycle, to make sure they both are in agreement. The Vital Software will turn off the input using the Vital Input test control. The results will be checked on every 3 sample sets, and require 2 in a row to fail before the test is deemed to have failed. If the test fails, the software will call shutdown.	Yes
		Shorted output transistor	Irrespective of FPGA command Vital Input status signal is always low.						
		Open Input LED	MOSFET cannot be turned ON hence loss of the Vital Input check.						
		Shorted Input LED	MOSFET cannot be turned ON hence loss of the Vital Input check.						
	Input Status Opto Coupler	Open output transistor	Irrespective of the Vital Input Check Control signal from FPGA the Vital Input status is always high.	8	5	5	200	The Vital Software will read the Vital Input hardware to get the status latch value containing the input states every 2ms. The Vital software will cross check the Vital Input state from Dual channel architecture for every execution cycle, to make sure they both are in agreement. The Vital Software will turn off the input using the Vital Input test control. The results will be checked on every 3 sample sets, and require 2 in a row to fail before the test is deemed to have failed. If the test fails, the software will call shutdown.	Yes
		Shorted output transistor	Irrespective of the Vital Input Check Control signal from FPGA the Vital Input status is always low.						
		Open Input LED	Irrespective of the Vital Input Check Control signal from FPGA the Vital Input status is always high.						
		Shorted Input LED	Irrespective of the Vital Input Check Control signal from FPGA the Vital Input status is always high.						

Table 6. (*Continued*)

#	Component	Failure Modes	Effects	S	O	D	RPN	Defense / Mitigation	Safe
	Output MOSFET	Open	Output cannot be enabled. No failure will be reported.	5	5	5	125	If the Vital Software determines that the output is energized from its reading of the Output feedback when it should be de-energized or vice versa then the Vital Software will put the Output into the failed state. If the Output is in failed state and the Output feedback still detects that the output is energized, the Vital Software shall Shutdown. If the Output is in failed state, and if no energy is detected, the Vital Software shall retry to turn Output ON after some determined time (15s), if the commanded state is energized. To protect against FPGA free-running the Output switch is implemented using a latch controlled from an output pin of the CPU. If the Vital Software determines that the Hardware circuit is generating an output when it's meant to be deenergized, the Vital Software can cut the output to the DC-DC converter by opening the Output Switch.	Yes
		Short Gate & Drain	Output always enabled.	5	5	5	125		
		Short Source & Drain							
		Short Gate & Source							
	Output Feedback Opto Coupler	Open output transistor	Irrespective of the Output the feedback signal to FPGA is always high.	8	5	5	200	If the Vital Software determines that the output is energized from its reading of the Output feedback when it should be de-energized or vice versa then the Vital Software will put the Output into the failed state. If the Output is in failed state and the Output feedback still detects that the output is energized, the Vital Software shall Shutdown. If the Output is in failed state, and if no energy is detected, the Vital Software shall retry to turn Output ON after some determined time (15s), if the commanded state is energized.	Yes
		Shorted output transistor	Irrespective of the Output the feedback signal to FPGA is always high.						
		Open Input LED	Irrespective of the Output the feedback signal to FPGA is always high.						
		Shorted Input LED	Irrespective of the Output the feedback signal to FPGA is always low.						
	Output Inductors	Open	Output cannot be enabled.	2	2	5	20	No need as de-energized state is safe state.	Yes
		Short	No immediate Safety effect. No error is reported.	1	2	5	10	No Defense necessary.	Yes

6 Conclusions

FMEA is one of the most effective safety analyses for achieving high quality system within specified timelines and budget constraints. A brain storming session was performed with the industry experts to conduct the FMEA of Automatic Door Operations. A case study of ADO module of the metro train is presented in this paper. The intension of this study was threefold: to create awareness for failures and their potential causes in order to prevent them, to point out how severe and critical potential failures may be and to show how they can be eliminated by offering solutions for different causes. As such, FMEA can be a time consuming process, but the results can be very worthwhile. However one has to obtain management support for the project, and the team leader's skills in keeping a team motivated and progressing through the project is essential to ensure the completion of a successful project.

Acknowledgement. Authors wish to thank Dr. Phanibhushan Sistu and Naveen Nair of Cognizant Technology Solution, Hyderabad, for their constant motivation, support and valuable feedback. Thanks are also due to Veerabhadra Prasad Gottimukkala and Vivek Diwanji for their constructive comments and suggestions which are very helpful in improving the paper.

References

1. Leveson, N.G.: Software Safety: Why, What, and How. Computing Surveys 18(2), 125–163 (1986)
2. Ericson, C.A.: Hazard Analysis Techniques for System Safety. Wiley Interscience, New Jersey (2005)
3. Maier, T.: FMEA and FTA to Support Safe Design of Embedded Software in Safety Critical Systems. In: 12th Annual CSR Workshop on Safety and Reliability of Software Based Systems, pp. 351–367 (1997)
4. Lutz, R.R., Woodhouse, R.M.: Bi-directional Analysis for Certification of Safety-Critical Software. In: Proceedings of International Software Assurance Certification Conference, Chantilly, VA, February 28–March 2 (1999)
5. Goddard, P.L.: Software FMEA Techniques. In: Proceedings of Annual Reliability and Maintainability Symposium, Los Angeles, CA, pp. 118–123 (2000)
6. Ericson, C.A.: Software and System Safety. In: Proceedings of the 5th International System Safety Conference 1, part 1, III-B-1–III-B-11 (1981)

A Generalized Model for Internet-Based Access Control Systems with Delegation Support

Utharn Buranasaksee, Kriengkrai Porkaew, and Umaporn Supasitthimethee

School of Information Technology
King Mongkut's University of Technology Thonburi, Bangkok
54500702@st.sit.kmutt.ac.th, {porkaew,umaporn}@sit.kmutt.ac.th

Abstract. In the web environment, web browsers use HTTP/HTTPS to communicate between users and web/application servers. However, many internet activities require interactions among three parties without compromising confidentiality. For example, an e-commerce transaction requires a buyer to authorize an e-commerce website to withdraw money from the buyer's bank account at an internet banking website. Although several existing works have been proposed to solve this problem, they are done in ad-hoc manners or lack of some important properties. This paper proposes a model, called PRA (Provider-Requestor-Authorizer), for generalizing three-party communication in the web-environment in order to identify desirable properties that can be used to measure the goodness of protocols for and classify them. We found that PRA model can generalize three-party communication protocols to a single model from conceptual level to implementation level.

Keywords: design, implementation, distributed access control, distributed system, classification, delegation.

1 Introduction

Since the invention of the Web, web-based applications have become ubiquitous because of its numerous advantages over traditional applications. One of the main reasons is that the users do not need any program other than web browsers which are commonplace. Unlike installing and using programs from unknown sources, with web browsers and HTTP/HTTPS, users do not need to worry about security. However, HTTP/HTTPS supports only communication between two parties. In some internet applications, three-party communications are necessary, such as e-commerce applications where there are a buyer, an e-commerce website and an internet banking website. These three parties need to communicate to one another in order to perform a buying transaction by delegating a right from one party to another without compromising confidentiality. For instance, the e-commerce website does not need to know about the bank account number of the buyer and the bank does not need to about what the buyer buys. Three-party communication can be built on top of two-party communication by having 3 pairs of two-party communication. Though there are many existing protocols and studies that support three-party communication [1-8], they are usually

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 975–988, 2013.

done in ad-hoc manners. Some protocols [9] even have security holes. Unlike the existing studies that focus on specific problems, this paper proposes a model that generalizes three-party communication called PRA model (Provider-Requestor-Authorizer) and identify desirable properties for three-party communication in the web environment so that it can be used to measure the goodness of such protocols and classify them. In addition, this paper also proposed implementation roles in PRA model which is an implementation level view of three-party communication in order to identify and generalize the process and required features of the protocol.

The remainder of this paper is organized as follow. In Section 2, we present the related work about the existing protocols. Then, the details of PRA model is discussed in Section 3. In Section 4, implementation role in PRA model are identified. We also analyzed the existing Internet-based access control protocol with delegation by PRA model and implementation roles in Section 5. Finally, the conclusions and future work are drawn in Section 6.

2 Related Work

There are many types of existing protocols that emulate three-party communication on the web. One type of these protocols is that a user is authorized by one system to access another system such as Single Sign On (SSO). SSO allows a user to use only one account to access multiple websites. In this type, OpenID [1] is one of the most popular SSO protocols available on the web. It is an open standard allowing anyone to be OpenID provider. Another protocol in SSO is Shibboleth [2]. Unlike OpenID, Shibboleth gained much attention in institution group since the protocol assertion has home institution feature. However, in enterprise, SAML [3] protocol was proposed to handle SSO and cross system authorization in XML format. SAML protocol was designed to be extensible to support many solutions, but the protocol does not gain much attention on the Internet due to its complexity.

Another type of three-party protocols is that a user is authorizing one system to perform a task at another system on a user behalf. Other than SSO, Gonzalez et al. [4] proposed an OAuth based protocol called Reverse OAuth. In Reverse OAuth, a user is authorized to access the resource on third party website while OAuth protocol [5, 6] allows a user to authorize one system to perform the tasks on another system without sharing user's credential. Due to its simplicity, OAuth is one of the most popular cross authorization protocols on the Internet. To support more in distributed system, Schiffman et al. [7] proposed DAuth protocol by introducing an agent to handle master token and sub-token to each sub-component in distributed system. In some situation, when there are complex delegation rules to delegate the user from many systems, xDAuth protocol [8] introduced a central agent called Delegation Service to handles predefined rules of delegation so as to reduce redundancy of delegation rules.

Though many of existing protocols are well-known and on the market, they were designed to focus on specific problems such as single sign on, cross system authorization, etc. None of the existing protocols were proposed to support the extension to solve general problem which could be different approach of authorization. For example, OAuth solves the problem of a user authorizing one system to access the resource on another system while OpenID solves the problem of a system authorizing a user to access another system. This study focuses on a generalized model called PRA model as well as its implementation guide that can be used to design generalized three-party communications protocol.

3 PRA Model

PRA model is a generalized model of three-party communication that proposes basic communication and required steps in order to make the interactions among the nodes complete without compromising confidentiality. In the Internet communication, when a node located in one domain requests the resources located on another domain, it would require third node that has accessible rights to grant the access. Therefore, we build Provider-Requestor-Authorizer model (PRA model) which based on the roles in the Internet communication.

According to PRA model, each scenario will have at least three roles involved in access control system with delegation support: Service Provider, Service Requestor and Service Authorizer. Each node performs the workflow according to its role. First, the Service Provider refers to the node that stores protected resource, or exposes some services. As we focus on Internet based access control model, Server Provider normally refers to a computer operator node since it electronically stores some resources on the Internet. Second, the Service Requestor refers to the node that receives the permission and accesses the resources or the services. Third, the Service Authorizer refers to the node that could physically or virtually own the protected resources located at the Service Provider node. According to our assumption, each node can either be human operator or computer operator. Therefore, there are 2^3 (2^N) = 8 scenarios that could occur in PRA model. Therefore, we identify the scenario according to Service Provider, Service Requestor, and Service Authorizer roles as the following:

3.1 CHH Scenario (Computer-Human-Human)

In CHH scenario, both Service Requestor and Service Authorizer are human operators while only Service Provider is computer operator. In this scenario, one user grants an access to another user on the same system. The grant typically is done in advance before an access, with no restriction and time limitation. Figure 1 shows that a user acting as Service Authorizer identifies himself, selects the sharing resource with privilege to another user. Then, another user acting as Service Requestor logs in, and requests for the shared resource afterwards. For example, a user shares a document to his colleague. After that, the user's colleague accesses the shared documents. The sharing usually allows user's colleague to edit the documents until the grant is revoked. The grant could also be adjusted according to privilege list depending on the resource type such as read only and read/write. There are many studies proposed the

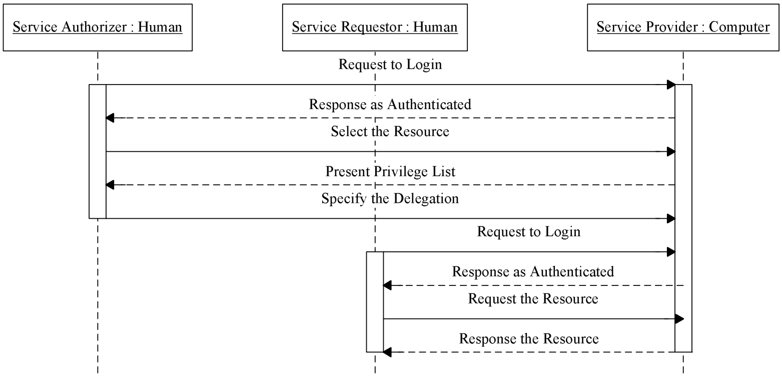


Fig. 1. CHH Scenario Basic Process

work on this scenario. For example, delegation in role based access control [10] and delegation in context-aware access control [11]. Furthermore, there are also many services in this scenario available on the Internet such as sharing documents on Google Docs [12], Facebook [13], etc. However, these previous studies were designed in ad-hoc manner. This scenario requires only Service Provider to have the interface for the user to share the resource and the storage to keep track the shares since the access control in this scenario does not have to pass secret information across the system.

3.2 CHC Scenario (Computer-Human-Computer)

In CHC scenario, both Service Provider and Service Authorizer are computer operators while Service Requestor is human operator. In this scenario, a user with the help of the server in one domain could get an access to protected resource in the server located on another domain. The process, as shown in Figure 2, starts by a user as Service Requestor requesting a resource on Service Provider. After that, a user is redirected to authenticate and to get authorized from Service Authorizer. Getting authorized, a user carries access token via URL redirection to Service Provider. Then, Service Provider verifies the authenticity of access token. If it is authentic, Service Provider returns a resource to a user. One of the situations that match this scenario is Single Sign On (SSO). SSO allows a user to use single credential as identification to access the resources on another systems. In enterprise, the applications could be run on many servers and the company maintain single identity provider to handle authentication process. When the user wants to access the application, he is redirected to identity provider to perform authentication. Furthermore, the logout status of the user needs to be propagated to recently used applications when the user logout. This scenario requires both Service Provider and Service Authorizer to have the same

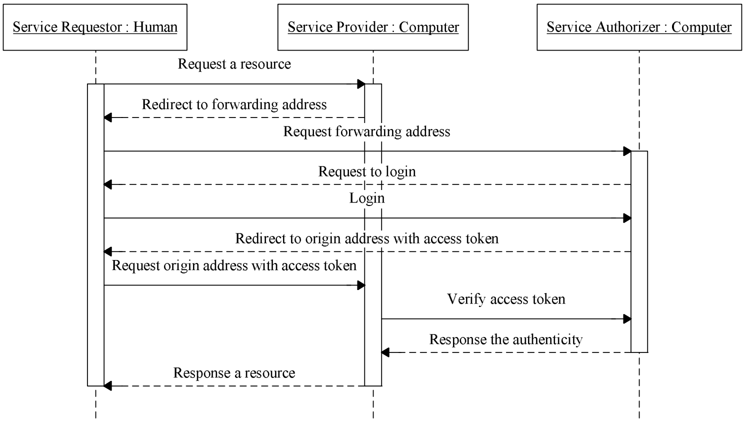

Service Requestor : Human
Service Provider : Computer
Service Authorizer : Computer
Request a resource
Redirect to forwarding address
Request forwarding address
Request to login
Login
Redirect to origin address with access token
Request origin address with access token
Verify access token
Response the authenticity
Response a resource

Fig. 2. CHC Scenario Basic Process

interface in order to communication to each other as well as Service Provider to trust Service Authorizer information. There are also many protocols available on the market such as OpenID [1], Microsoft Account [14], SAML [3], Shibboleth [2], etc.

3.3 CCH Scenario (Computer-Computer-Human)

In CCH scenario, both Service Provider and Service Requestor are computer operators while Service Authorizer is human operator. In this scenario, Service Requestor and Service Provider are likely to be located in different domains while the user could have an action in both parties so that the user has one computer perform on another computer on the user behalf. The process, as shown in Figure 3, starts by a user requesting Service Requestor to perform the task on Service Provider. Then, a user that carries requested task information from Service Requestor is redirected to Service Provider. Accessing Service Provider, a user is presented the login form. After logging in, a user is asked to authorize the task that Service Provider processes from the redirecting parameters. At this stage, a user could see requested operations and resources required by Service Requestor. After the operation is authorized, Service Provider returns access token and redirects a user to Service Requestor. Finally, Service Requestor uses that access token to access the resource on Service Provider, and renders the output to a user.

Since the authorizer in this scenario is human, the delegation is usually done in a real time rather than in advance. For example, asking to authenticate and authorize at Gmail website, Facebook fetches the user's contacts from Gmail and adds to Facebook friends immediately. From the nature of workflow in this scenario, the grant is given by once rather than giving always allowed. Therefore, the restriction of

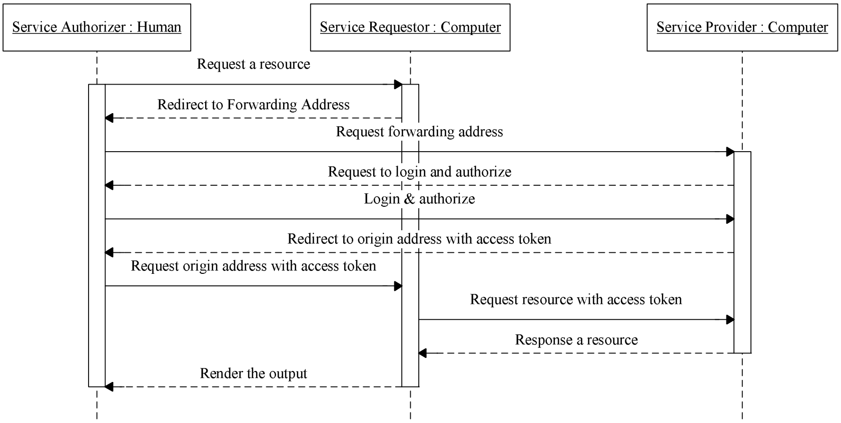

Service Authorizer : Human
Service Requestor : Computer
Service Provider : Computer
Request a resource
Redirect to Forwarding Address
Request forwarding address
Request to login and authorize
Login & authorize
Redirect to origin address with access token
Request origin address with access token
Request resource with access token
Response a resource
Render the output

Fig. 3. CCH Scenario Basic Process

the delegation such as time limit should be a necessary requirement. Another issue in this scenario generally involves confidentiality of the user. Unless a proper delegation transfer is deployed, Service Requestor could gain over the access since it could misuse the user's credential in Service Provider.

Therefore, this scenario requires Service Provider to have appropriate method that could differentiate computer operator from human direct access from browser in addition to on-the-fly authorization interface for Service Authorizer. Furthermore, Service Requestor and Service Provider should know each other which parameters to pass to the other via the user browser without compromising integrity and confidentiality. There are also many protocols proposed in this scenario such as OAuth [5, 6], Reverse OAuth [4], DAuth [7], xDAuth [8], etc.

3.4 CCC Scenario (Computer-Computer-Computer)

Unlike another scenario, CCC scenario has all the nodes including Service Provider, Service Requestor, and Service Authorizer performing in computer operators. Therefore, this scenario has no human interaction. Comparing to another scenarios, the basic process in CCC scenario could be applied from both CHC and CCH scenarios by changing human to computer operator instead. Therefore, CCC scenario could have two types of the basic process depending on what role the first party that initiates the communication takes place. If the first party takes Service Requestor role, the basic process matches CHC scenario. If the first party takes Service Authorizer role, the basic process matches CCH scenario. Even though CCC scenario could also be applied from CHH scenario, we do not consider this as three-party communication. If Service Authorizer grants the privilege beforehand, there would be only Service Requestor interacting with Service Provider. Unfortunately, at the time of writing this paper, we have not found any standard three-party communication protocol proposed since the communication is generally done in two-party since the application of three-party communication still is not clear.

Regarding another four categories in PRA model in which Service Provider is human, we are unaware of those scenarios and they are out of scope in this paper since we restrict that Service Provider roles would electronically store the protected resource online.

Though every parties in these scenarios perform the same concept of communication, the interactions of each party in these scenario does quite different when compare to that of another. For example, the interactions of Service Provider in CHC scenario is quite different when compare to that in CCH scenario. However, the protocols in each scenario also share some common implementation. Therefore, to generalize the process, we proposed implementation roles that could be used to generalize the function of three parties in PRA model.

4 Implementation Roles in PRA Model

By the nature of the protocols in CHH scenario, they generally involves the grant that is predefined in advance rather than immediate action and it does not involve cross system communication. For example, a user shares the document to another user. This kind of authorization is different from that in the other scenarios where the authorization is done on-the-fly and real time manner. Therefore, the generalization of the protocols in CHH scenarios is out of scope.

Implementation role refers to the role in PRA model that considers the implementation function such as redirecting a user, passing parameters, generating token, etc. rather than conceptual level such as authorizing node, requesting node, providing resource node. To illustrate this, comparing Figure 2 and Figure 3, the basic process in these two scenarios perform mostly the same pattern except that the ones that perform the same task are in different role. For instance, Service Provider in CHC scenario conceptually performs the same request/response as Service Requestor in CCH scenario does, and Service Authorizer in CHC scenario conceptually performs the same request/response as Service Provider in CCH scenario does. Therefore, we could see that, regardless which role the node takes in communication, each node has some specific interactions as well as information exchanging with another node as shown in Figure 4. To generalize the implementation requirements in the three scenarios, we proposed the implementation roles – Ticket Holder, Ticket Checker, and Ticket Master.

4.1 Ticket Holder

Ticket Holder refers to a user who first initiates the request to the resources or services. In this implementation role, a user would have nothing to implement on the server but hold the ticket via URL redirection between Ticket Checker and Ticket Master.

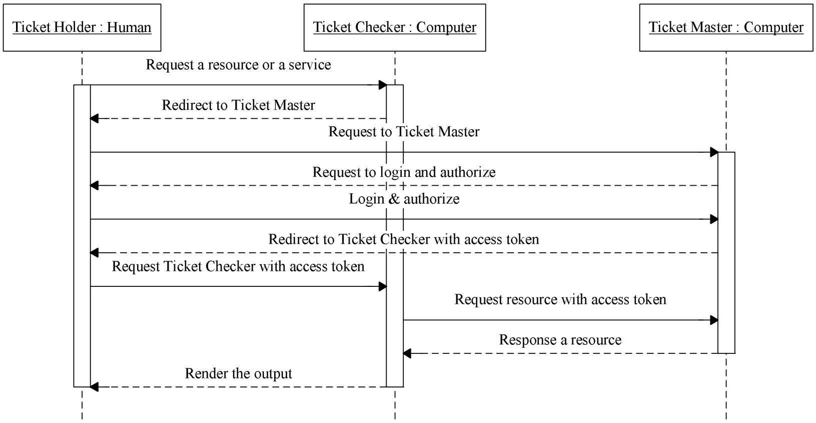

Ticket Holder : Human
Ticket Checker : Computer
Ticket Master : Computer
Request a resource or a service
Redirect to Ticket Master
Request to Ticket Master
Request to login and authorize
Login & authorize
Redirect to Ticket Checker with access token
Request Ticket Checker with access token
Request resource with access token
Response a resource
Render the output

Fig. 4. Implementation Role in PRA Model

4.2 Ticket Checker

Ticket Checker refers to the node that obtains information from another node by access token so that it could use this information to perform some tasks or provide some services. Unlike Service Requestor role that conceptually refers to the node that directly accesses the final output, Ticket Checker in implementation role concerns the node that forwards the user to obtain access token, and uses access token to access some information on another server, and response a finishing resource or service to a user. In addition, Ticket Checker is responsible to verify the authenticity and integrity of access token and information. Therefore, when implementing Ticket Checker, the Ticket Checker has to provide a user interface to authenticate and request for a resource or a service. Furthermore, it should also be capable to request, verify, and use the token to integrate the service from Ticket Master.

4.3 Ticket Master

Ticket Master refers to a node that produces information to another node by generating the tokens so that it could allow another node to obtain information and process the output to the user. Unlike Service Provider role that conceptually refers to the node that electronically stores protected resources and that protect resources would be accessed by Service Requestor, Ticket Master in implementation role refers to the node that generates token to another. After that, whenever Ticket Checker uses generated token to obtain the resource, Ticket Master returns the resource or verify the authenticity of an access. Therefore, when implementing Ticket Master, it should be capable to generate, verify the token as well as maintaining generated tokens. In addition to providing user interface for Ticket Holder to authenticate, Ticket Master should provide different approach that is suitable format for Ticket Checker to fetch the information when accessing with the generated token rather than returning a HTML document.

It is not only that both parties are capable to perform the required function, but also they should trust to each other. The trust might not be fully-trusted, but at least public-private key or shared secret key. For example, both parties could verify that the value in passing parameters have confidentiality, integrity and authenticity since Client could capture and change the value while redirecting. This could be done by deploying digital signature or encryption.

5 Example Analysis of Existing Protocols

We give the brief characterization of existing access control protocols, and explain the process of the protocol. Then, we apply PRA model into the protocol in order to categorize the protocol into suitable scenario. After that, we compare each protocol process to basic process in PRA model, and discuss how the protocol could be generalized to PRA model in implementation level using implementation roles. Due to space limitation, we pick some of the protocols in each type and analyze in this paper.

5.1 OpenID

OpenID is a decentralized open authentication protocol that allows users to identify themselves on any OpenID-enabled website. In OpenID context, there are three nodes which are OpenID provider, relying party, and user. The process of OpenID protocol,

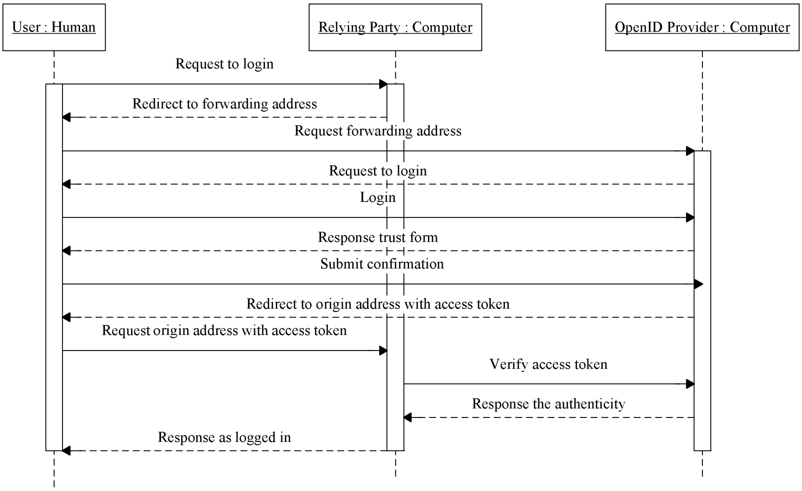

User : Human
Relying Party : Computer
OpenID Provider : Computer
Request to login
Redirect to forwarding address
Request forwarding address
Request to login
Login
Response trust form
Submit confirmation
Redirect to origin address with access token
Request origin address with access token
Verify access token
Response the authenticity
Response as logged in

Fig. 5. OpenID Protocol Process

as shown in Fig. 5, starts by a user wanting to access the resource on relying party by providing the OpenID account information. Then, relying party redirects a user to OpenID provider page. Then, a user gets authenticated and confirms the trust to relying party at OpenID provider. After that, relying party verifies if the account is not blacklisted, the user is allowed to access the resource on relying party based on OpenID account. Therefore, in OpenID protocol, OpenID acts as Service Authorizer that authorizes end user as Service Requestor to protected resources located at relying party which acts as Service Provider. Thus, OpenID matches CHC scenario.

Comparing to CHC scenario basic process, OpenID protocol does quite the same as CHC basic process does except that relying party does not only request for authenticity from OpenID provider, but also the user's account information. Furthermore, when compare to implementation roles, relying party and OpenID provider could take Ticket Checker and Ticket Master role accordingly while a user takes Ticket Holder role.

5.2 OAuth 2.0

OAuth 2.0 is one of the most popular web authorization protocols. OAuth protocol was proposed to allow delegation the resources without giving credential. In OAuth context, there are three nodes which are Service Provider, OAuth consumer, and the user. The process of OAuth protocol, as shown in Fig. 6, starts by OAuth consumer requesting the resource at service provider in exchange of request token and redirecting user to authorize. After the user authorizes the request token at service provider, the service provider returns access token to OAuth consumer so that OAuth consumer could use this access token to access protected resource afterwards.

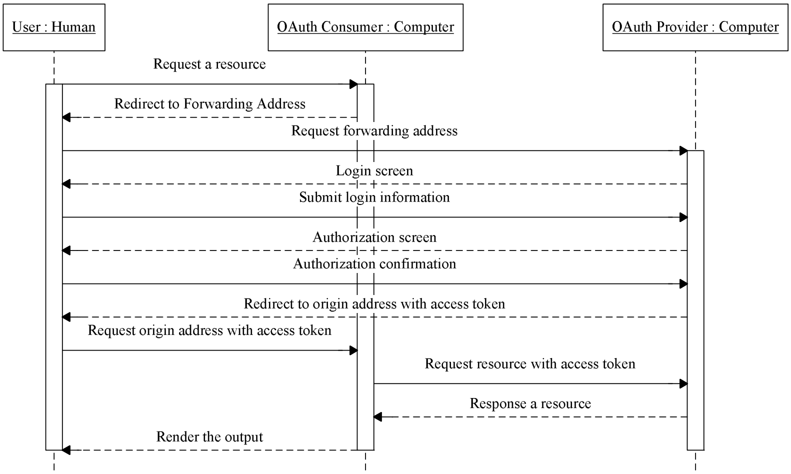

User : Human
OAuth Consumer : Computer
OAuth Provider : Computer
Request a resource
Redirect to Forwarding Address
Request forwarding address
Login screen
Submit login information
Authorization screen
Authorization confirmation
Redirect to origin address with access token
Request origin address with access token
Request resource with access token
Response a resource
Render the output

Fig. 6. OAuth Protocol Process

According to PRA model, the user acts as Service Authorizer that authorizes OAuth consumer as Service Requestor to access the resource at Service Provider. Therefore, OAuth protocol matches CCH scenario. However, when redirecting a user to OAuth provider, OAuth consumer also has to specify required permission, called scope, in addition to forwarding address and request token.

When applying implementation roles to OAuth, OAuth consumer could be implemented in Ticket Checker role and OAuth provider could be implemented in Ticket Master role. Furthermore, OAuth protocol process does the same as CCH scenario basic process does since there are only three parties.

5.3 Reverse OAuth

Reverse OAuth is a delegation access control protocol that based on OAuth protocol to support different requirement. Unlike OAuth which allows a user to grant one system to access another system, Reverse OAuth allows a system to grant a user to access another system on-the-fly using token. In Reverse OAuth, there are three nodes which are student, web tool, and LMS system. The process of Reverse OAuth, as shown in Fig. 7, starts by a student requesting a service at web tool via LMS system. Then, a student authenticates himself at web tool using request token generated from LMS system. Then, web tool asks LMS system if a student who uses this request token have privilege to access web tool. If so, a student could have an access to web tool. According to PRA model, LMS system acts as Service Authorizer that authorizes student as Service Requestor to access web tool as Service Provider. Therefore, Reverse OAuth matches CHC scenario in PRA model.

According to implementation role, a student as a human takes Ticket Holder role. LMS system does the job generating an access token to a student. Therefore, LMS system takes Ticket Master role while web tool that uses access token to render a finishing output to a student takes Ticket Checker role. Comparing to CHC scenario basic process, Reverse OAuth protocol has more complicate process several points. First, a student could not just directly access to the web tool, he needs to request to LMS system before requesting to web tool. This is because Gonzalez et al. designed the protocol to support controlling usage of the web tool such as time period, scope of usage. However, with CHC basic process, Ticket Master could still validate the right of the user before generating an access token to Ticket Holder anyway. Second, the authenticity and the authorization of access token are checked by Ticket Checker before passing to Ticket Holder. However, in CHC basic process, the authenticity and authorization of access token whenever Ticket Holder requests a service. This is just a design differentiation since PRA model could be adjusted to flexible enough.

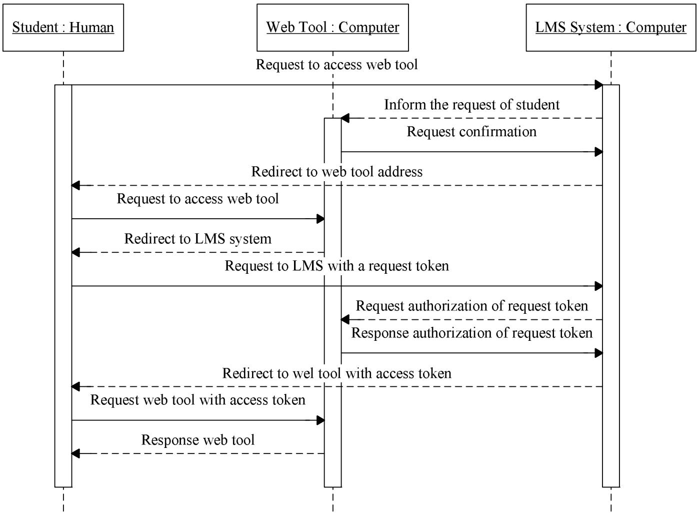

Student : Human
Web Tool : Computer
LMS System : Computer
Request to access web tool
Inform the request of student
Request confirmation
Redirect to web tool address
Request to access web tool
Redirect to LMS system
Request to LMS with a request token
Request authorization of request token
Response authorization of request token
Redirect to wel tool with access token
Request web tool with access token
Response web tool

Fig. 7. Reverse OAuth Protocol Process

5.4 DAuth

DAuth is an authorization delegation access control for distributed web application. The protocol assumes that one web application would have many sub-components and addresses the problem of fine-grained authorization on each sub-component by introducing a new component called DAuth Agent. Therefore, there are four types of node in DAuth protocol which are Server Provider, DAuth Agent, consumer components, and user. The process starts by, as shown in Fig. 8, a user requesting a resource to DAuth Agent. Then, DAuth agent requests for a request token from Service Provider and redirect a user to authenticate and authorize at Service Provider. After authorizing the request token, Service Provider generates an access token and passes it to DAuth Agent via a user. Obtaining a master access token, DAuth Agent acts as a proxy of consumer components by requesting access sub-token and transmits to consumer component. Whenever consumer component needs to access the resource at Service Provider, it just requests access sub-token from DAuth Agent.

According to PRA mode, the user acts as Service Authorizer that authorizes both DAuth Agent and consumer components to access the resource at Service Provider. Therefore, DAuth protocol matches CCH scenario. Since DAuth has 4 parties involved in the protocol, the process in the protocol is more complicated than that of CCH basic process. It could be called as additional features to basic protocol which allows DAuth Agent to handle distributed components.

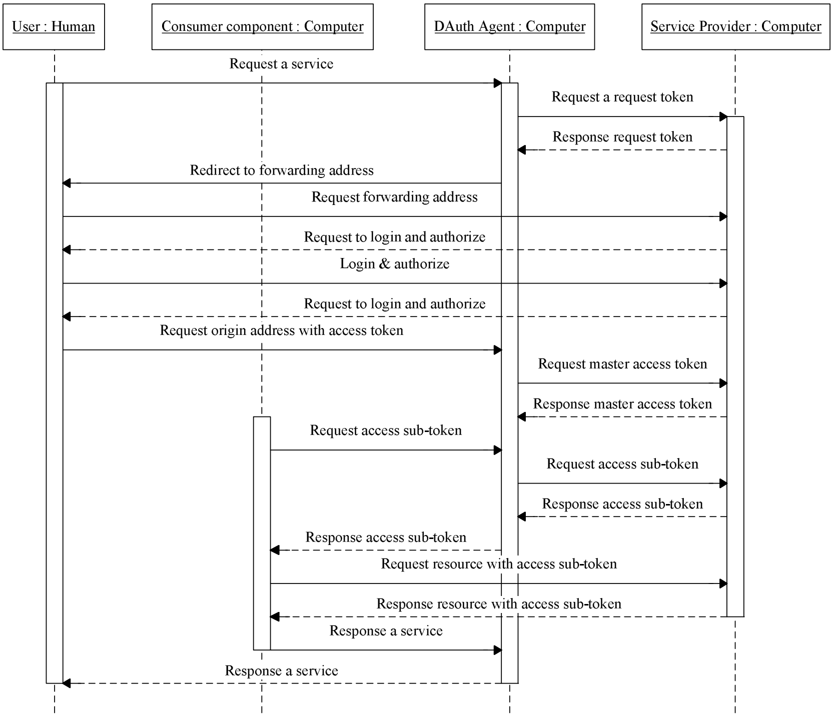

User : Human
Consumer component : Computer
DAuth Agent : Computer
Service Provider : Computer
Request a service
Request a request token
Response request token
Redirect to forwarding address
Request forwarding address
Request to login and authorize
Login & authorize
Request to login and authorize
Request origin address with access token
Request master access token
Response master access token
Request access sub-token
Request access sub-token
Response access sub-token
Response access sub-token
Request resource with access sub-token
Response resource with access sub-token
Response a service
Response a service

Fig. 8. DAuth Protocol Process

According to implementation roles, both consumer components and DAuth Agent that are located in the same domain operate by obtaining the token from Service Provider and using it to access the resource. Therefore, both of them act as Ticket Checker. Second, Service Provider that generates request token and access token to another party would take Ticket Master role. Finally, a user takes Ticket Holder role since the generated token is passed via URL redirection. This could be seen as CCH scenario basic protocol with sub token extension.

6 Conclusion and Future Work

We have proposed a generalized model for Internet-based access control systems, called PRA model, which built for three-party communication. The model categorizes access control into scenario. The access control protocols that fall into the same category have the same approach of handling authorization. After proposing PRA model, we found that CHC, CCH, and CCC scenarios could be generalized the process by introducing implementation role. According to implementation role in PRA model, the node that takes particular role would have specific guideline of implementation regardless which scenario the protocol is. After reviewing the existing protocols by listing the process of the protocol and applying both the concept and implementation roles of PRA model into each protocol, we found the common implementation of each protocol that could be categorized to the same category in PRA model. Furthermore, the existing protocols that fall into CHC, CCH, and CCC protocols could also be generalized to PRA model in terms of both concept and implementation.

Acknowledgement. This work was supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission and IBM Faculty Award.

References

1. OpenID Authentication 2.0, `http://openid.net/specs/openid-authentication-2_0.html` (accessed 30 June 2012)
2. Morgan, R.L., Cantor, S., Carmody, S., Hoehn, W., Klingenstein, K.: Federated Security: The Shibboleth Approach. In: EDUCAUSE Quarterly, vol. 27, pp. 12–17 (2004)
3. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0., `https://www.oasis-open.org/committees/download.php/35711/sstc-saml-core-errata-2.0-wd-06-diff.pdf` (Accessed 30 August 2012)
4. González, J.F., Rodríguez, M.C., Nistal, M.L., Rifón, L.A.: Reverse OAuth: A solution to achieve delegated authorizations in single sign-on e-learning systems. Computers & Security 28, 843–856 (2009)
5. OAuth Core 1.0a, `http://oauth.net/core/1.0a/` (accessed 30 June 2012)
6. The OAuth 2.0 Authorization Framework, `http://tools.ietf.org/html/draft-ietf-oauth-v2-30` (accessed 30, June 2012)
7. Schiffman, J., Xinwen, Z., Gibbs, S.: DAuth: Fine-Grained Authorization Delegation for Distributed Web Application Consumers. In: IEEE International Symposium on Policies for Distributed Systems and Networks (POLICY), pp. 95–102 (2010)
8. Alam, M., Zhang, X., Khan, K., Ali, G.: xDAuth: a scalable and lightweight framework for cross domain access control and delegation. In: Proceedings of the 16th ACM Symposium on Access Control Models and Technologies, SACMAT 2011, pp. 31–40. ACM, New York (2011)
9. OAuth 2.0 Threat Model and Security Considerations, `http://tools.ietf.org/html/draft-ietf-oauth-v2-threatmodel-07` (accessed 20 August 2012)
10. Crampton, J., Khambhammettu, H.: Delegation in Role-Based Access Control. In: Proceeding of the 11th European Symposium on Research in Computer Security, pp. 174–191 (2006)
11. Toninelli, A., Montanari, R., Kagal, L., Lassila, O.: A Semantic Context-Aware Access Control Framework for Secure Collaborations in Pervasive Computing Environments. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 473–486. Springer, Heidelberg (2006)
12. Google Docs, `http://www.google.com/google-d-s/b1.html` (accessed 30 August 2012)
13. Facebook, `http://www.facebook.com` (accessed 30 August 2012)
14. Microsoft account, `https://account.live.com/` (accessed 30 August 2012)

QoS Support Downlink for WiMAX Network

Pooja Gupta, Brijesh Kumar, and B.L. Raina

Lingaya's University, Fardabad, India
{poojagupta29,muskanbrijesh}@gmail.com, rbushan@rediffmail.com

Abstract. We develop new scheduling algorithms for the IEEE 802.16d based broadband wireless access system, in which radio resources of both time and frequency slots are dynamically shared by all users. Our objective is to provide a fair and efficient allocation to all the users to satisfy their quality of service. IEEE 802.16 based wireless access scheme(commonly known as WiMAX) is considered as one of the most promising wireless broadband access for communication networks in metropolitan areas today. Since this broadband wireless access system defines the concrete quality of service(QoS) requirement, a fair scheduling(FS) scheme is necessary to meet the QoS requirements. Many Scheduling schemes have been proposed earlier with the purpose of throughput optimazation and fairness enhancement. Here we present FS to derive its performance bounds. Our analysis demonstrates that FS support the delay requirement.This scheduler proposes a new scheduling scheme reflecting the delay requirement of rtps connections with respect to the various nrtps connections to achieve the optimal QoS requirement, without the excessive resource consumption since it 1) achieves low average as well as maximum delay for low-throughput applications 2) provides fairness regardless of variation in server capacity3) is computationally efficient.

Keywords: QoS, IEEE 802.16, WiMax, rtps, nrtps.

1 Introduction

IEEE 802.16[1] architecture includes one Base station (BS) and Multiple Subscriber Station(SS) Communication occurs in two directions: from BS to SS is called Downlink and from SS to BS is called Uplink. During downlink, BS broadcasts data to all subscribers and subscribers selects packets destined for it. IEEE 802.16[2] standard also known as worldwide interoperability for microwave access (WiMAX) defines two modes to share wireless medium: point-to-multipoint (PMP) mode and mesh mode. In the PMP mode, a base station (BS) serves several subscriber stations (SSs) registered to the BS. In IEEE 802.16, data transmission is on the fixed frame based. The frame is partitioned into the downlink subframe and the uplink subframe. The frame duration and the ratio between the downlink subframe and the uplink subframe are determined by the BS. In the PMP mode, the BS allocates bandwidth for uplink and downlink. The BS selects connections to be served on each frame duration. IEEE 802.16 defines four classes of service type such as unsolicited grant service (UGS), real-time polling service (rtPS), non-real-time polling service (nrtPS) and best

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 989–997, 2013.

effort (BE) service. Each service class has requirements to be met to serve the applications that belong to the category. The UGS is designed to serve the applications having stringent delay requirement, like voice over IP (VoIP). The rtPS is designed for the applications having the less or stringent delay requirement, like video or audio streaming service. The nrtPS does not have the delay requirement; however, it has the minimum reserved rate requirement. To satisfy these QoS requirements, we need a well-designed scheduling scheme. However, IEEE 802.16 specification does not describe the scheduling scheme, and it leaves the implementation of a scheduling scheme to device manufacturers' decision. The scheduling scheme plays an important role in the quality of service (QoS) provision. Many scheduling schemes have been proposed. An overview of scheduling schemes in wireless networks is presented in[3][4][5]. There are many papers suggesting scheduling schemes[6][7] to reflect the QoS requirement. The proportional fair scheduling has been introduced in [7][8].The concept of the proportional fair scheduling is widely accepted in scheduling design. In the current work we have proposed an alternate scheduling scheme based on proportional fairness. The scheduling parameters have been selected based on the number of connections in the network.

2 System Model

PMP mode and mesh mode are the two types of operating modes define for IEEE802.16. In the PMP mode SSs are geographically scattered around the BS. The performance of IEEE 802.16 in the PMP mode is verified in[8][9]. Our system model is based on a time-division-duplex (TDD) mode. The IEEE 802.16 frame structure is illustrated in Fig.1[2]. The downlink subframe starts with preamble followed by frame control header (FCH), downlink map (DL-MAP), uplink map (UL-MAP) messages and downlink burst data. The DLMAP message defines the start time, location, size and encoding type of the downlink burst data which will be transmitted to the SSs. Since the BS broadcasts the DLMAP message, every SS located within the service area decodes the DL-MAP message and searches the DL-MAP information elements (IEs) indicating the data bursts directed to that SS in the downlink subframe. After the transmit/receive transition gap (TTG), the uplink subframe follows the downlink subframe. IEEE 802.16 provides many advanced features like adaptive modulation coding (AMC), frame fragmentation and frame packing. In the current work, the focus is on the downlink scheduling scheme.

3 Multi User Scheduler of the MAC Layer

In this section, a multiuser scheduler is designed at the medium access control (MAC) layer. Delay requirement is taken into account in the scheduler design. The AMC, packet fragmentation and packet packing have not been considered. In case of the UGS traffic, the required bandwidth is reserved in advance. Hence, only rtPS, nrtPS and BE connections are focused in the design.

3.1 Proportional Fair Scheduling

The proportional fair scheduling [10] has shown an impressive guideline in scheduler design because it maximizes the total sum of each SS's utility. In the proportional fair scheduling, the metric for each connection is defined as follows

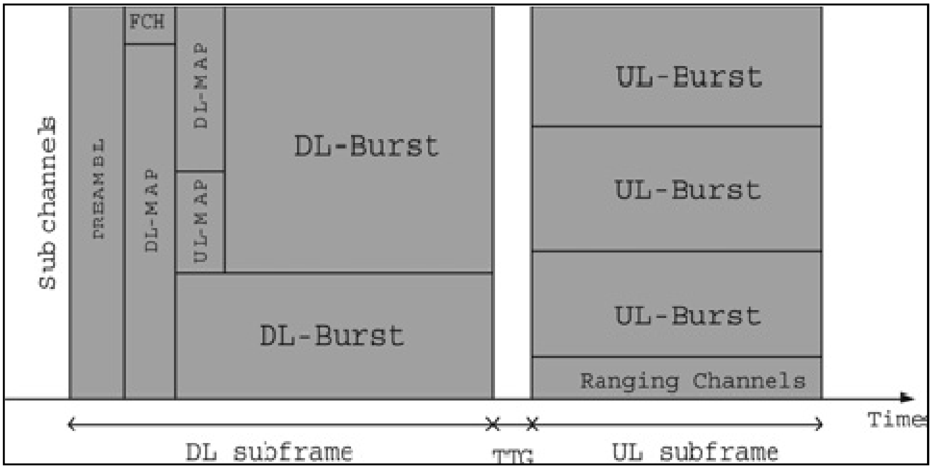


Fig. 1. IEEE 802.16[2] frame structure

$$\phi_i(t)=DRC_i(t)/R_i(t). \qquad (1)$$

where DRC_i [12] is the rate requested by the SS_i and R_i is the average rate received by the SS_i over a window of the appropriate size T_c [2][12]. The average rate R_i is updated as

$$R_i(t+1)=(1-1/T_c)R_i(t)+1/T_c \text{ * current transmission rate.} \qquad (2)$$

where T_c is the window size to be used in the moving average. The proportional fair scheduler selects the connection that has the highest metric value.

3.2 Proposed Fair Scheduling (FS)

In the proportional fair scheduling, the strict fairness is guaranteed, however the QoS requirement is not reflected. To the knowledge of authors normally various rtps connections for QoS have been discussed in the literature with regard to one specified nrtps connection. The present authors have generalized this concept by associating various parameters of x_i defined as (various) rtps connections to the parameter k_i associated to the (various) nrtps connections Thus, the general Fair Scheduling(FS) scheme is being introduced that satisfies the delay requirement.

The metric value of the rtPS connections with the delay requirement should be increased as the queuing delay increases because the scheduler selects the connection with the highest metric value with BE connections, because BE connections are in the lowest priority. For the above mentioned condition the equations for rtps,nrtps and BE are proposed by the authors in paper [2].Here we are generalizing the equation by proposing a new scheduling scheme based on the following metrics for rtPS, nrtPS and BE connections given as :

$$\Phi_{rt,i}(t)= 1/R_{rt,i}(t)+C(1+2/\pi{*}\arctan(|d|)). \quad \text{if } q_i >0 \text{ and } d \geq d_{min.} \tag{3}$$
$$=1/R_{rt,i}(t)+ C. \quad \text{if } q_i >0 \text{ and } d< d_{min.}$$
$$0 \quad \text{if } q_i =0$$

$$\Phi_{nrt,i}(t)= 1/R_{nrt,i}(t)+ C. \quad \text{if } q_i >0 \tag{4}$$
$$=0 \text{ if } q_i =0$$

$$\Phi_{BE,i}(t)= 1/R_{BE,i}(t). \quad \text{if } q_i >0 \tag{5}$$
$$=0 \quad \text{if } q_i =0$$

The parameter d is the queuing delay and C means the intensity of the delay requirement in the rtPS connection. The parameter d_{min} is the minimum delay that triggers the service differentiation between the rtPS connection and nrtPS connection, and q_i means the queue length of the connection i. Note that R_{rt}, R_{nrt} and R_{BE} are updated in the same manner as in the proportional fair scheduling, that is

$$R_{rt,i}(t+1)= (1-1/T_c)R_{rt,i}(t) + r/T_c \text{ if connection } i \text{ is scheduled.} \tag{6}$$
$$=(1-1/T_c) R_{rt,i}(t) \quad \text{otherwise}$$

where T_c is the window size to be used in the moving average and r is the current transmission rate requested by the SS. The long-term rate is the average sum of the previously scheduled transmission rates during the time window T_c, where the high T_c value means that the long-term rate changes slowly because the average is taken over many previous transmission rates. The long-term rate of a connection decreases exponentially before the connection is scheduled, and it increases when the connection is scheduled. We do not consider the AMC, so r is a constant. On every frame, the scheduler selects the connection that has the highest metric value. Owing to the delay requirement term in the rtPS metric, rtPS connections are served more frequently than other connections when the queuing delay increases.

3.3 Determination of Novel Parameters with Analysis

The scheduling ratio x as the average number of scheduling times for rtPS connection per k nrtPS scheduling has been defined. If rtPS and nrtPS connections are scheduled equally, the scheduling ratio x equals k, and if rtPS connection is scheduled more frequently than nrtPS connection, the scheduling ratio x becomes larger than k. The average scheduling interval in the rtPS connection is ((k+x)/k) frames because, on the average, k nrtPS is scheduled corresponding to the scheduling of x rtPS connections.

As a result of this, the average scheduling interval in nrtPS connection is (k+x) frames. At the steady state, the average long-term rates of rtPS and nrtPS connections at the scheduling instance are as follows:

$\overline{R_{rt}} = \overline{R_{rt}}(1-(1/T_c))^{(k+x)/x} + (r/Tc)$, at the steady state, we obtain

$$\overline{R_{rt}}=(r/T_c)/\ (1-(1-(1/T_c))^{(k+x)/x} \qquad (7)$$

Analogously, Since $\overline{R_{nrt}} = \overline{R_{nrt}}(1-(1/T_c))^{(k+x)} + (r/Tc)$ at the steady state, we obtain

$$\overline{R_{nrt}}=(r/T_c)/\ (1-(1-(1/T_c))^{(k+x)} \qquad (8)$$

We consider the same assumption as in[11] that the average metric value for each connection at the scheduling instance becomes similar to each other. Hence,

$$1/\overline{R_{rt}}(1-(1/T_c))^{(k+x)/x} + C(1+(2/\pi)\arctan(|d|)).$$
$$\approx 1/\overline{R_{nrt}}(1-(1/T_c))^{(k+x)} + C. \qquad (9)$$

From (7) and (8) , (9) can be written as

$$(1-(1-(1/T_c))^{(k+x)/x})*T_c/\ (r*(1-(1/T_c))^{(k+x)/x}) + C(1+(2/\pi)\arctan(d)).$$
$$\approx (1-(1-(1/T_c))^{(k+x)})*T_c/(\ r*(1-(1/T_c))^{(k+x)}) + C. \qquad (10)$$

Put $(1-1/T_c)=X$, $L=1+(2/\pi)\arctan(d)$, therefore from above equation we have

$$(1-(X)^{(k+x)/x}\ )*T_c/(r*(X^{(k+x)/x}) + C*L$$

$$(1-(X)^{(k+x)})*Tc\ /(\ r*(X^{(k+x)}) + C$$

$$\text{i.e. } C*(L-1)=(Tc/r)*((1-(X)^{(k+x)}/X^{(k+x)} - (1-(X)^{(k+x)/x})/X^{(k+x)/x}))$$

$$C*(2/\pi)*\tan^{-1} d=(Tc/r)*((X^{(k+x)/x} - X^{(k+x)})/X^{((x*x+k*x+k+x)/x)}) \qquad (11)$$

Now with the help of L and X as defined above and with little algebra, the set of values of delay represented by $d=d_i$ correspond to different sets of values of x,k and C, from equation (11) we have for $d \geq 0$,

$$d= \tan(((\pi*Tc)/(2*r*C)*[\ ((1-1/Tc)^{(k+x)/x} - (1-1/Tc)^{(k+x)})/(1-1/Tc)^{((x*x+k*x+k+x)/x)}] \qquad (12)$$

Now generalizing the above equation if d_i represents the various delays for i iterations corresponding to the above parameters associated to number of rtps,nrtps and intensity such that $d \geq 0$.Thus we have the main result as :

$$d_i=\tan(((\pi*Tc)/(2*r*C)*[\ ((1-1/Tc)^{(k+x)/x} - (1-1/Tc)^{(k+x)})/(\ 1-1/Tc)^{((x*x+k*x+k+x)/x)}],$$
$$\text{However } d_i \geq 0 \qquad (13)$$

here $x_i=i, 0 \leq i \leq 10$.However,d_i, C_i, k_i all will take real values under the investigation as given below:

Now we determined the solution set (d_i) corresponding to the various parameters C_i, x_i and k_i . As the parameter C_i increases, the delay d_i decreases because d_i and C_i are inversely proportional to each other. Interestingly we find delays obtain corresponding to the value C_i and x_i and k=1 turn out to be the same in the below given tables[1-3] as given in [2] and for the values other than k=1 we obtain the various forms of delays with regard to rtps via-vice nrtps in subscribed paper. Further we note that if the delay requirement is d_{req}, then we select the smallest parameter C_i, which satisfies $d_i \leq d_{req}$.

4 Simulation Result

The analysis has been done using Matlab for values of d_i(delay) corresponding to different prescribed values of x_i, k_i and $C_{i.}$,for $1 \leq i \leq 4$ as given in the following tables:

Case I:

Table 1. C_i = .05(intensity of the delay requirement in the rtps connection)

k \ x	1	2	3	4
1	0	$0.0845*10^{-3}$	$0.1501*10^{-3}$	$0.2111*10^{-3}$
2	0	0.0085	0.0141	0.0190
3	0	$0.3865*10^{-3}$	$0.6183*10^{-3}$	$0.8113*10^{-3}$
4	0	0.0065	0.0101	0.0129

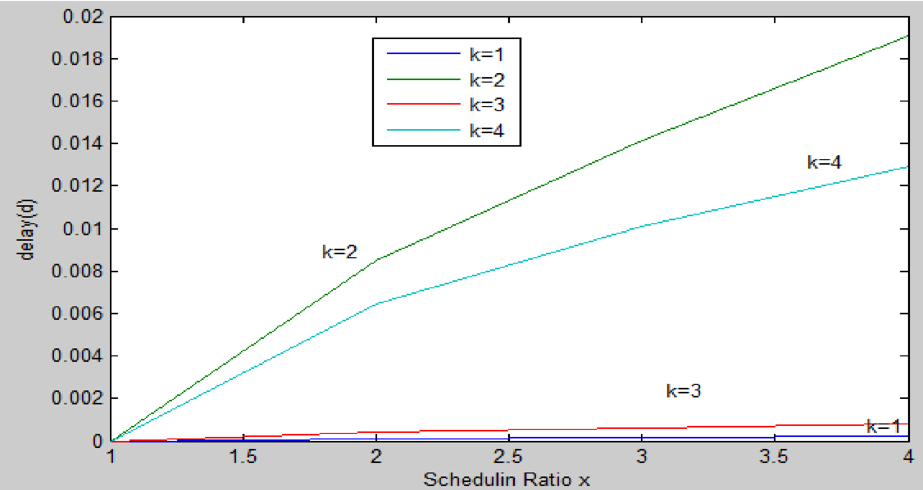


Fig. 2. Delay against Scheduling ratio x when C=.05

Case II:

Table 2. C_i = 0.08(intensity of the delay requirement in the rtps connection)

k x	1	2	3	4
1	0	$0.0527*10^{-3}$	$0.0937*10^{-3}$	$0.1317*10^{-3}$
2	0	0.0025	0.0041	0.0056
3	0	0.0028	0.0045	0.0059
4	0	$0.1799*10^{-4}$	$0.2799*10^{-4}$	$0.3598*10^{-4}$

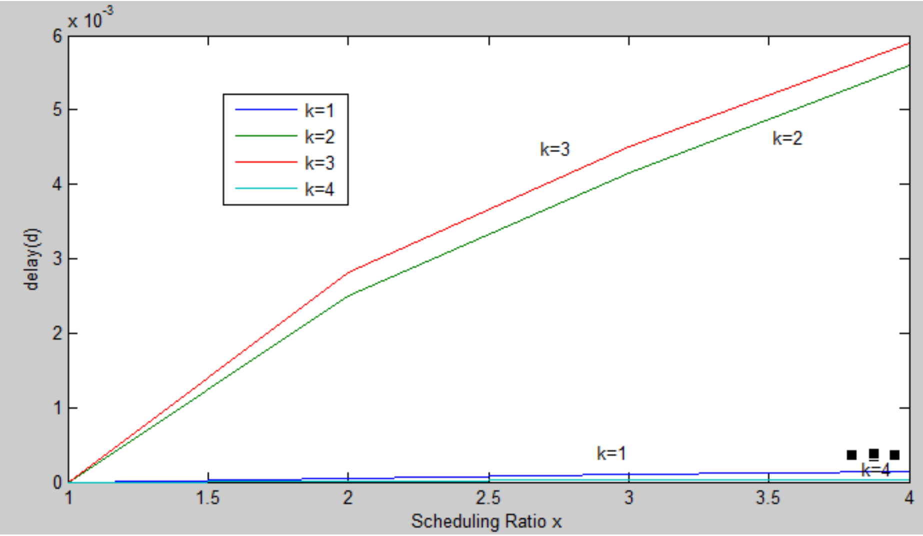


Fig. 3. Delay against Scheduling ratio x when C=.08

Case III:

Table 3. C_i = 0.1(intensety of the delay requirement in the rtps connection)

k x	1	2	3	4
1	0	$0.0422*10^{-3}$	$0.0749*10^{-3}$	$0.1053*10^{-3}$
2	0	.0016	.0026	.0035
3	0	.0146	.0233	.0306
4	0	.0015	.0023	.0029

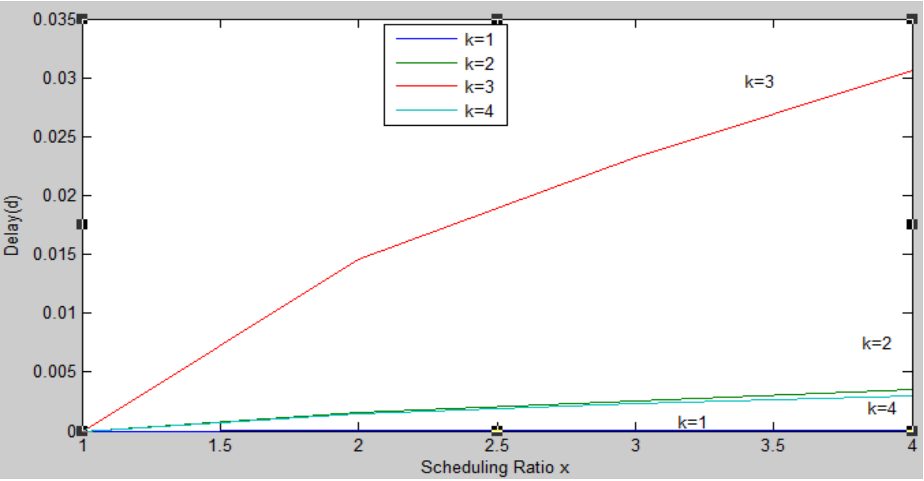


Fig. 4. Delay against Scheduling ratio x when C=.1

Table 4. Simulation Information

Parameter	Value
Packet Size	1500 bytes
Number of nodes	10
Delay requirement	30ms

5 Conclusion

In this paper, scheduling scheme in IEEE802.16 Network has been proposed and simulation results have been discussed. To support the QoS requirement the delay requirement term in the proportional fair scheduling scheme has been added. The main contribution of this paper is that a method has been proposed which will generalize the delay requirement by associating various parameters of x_i defined as (various) rtps connections to the parameter k_i associated to the (various) nrtps connections. The suggested general scheduling scheme satisfies the delay requirement. One can find the appropriate parameter C according to the traffic condition of the networks. After fine tuning of the operating parameter, the delay requirement can be satisfied without excessive sacrifice in the nrtps connection performance.

References

[1] Jain, A., Verma, A.K.: Comparative Study of Scheduling Algorithms for WiMAX. Published in TechRepublic (2008)

[2] Kim, T., Lim, J.T.: Quality of service supporting downlink scheduling scheme in worldwide interoperability for microwave access wireless access systems. Published in IET Communications 4(1), 32–38 (2010)

[3] Rashwan, A.H., ElBadawy, H.M., Ali, H.H.: Comparative Assessments for Different WiMAX Scheduling Algorithms. In: Proceedings of the World Congress on Engineering and Computer Science, October 20-22, vol. 1, WCECS, San Francisco (2009)

[4] Belghith, A.: Pricing-Based Schedulers for WiMAX. In: IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (2009)

[5] Wee, K.K., Lee, S.W.: Priority Based Bandwidth Allocation Scheme For Wimax Systems. In: Proceedings of IC-BNMT (2009)

[6] Zhao1, C., Hu2, J., Zhou2, J., Shi2, J., Dutkiewicz3, E.: Channel Quality Dependent Rate-limited Scheduling Algorithm for IEEE 802.16 Wireless Networks. In: International Conference on Communications and Mobile Computing (2009)

[7] Lakkakorpi, J., Sayenko, A., Moilanen, J.: Comparison of Different Scheduling Algorithms for WiMAX Base Station. In: Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2008), Las Vegas, Nevada, USA, 31 March -3 April, pp. 1991–1996 (2008)

[8] Sushil, K.: Hybrid scheduling For QoS in WiMAX. Master Thesis, Thapar University, Patiala,June (2009)

[9] IEEE Standard for Local and metropolitan area networks, Part 16: Air Interface for Broadband Wireless Access Systems

[10] CDMA-HDR a high efficiency-high data rate personal communication wireless system. In: Proc. IEEE Vehicle Technology Conf., Tokyo, vol. 3, pp. 1854–1858 (May 2000)

[11] Hou, F., Ho, P., Shen, X., Chen, A.: A novel QoS scheduling scheme in IEEE 802.16 networks. In: Proc. IEEE WCNC, Hong Kong, pp. 2457–2462 (March 2007)

[12] XIaojing, M.: An Efficient Scheduling for Diverse QoS Requirements in WiMAX. Thesis presented in the University of Waterloo, Ontario (2007)

A Framework of Service Selection and Composition for Flexible Network Architecture

Akhilendra Pratap Singh, O.P. Vyas, and Shirshu Varma

Indian Institute of Information Technology Allahabad-211012, India
akhil121282@gmail.com, {opvyas,shirshuvarma}@iiita.ac.in

Abstract. Internet becomes a back bone of the communication network in the current era. People are very much dependent on the internet to connect from one end to another globally in todays environment. Internet suffers from the problem with the tightly coupled layered architecture. So there should be flexible architecture of internet. Flexible architecture will reduce the tight coupling of the layers. Selection of service and Composition of service is also an issue which will be solve through our framework. So we propose a framework of flexible network architecture which will provide flexibility to user for their application. This framework has many advantages like availability of services, load distribution, service discovery through the cloud technology. Cloud has a group of nodes which will act as single node to discover service and provide a service .It is a capable to create a mirror image of the broker node. Flexible network architecture is a combination of blocks. These blocks are functional module. Our framework will maintain also performance of the network.

Keywords: Future Internet, Service oriented architecture, Flexible network,Cloud management,Service selection and composition.

1 Introduction

Nowadays Internet is a well known term in the communication environment. Internet plays an important role as a medium to transfer the information from one end to another end of the world. In the current real communication, internet plays a role as a back bone and use to connect the entire world. Today in fast growing environment people are dependent on the internet to perform a various task such as information gathering, searching, resource sharing, software, audio, video data, selling and buying a goods like books etc. Internet provides a platform to develop a new thing such as software, instruments, architecture etc. which will be fruitful for the society.

In 1985 internet was a collection of 50 sites and 1000 nodes approx [1] to connect the group of people and researchers but today internet spread itself with very fast growing speed due to its characteristic. Billions of user depends on the internet as a logical resource. So it is responsibility, today's present researchers that all resources must be available, up-to-date, fault tolerance reliable. All these important factors can affect current architecture of the internet. Current internet

K. Singh, A.K. Awasthi, and R. Mishra (Eds.): QSHINE 2013, LNICST 115, pp. 998–1007, 2013.

architecture has various advantages but there is also some limitation. Current architecture of internet is layered architecture, one layer depends on other adjacent layer. In layered architecture upper layer use the service of lower layer and vice versa. Header format of the packet is also another overhead of the network which consumes the extra bandwidth to transfer the information.

Layered architecture is tightly coupled from each other if any change is perform in one layer then there must be a change in the adjacent layer. Always it is difficult to change in the functional module of all layers. It will increase the complexity in modules of the layers. In the current Scenario people thoughts, way of thinking and requirement dynamically change with fast growing speed at technological level. In layered architecture every time to fulfill the requirement of people is not easy because it require changes in every layer. So some researchers suggest an idea of loose coupling to remove these problems from the layered architecture. There is no need to change at every layer when user will perform the change at any layer of architecture. So they propose a new architecture after realizing these problems of layered architecture which is a loosely coupled architecture. It is also known as today's Flexible network architecture [16].

Flexible network architecture [22] is a block wise information exchange communication system. It is based on the loosely coupled approach that is a reason to say the flexible network architecture. Service oriented architecture is defined as a loosely coupled system. Flexible network architecture provides a facility to change the function or extend the function of module and there is no need to modify in the adjacent layers. Users are facilitated with flexible network architecture to do the modification according to rapid change at technological point. Future internet [27] is rich from the components (module),these components can be a procedure, structure of data etc. Future internet is very much useful for the future application. Modules in the future internet is described as block and these blocks are collection of functions. These services are used to perform a task which will select the users. Service oriented Architecture principle is helpful to remove the rigidity of the current network architecture.

Why Redesign the Network Architecture?

* The goal of network design is to improve the performance of network architecture [20], flexibility of user and selection of a service to complete the requirement of application. Service model is a solution to complete the requirement of application. Design of network architecture is crucial issue because various type of questions will arise such as:

- What should be the parameter?
- Why we should change the architecture of existing one?
- How we will evaluate the performance of network architecture?
- Which type of architecture will be feasible to fulfill the requirements of new application?

There are various other important Parameters on which network architecture will evaluate like speed, delay, packet loss, availability, reliability, flexible to select a service etc.

Issues in Future Internet

Internet architecture was developed some years ago to provide a communication between the people but in current situation it has changed the shape of modern society. In the past there were less user, whose requirement were fixed but the present scenario has witnessed a change. Users demands are also increasing along with the growing population and education. There are various issues [25] which can be come in the future. Some issues are present and some will arise in future. Flexible network architecture can play an important role to solve theses issue which are given below:

- Processing power.
- Service impact on the network.
- Security.
- Context and location awareness of service.
- Mobility.
- Addressing scheme.
- Multimedia support.
- Heterogeneity.

1.1 Impact of Flexible Network on the Global Environment

Today good investment is not only reducing the cost at operational point, there are also various points for eg.reducing the management challenge, technology cost and energy consumption. Flexible network architecture can give a good result on the IT industry. Every organization needs the deployment of product, its services and use with very fast speed in current scenario. This will help to cut the investment of organization in terms of maintenance and management. When organization will monitor the cost then it will be easy to validate at architectural level. These are the reasons to validate like:

- Enhancement of resource sharing capability.
- Service and components are reusable.
- Interoperability between heterogeneous components.
- Integration of system at application level.
- Services are manageable using cloud technology.

All above things will be beneficial to improve the performance of organization and help to cut the budget of investment. Theses all reasons will affect the society in future directly or indirectly.

1.2 Technological View

Current Internet Architecture has various major issues which can be resolved with the Flexible network architecture. Loose coupling is a method which has been used to resolve the problem [26] of tightly coupled architecture. Loose coupling is a characteristic of Service oriented architecture. In this architecture there are three entities Service Provider, Service consumer and Service broker. Service Provider provide a facility of available Service, Service Consumer Send a request to consume the Service and Service broker provides information of available services. There are instances when services broker suffers from many request at the same time, Broker can be off due to failure of hardware or power. Service provider may also suffer from the same problem. These are the major problems related to the availability of t resources.

From the above study, the major issues essential for optimal procedures are:

- Efficient Service discovery using broker.
- Providing Fault tolerance broker mirroring and peer broker.
- Load balancing among the Peer broker and service broker.
- Auto configuration of peer broker during very high load period.

2 Related Work

Till now Service selection and service composition, availability, service discovery and delay management [24] are challenging in flexible network architecture. There are very few research suggestions present to address these issues till date.

Flexible network architecture is shaped to solve the problem of rigidity in layered architecture. Very few researchers discussed the issue of internet architecture. Scott Shenker [1] has been discussed some issue of current internet architecture will arise in future. Internet should accept a new service model, various other issues also covered [3]. What will be the cost of framework [19] and options of architecture? MIT Lab projected a research report [2] in which researchers tackle the problems like design of architecture from existing one, process to reshape that etc. In this report researchers discussed that we should work for future not for a a short period of time in order to get a result. Some little modification proposed to resolve the problem for short period of time but it is not a future vision because with these modifications in the current architecture some other difficulty may occur which will be tackled in future. According to the report there must a complete solution but previous solution are dependent to each other like the protocol for application is modified then the application needs to be modified. Service oriented model [5][9][10] gives a facility to select and configure the protocol separately according to the situation. Rudra Dutta et al [6] have suggested an idea for architecture to integrate, organize and optimize the service. In this approach framework is a group of building blocks [15], open control to connect these blocks for the communication and control fundamentals will provide a facility to interact at cross-layer [23]. Holistic concept [25] is used to do the service selection, flexibility of building block according to the

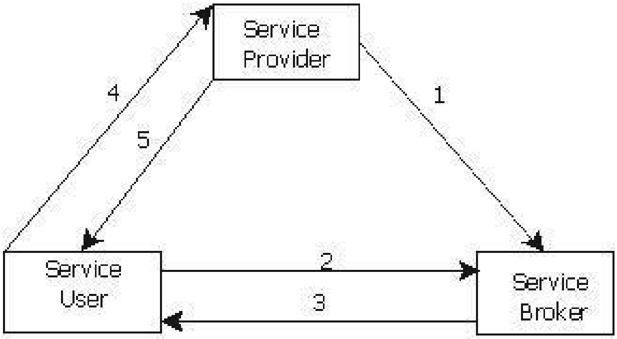


Fig. 1. Service Oriented Model

requirement of application. Flexible architecture can be modified at user end to support internet and integration process is used to solve the security issue with networking stack. Michael P. Papazoglou et al proposed an idea [7] [11] for service oriented environment to join the building block of application. Loose coupling is a key factor for flexibility and dynamic change in the architecture. Service oriented computing environment facilitate user to develop a new architecture.

Mike P. Papazoglou and Willem-Jan van den Heuvel [8] have introduced the necessities of distributed computing with some other factor like standard and loose coupling. To construct service oriented architecture, distributed communication environment and integration of components [30] are highly required. This paper gives an idea of Enterprise service bus which will work as a back bone to connect the different homogeneous and non homogeneous system with its component [21]. Joseph D. Touch et al [10] shows the relation between network architecture and protocol through recursive network architecture. In this paper [31] author provide an interaction with cross layer to sustain dynamic service composition and discovery which will not generate another problem in future direction. Bernd Reuther et al [4] have introduced a model given in figure1. International telecommunication unit presented a report [12] where internet architecture was designed in 1970's with a simplified model to support protocols and its functioning at different layers. MIT Professor Dave Clark in 80s presented an article in which Internet is broken and will appear in future. Various modifications have been proposed to reshape the current network architecture like network virtualization, cross layer design, Security, Reliability etc. National science foundation has invested 20 Million for two projects, these projects are:

- Future Internet Design.
- Global Environment for network improvement.

Many other research communities are also working to redesign the structure of existing internet architecture.

Dennis Schwerdel et al [13] found that protocols have been developed many years ago without keeping in mind todays problem. Previous protocols configuration and implementation is not easy because of fixed structure. So author proposed a way to select and compose dynamically based on the SOA principle. Thomas Ristenpart et al [14] have proposed an approach to manage the memory and operation on the virtual machine that may be placed in the internal network or beyond it. Paul Muller [17] has suggested an idea to manage the architecture of internet and cloud [29] with the service oriented principle. Internet is software system which is deployed in distributed environment to achieve a goal for future internet [32]. Ivan Seskar et al proposed approach [18] to solve the issues of mobility and wireless access. They provides a method to any cast, multicast, multipath [28] for the application of internet. In this author has implemented a mobility stack of protocol on the GINI platform.

3 Proposal to Major Issue

Many proposals have been give to solve the problems of current architecture of internet. We proposed a framework for flexible network architecture to solve the issue of the existing internet architecture. In this framework we are using the principle of service oriented architecture to provide the flexibility in the layered network architecture. This framework is able to tackle the issue of current layered architecture. Framework consists with various types of services which will facilitate to the internet application. These services provide flexibility to the user select the service as per the requirements. The proposed frame work is given in fig 2.

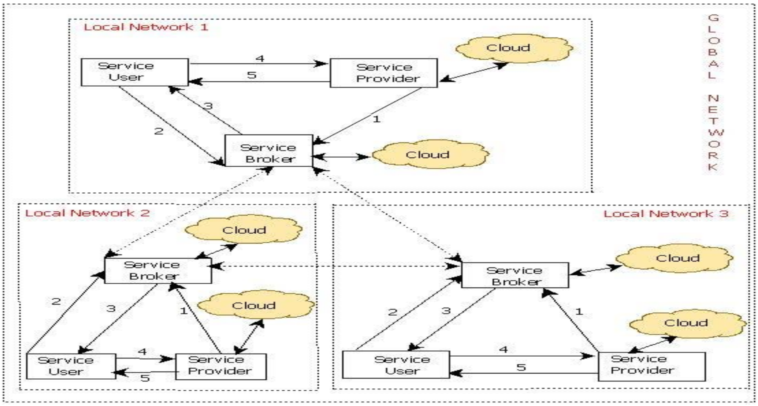


Fig. 2. Framework of Service selection and Service composition

Proposed framework consists of four major modules which are 1.Service provider 2. Service user 3. Service broker 4. Cloud of service. In this framework following process should be follow to get the reply for a service.

1. Service provider will develop a service and register itself to the service broker.

2. Service user will request to the broker through the API to Service broker.
3. If service identification details are present the broker will reply to the user.
4. When user will get the reply from broker then user will request to the service provider.
5. Service provider will reply to the user after execution of request.

Above framework has a facility of service cloud which will manage the services, that is not available at service provider. It is capable to manage the availability of the resources. Availability of Service provider and Service broker will be controlled through the cloud management. Management of service broker is also issue because if broker will goes down then what will happen for request. To manage service broker ,we will use the principle of cloud computing but in the cloud computing, security issue will arise. Security issue can be solved in that scenario through the authorized and authenticated system. Geographical area will also be one of the constraints which can affect the efficiency of network management. Cloud of the system cannot be far away from another service broker. Request of many users at a time will also be handled with the help of cloud computing. If service provider has many requests than its capability, then performance of the service provider will decrease and then it can transfer the request to another service provider of the cloud to balance performance of network . Each system of the cloud has some threshold value according to its hardware and software devices.

Request of the users, if greater than threshold then Service provider will transfer the request to another provider, maintain the balance of execution otherwise Service provider process and reply. So from the above discussion we can say that our proposal is feasible to target the given issues.

Proposed Algorithm

The algorithm explain various phases to find the service which is the part of proposed Framework. In this requester query will be raised to discover the service at broker site. Broker will search the service in to reply the user . User will select the service from list of services and execute at service provider end. Service providers will be reply to user in the form of service. Hence the algorithm will break in three phase .

1. Service provider to Service broker.
2. User to Service broker.
3. User to Service Provider.

Service_id is a unique identification number of service. *Service_list* is the collection of all existing service in database. List is structure type data which consist of address of service provider,type, *service_id* and some other essential parameter. Broker cloud is a function which search *broker_id* of appropriate service for requester.

Algorithm 1. Algorithm of Service selection and composition

```
                                                    ▷ Requester to Broker
while true do
    if ($Service\_id = true$) then
        $list \leftarrow list[Addr, Service\_id, Type, parameter]$ return $list$
    else
        Broker_cloud($Service\_id$)
    end if
end while
function BROKER_CLOUD
    list_Broker;
    while true do
        if ($Service\_id = Broker\_service$) then return $Service\_list$
        end if
    end while
end function
```

New_service is type of newly developed service which will identified by *Service_id*.

```
                                                    ▷ User to Service provider
while true do
    if ($Service\_P\_id = Addr$) then
                    ▷ $Service\_P\_id$ is a unique identifier for service provider
        $execute(Service\_request)$ return $Service$
    end if
end while
                                                    ▷ Service provider to Broker
if ($New\_Service = true$) then
    while $New\_service\_type = exist\_service\_type$ do
        Add $Broker\_list[] = New\_service$
    end while
end if
```

4 Advantage of Proposed Approach

A proposed framework is feasible for service selection and composition. It has various advantageous over the existing layered architecture. This framework provides a flexibility to select the service among various other related services. Due to this reason users have options for selection and composition of a best service according to requirements. Framework provides a facility to connect the global environment in terms of services. Each and every user think and perform the things in terms of service on the internet .In this user can change the functional module of service according to requirement on application layer but there is no need to change in adjacent layer. Framework will support the dynamicity of user and environment to fulfill the demand of users. This framework will remove the rigidity of tightly coupled layered architecture.

5 Conclusion

In this paper we have converse all the issue in our solution. Internet will have various issues in the future, some of them has been discussed in this paper. Our framework is able to handle the issues of the current internet architecture. This proposed framework will maintain the performance of the network and provide a facility to user for selection and composition of a best services according to the requirements of application. So from above discussion management related issue of framework like resource updation, availability and reliability will be solved in future.

References

1. Shenke, S.: Fundamental Design Issues for the Future Internet. IEEE Journal on Selected Areas in Communication 13(7) (September 1995)
2. Debany, W.: New Arch: Future Generation Internet. In: MIT Computer Science and AI Lab. Final Technical report (August 2004)
3. Jain, R.: Internet 3.0: Ten Problems with Current Internet Architecture and Solutions for the Next Generation. In: Proceeding of IEEE Military Communication Conference, Washington, October 23-25 (2006)
4. Reuther, B., Henrici, D.: A Model for Service-Oriented Communication Systems. In: Proceeding of 32nd EUROMICRO Conference on Software Engineering and Advanced Applications (2006)
5. Matthew MacKenzie, C.: Adobe Systems Incorporated Reference Model for Service Oriented Architecture 1.0 In the report of MITRE Corporation (October 12, 2006)
6. Dutta, R., Rouskas, G.N., et al.: AThe SILO Architecture for Services Integration, control, and Optimization for the Future Internet. In: The Proceeding of IEEE Communications Society Subject Matter Experts for Publication in the ICC (2007)
7. Papazoglou, M.P., Traverso, T.P., et al.: Service-Oriented Computing: State of the Art and Research Challenge. The Proceeding of IEEE (November 2007)
8. Papazoglou, M.P., van den Heuvel, W.-J.: Service oriented architectures: approaches, technologies and research issues. The Proceeding of Springer, VLDB Journal (2007)
9. Muller, P., Reuther, B.: Future Internet a service oriented Approac. The Project Report of AGICSY (July 2007)
10. Touch, J.D., Pingali, V.K.: The RNA Metaprotocol. The Proceeding of IEEE (2008)
11. Reuther, B., Henrici, D.: A model for service-oriented communication system. The Proceeding of Journal of System Architecture (December 12, 2007)
12. Mueller, P., Reuther, B.: Future Internet Architecture A Service Oriented Approach (2008)

13. Schwerdel, D., et al.: Composition of Self Descriptive Protocols for Future Network Architectures. In: The Proceeding of IEEE, 35th Euromicro Conference on Software Engineering and Advanced Applications (2009)
14. Ristenpart, T., et al.: Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds. The Proceeding of ACM, November 12-15 (2009)
15. The Future Internet ITU-T Technology Watch Report (April 10, 2009)
16. Harai, H., et al.: New Generation Network Architecture AKARI Conceptual Design (ver2.0), AKARI Architecture Design Project Original Publish (August 2009) (Japanese)
17. BB-Framework How to Version1.0 (February 1, 2010)
18. Pan, J., et al.: A Survey of the Research on Future Internet Architectures. The Proceeding IEEE Communications Magazine (July 2011)
19. Seskar, I., et al.: MobilityFirst Future Internet Architecture Project. In: The Proceeding of Mobility First Project, Proc. ACM AINT 2011 (2011)
20. Khondoker, R., et al.: A Description Language for Communication Services of Future Network Architectures. In: The Proceeding of International Conference on the Network of the Future (2011)
21. EC FIArch Group: Fundamental Limitations of current Internet and the path to Future Internet. In: The Proceeding of Current Internet and the Path to Future Internet (March 2011)
22. Salvador, T.S.: Dynamic and Heterogeneous Wireless Sensor Network for Virtual Instrumentation Services. In: The Proceeding Eighth IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (2011)
23. Gonzalez, A.J., et al.: Costing framework for service oriented Future Internet architectures: empowering requesters choice. The Proceeding of ACM (2011)
24. Mueller, P.: Future Internet Architecture and clouding Computing: A Service Oriented Architecture. Future Internet Architectures Poznan (October 11, 2011)
25. Manu, A.P., et al.: Design and Implementation Issues of Flexible Network Architecture. In: International Conference on Computational Intelligence and Communication Systems (2011)
26. Khondoker, R., et al.: Describing and Selecting communication services in a Service Oriented Network Architecture. In: The Proceeding of ITU-T Kaleidoscope Academic Conference (2010)
27. Becke, M., et al.: A Future Internet Architecture supporting Multipath Communication Networks. The Proceeding of IEEE (2010)
28. Baldine, I., et al.: A Unified Software Architecture to Enable Cross-Layer Design in the Future Internet. The Proceeding of IEEE
29. Giese, H., et al.: Software Engineering for Self-Adaptive Systems. In: The Proceeding of Dagstuhl Seminar Software Engineering for Self-Adaptive Systems (2011)
30. A Report on the Future of Internet, ICT European Project the EU 7th Framework Program (2011)
31. Yanggratoke, R.: The Amazon Web Services, TKK T- 110.5190 Seminar on Internetworking (2009)
32. Siddiqui, A.: Tradeoffs in Selection and Composition Approaches for Future Internet Architectures. The Proceeding of IEEE (2011)

Author Index